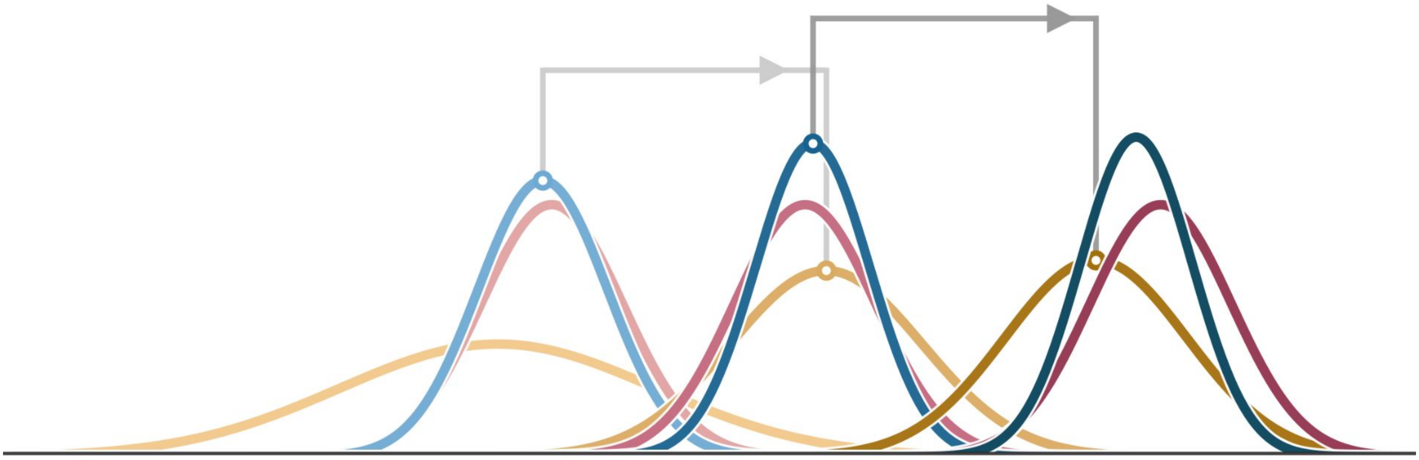
Bayesian Models of Perception and Action

An Introduction



Wei Ji Ma

Konrad Paul Kording

Daniel Goldreich

Preamble

Bayesian models of perception and action. (Previously also distributed as *Bayesian models of behavior*.) Copyright Wei Ji Ma, Konrad Kording, Daniel Goldreich, 2021, with original artwork by Brennan Klein.

This is only a draft. Comments are welcome on www.bayesianmodeling.com. The book will be published by MIT Press.

The LaTeX template for this draft was based on The Legrand Orange Book, downloaded from `http://www.LaTeXTemplates.com`.

Dedication
We dedicate this book to the memory of David Knill (1961-2014). All three of us have learned a good part of what we know about Bayesian modeling of perception and action from him. As a caring and patient mentor and as an excellent teacher, he also made studying this topic a lot more enjoyable for all of us. The field of Bayesian modeling of perception and action would not be where it is without him and this book would probably never have been written.

Acknowledgments
Numerous. Will go here.

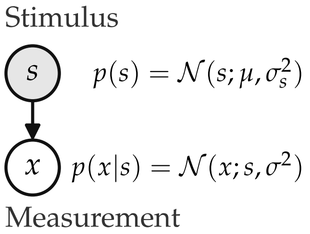
The four steps of Bayesian modeling

Example from Chapters 3 - 4

Step 1: Generative model

a) Draw a diagram where each node is a variable and each arrow a statistical dependency. Observation/measurement is at the bottom.

b) For each variable, write an equation for its probability distribution. For the observation, assume a noise model. For others, get the distribution from your experimental design. If there are incoming arrows, the distribution is a conditional one.



$$p(s) = \mathcal{N}(s; \mu, \sigma_s^2)$$

$$p(x|s) = \mathcal{N}(x; s, \sigma^2)$$

Step 2: Bayesian inference (decision rule)

a) Compute the posterior over the world state of interest given an observation. The optimal observer does this using the distributions in the generative model. Alternatively, the observer might assume different distributions (natural statistics, wrong beliefs). Marginalize (average) over variables other than the observation and world state of interest.

b) Specify the read-out of the posterior. Assume a utility function, then maximize expected utility under posterior. (Alternative: sample from the posterior.) Result: decision rule (mapping from observation to decision). When utility is accuracy, the read-out is to maximize the posterior (MAP decision rule) and the decision is a world state estimate.

$$\mathcal{L}(s; x) = p(x|s)$$

$$p(s|x) \propto \mathcal{L}(s; x) p(s)$$

$$p(s|x) = \mathcal{N}\left(s; \frac{J_s\mu + Jx}{J_s + J}, \frac{1}{J_s + J}\right)$$

$$\hat{s} = \frac{J_s\mu + Jx}{J_s + J}$$

$$J_s \equiv \frac{1}{\sigma_s^2} \text{ and } J \equiv \frac{1}{\sigma^2}$$

Step 3: Response probabilities

For every unique trial in the experiment, compute the probability that the observer will choose each decision option given the stimuli on that trial. For this, use the distribution of the observation given those stimuli (from Step 1) and the decision rule (from Step 2).

- Good method: Sample observations according to Step 1; for each, apply decision rule; tabulate responses.
- Better: Integrate numerically over observation.
- Best (when possible): Integrate analytically over observation.

• Optional: Add response noise or lapses.

$$p(\hat{s}|s) = \int p(\hat{s}|x) p(x|s) dx$$

$$= \mathcal{N}\left(\hat{s}; \frac{J_s\mu + J_s}{J_s + J}, \frac{J}{(J_s + J)^2}\right)$$

Step 4: Model fitting and model comparison

a) Compute the parameter log likelihood, the log probability of the subject's responses across all trials for a hypothesized parameter combination.

b) Maximize the parameter log likelihood. Result: parameter estimates and maximum log likelihood. Test for parameter recovery and summary statistics recovery using synthetic data. Use more than one algorithm.

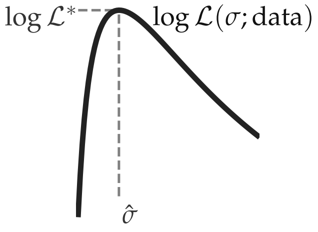
c) Obtain fits to summary statistics by rerunning the fitted model.

d) Formulate alternative models (e.g. vary Step 2). Compare maximum log likelihood across models. Correct for number of parameters (e.g. AIC). Test for model recovery using synthetic data.

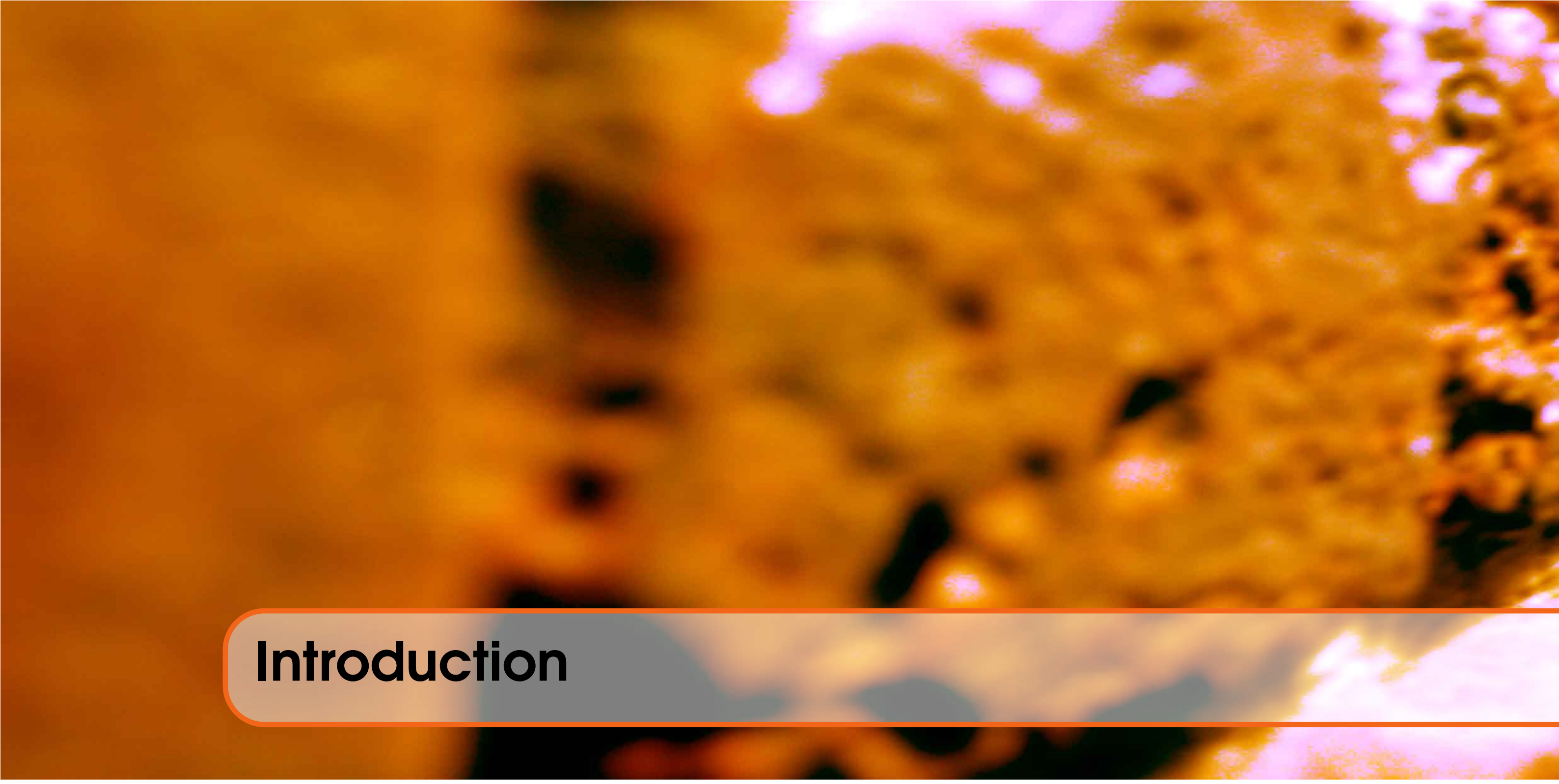
e) Check model comparison results using summary statistics.

f) Optional: evaluate absolute goodness of fit.

$$\log \mathcal{L}(\sigma; \text{data}) = \sum_{i=1}^{n_{\text{trials}}} \log p(\hat{s}_i | s_i, \sigma)$$



Contents



Introduction

This book provides a gentle introduction to a rigorous quantitative framework for understanding the role of probabilistic inference in perceptual decision-making and action. *Probabilistic, or Bayesian, inference* is a method used to draw conclusions from uncertain evidence. This book explains how many forms of perception and action can be mathematically modeled as Bayesian inference. According to these models, the human mind behaves like a capable data scientist (or crime scene investigator, or diagnosing physician, . . .) when dealing with noisy and ambiguous data. In recent decades, the Bayesian approach to perception and action has become increasingly popular and widely tested.

Inference plays a central role in perception. Our eyes, ears, skin, and other sensory organs register physical signals, and convert these into electrical impulses that travel towards the brain, a sort of neural Morse code. The brain must decode these signals and draw inferences from them regarding the state of the world. The eyes register patterns of light, but do not identify the visual scene. The skin senses pressure and vibration, but does not identify the external object causing these stimuli. The muscles sense tension, but do not unambiguously signal the configuration of the body. The ears detect sound waves, but do not indicate their meaning. The brain undertakes these difficult interpretive tasks, coming up with a perceptual best guess about the world from the sensory information it receives.

The sensory information on which perception is based is typically open to multiple possible interpretations. When recognizing sound waves as spoken words or as the chirps of birds, when interpreting a visual image as the face of a friend or as a clock on the wall, when navigating our environment on foot or in a car, we are guided by our inferences. Our inferences are usually correct, but in perceptual illusions, the lack of full knowledge becomes apparent. The central tenet of this book is that perception is a form of probabilistic inference: from incomplete and ambiguous sensory observations, the brain strives to figure out the state of the world.

Perception is a process of probabilistic inference because the sensory information available to the brain is typically only partially informative. For instance, the sensory input might be of low quality (objects might be poorly lit, far away, moving fast, similar to the background, etc.) or, even when its quality is high, the input might be compatible with two or more interpretations (e.g., the equivalently pronounced words "red" and "read," or a two-dimensional retinal image that is compatible with multiple three-dimensional objects). For these reasons, two or more interpretations are often plausible for the same sensory data. This ambiguity leaves the observer uncertain as to the state of the world.

Clearly, performing probabilistic inference under such circumstances can never be error-free. However, among all possible strategies that can be used for solving a perceptual inference task, there is always a *best possible* one. This strategy is called optimal probabilistic inference or optimal Bayesian inference. It consists of computing the probability of each possible interpretation of the observations, and then acting in a manner that has the greatest expected benefit. For instance, if the goal is to make as few errors as possible, the best strategy is for the organism to perceive the interpretation that has the greatest probability of being correct. Optimality does not mean making no mistakes. It means drawing the best conclusion possible, given the information available to the observer. (Suboptimal Bayesian inference also exists, as we will discuss.)

The Bayesian approach to modeling perception explored in this book is exciting to us because it explains a wealth of data and has successfully accounted for the results of many experiments. Within the Bayesian framework, the goal of the organism is to compute probability distributions over parameters describing the state of the world. This computation is based on sensory information and knowledge accrued from experience. The particular sensory information and prior knowledge are specific to the task at hand, but the computation conforms to the same rules of probability calculus in every case. The Bayesian approach thus unifies an enormous range of otherwise apparently disparate behavior within one coherent framework.

Positioning Bayesian models within the space of models

There is a rich history of models of behavior and hence, scientists have long sought ways of characterizing the logical structure of models. One particularly useful scheme is to categorize models as either *descriptive models, process models* or *normative models*.

A *descriptive model* is a mathematical description of behavioral variables (such as accuracy or reaction time) in terms of input variables (such as intensity of a stimulus or a quantified personality trait). These descriptions may take the form of a regression or a generalized linear model or even a deep learning model. The kind of statement supported by fitting descriptive models to data is simply that a model exists that can fit data with a certain amount of error.

A *process model* is more ambitious: it tries to dissect the mapping from input to output into generalizable, often psychologically meaningful component processes, expressly specifying how the observer/agent makes a decision based on the information available. Examples of such component processes would be "add Gaussian measurement noise to a sensory variable" or "the observer maps the decision variable to a decision by applying a criterion". One popular model from this class is the drift-diffusion model. The goal of a process model is not just to fit the data, but to understand the structure of information processing. Process models are often non-linear, but not in arbitrary ways: the specific non-linearities are direct consequences of the assumptions used in constructing the model. The kinds of statements supported by fitting process models to data are that if the process is as hypothesized then the measured behavior would result.

A *normative model* is ambitious in yet another way. In a normative model, we ask why behavior is the way it is. More specifically, we ask why, in a particular ecological niche, certain behaviors are beneficial. For example, a normative model may (explicitly or implicitly) assume that maximizing accuracy is important in our lives. Such a model might then make assumptions about the world that are viewed as immutable (e.g. that there is a fixed amount of noise in vision). The model could then derive the optimal solution to the problem and compare this with actual behavior. The kinds of statements supported by fitting normative models to data are that if the world or ecological niche is as hypothesized, then the measured behavior would be beneficial.

In light of the above model classification scheme, Bayesian models can be viewed as both normative models and process models. This will become apparent to the reader, as the modeling approaches highlighted in this book address the question "what computations *should* the brain carry out to perceive optimally?"By having a successful mathematical model of optimal performance in a particular perceptual task, we then hope to constrain our understanding of the underlying neural processes and to affect algorithm and implementation level models. This contrasts with a bottom-up approach, in which one might start by modeling small circuits of neurons in biophysical detail, and then attempt to build up the models by combining multiple small circuits. These two approaches are complementary, and both contribute to the understanding of brain function.

The scope and structure of the book

While this book focuses on perception and action, Bayesian models are widely useful in other realms of cognitive science and psychology. In particular, there is a rich history on Bayesian models in higher-level cognition [2], dating back at least to the work of Jonathan Evans [43] and John Anderson [21], with great contributions to the understanding of cognitive development [53, 123]. Higher-level cognition makes occasional appearances in this book, especially in Chapters 6 (Learning), 12 (Inference in a changing world), and 13 (Combining inference with utility). In Chapter 15 (Bayesian models in context), we comment on differences between perception and cognition.

We decided to write this book because there was no accessible text that teaches the reader to build Bayesian models. This is not to suggest that excellent Bayesian materials are unavailable. However, review papers are generally too qualitative and focused on recent results to be practical to an aspiring modeler, and chapters in contributed books on Bayesian inference often assume extensive background knowledge or mathematical expertise that pose difficulties for newcomers to the field. This textbook attempts to fill this gap by providing an elementary introduction.

No previous knowledge of Bayesian inference is expected or required of the reader, but this book does necessarily involve mathematics. If this is your first foray into mathematical modeling: congratulations! Bayesian models are an excellent place to start. The elegant and powerful language of mathematics avoids ambiguity, and mathematical models yield quantitative, testable, predictions. We hope that you embrace the math and are not intimidated by it. Every equation in a good model has an intuitive explanation, and we have tried our best to provide such explanations throughout. Readers with a basic understanding of calculus will find the book accessible; those who are uncomfortable with calculus will still be able to understand the majority of the content.

We recommend that readers take the time to work through the within-chapter exercises and the end-of-chapter problems (which are a mix of concepts, math, and simulations). For building understanding, there is no substitute for struggling with concepts, equations and computer implementations.

A note on Bayesian statistical analysis

In Bayesian models of behavior (or brain function), an observer tries to infer the state of the world from sensory observations. This contrasts with Bayesian statistical data analysis, in which an experimenter tries to infer the value of a model parameter from collected data. The mathematical formalism is the same, but in this book, we focus on how the brain perceives or decides, not on how statistical data can be analyzed. That being said, Bayesian models, like all models, have parameters whose values need to be inferred. For this reason, we include Appendix C on model fitting and model comparison; this appendix, however, does not emphasize Bayesian methods for data analysis.

A note on citations

In this book, we cite the works of many excellent scientists. Without detracting from the achievements of these scientists, we wish to take this opportunity to recognize that academic science has for long – whether deliberately or inadvertently – erected barriers to the participation of individuals from many groups, including women, people of color, immigrants, and individuals of lower socioeconomic status. Consequently, members of these groups have had, and continue to have, far fewer chances to come up with the same ideas or findings, and even when they do, they often face greater challenges in receiving recognition for their work. We hope the reader will keep this in mind when reflecting on citations.

In conclusion

We hope you enjoy the book, and we welcome your feedback on the book's website, www.bayesianmodeling.com. Extra material, including solutions to problems, interactive demonstrations, and further reading, will in the future be available on the same website.

1. Uncertainty and inference

How do we transform our sensory observations into beliefs about the state of the world?

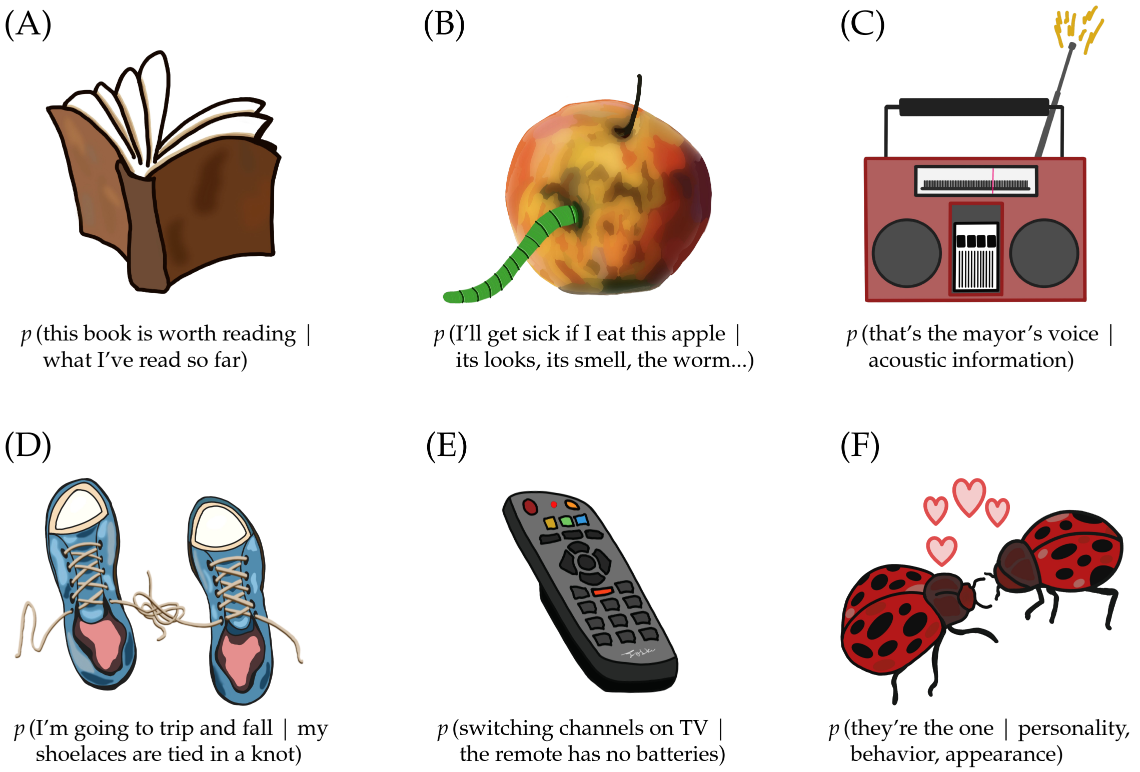
Whenever we perceive something, make a prediction, or deliberate over a decision, we are reasoning with probabilities, even if we do not realize it. We are using the information we have at hand to infer or estimate something else that interests us. The information we have is usually incomplete or noisy, so our inference is not certain. For instance, if we see a floor that is shiny (information we have) this *suggests* that it may be wet (the focus of our interest). Using the available sensory information and any relevant knowledge we may have, we must determine the probability of each interpretation (wet or dry). How can we make sound judgments in such situations? This book instructs the reader in the optimal method for performing inference.

Plan of the chapter

We outline the perceptual inference process, emphasizing the uncertainty that is inherent in perception. Using simple examples, we introduce the probabilities involved in perceptual inference, the likelihood, the prior, and the posterior, focusing on the underlying intuitions. We then illustrate the ubiquity of perceptual inference in daily life with a series of examples involving visual and auditory perception. We do not use mathematics in this chapter but instead explore each example qualitatively and graphically. Our goal is to provide an intuitive understanding of the perceptual inference process, which will serve as a foundation for the more rigorous mathematical treatments in the following chapters.

1.1 The goal of perception

Humans and other animals are endowed with a collection of exquisite sensory organs through which they detect the environment. Sensory organs respond to physical properties as diverse as light (eyes), sound (ears), temperature (skin), material texture (skin), chemical composition (nose, tongue), and body position (joint and muscle receptors, vestibular organs). Our sensory organs form an integral part of ourselves, so much so that we usually take their presence for granted. To appreciate the role that our senses play, try to imagine life without vision, hearing, touch, smell, or taste.



p (this book is worth reading | what I've read so far)

p (I'll get sick if I eat this apple | its looks, its smell, the worm...)

p (that's the mayor's voice | acoustic information)

p (I'm going to trip and fall | my shoelaces are tied in a knot)

p (switching channels on TV | the remote has no batteries)

p (they're the one | personality, behavior, appearance)

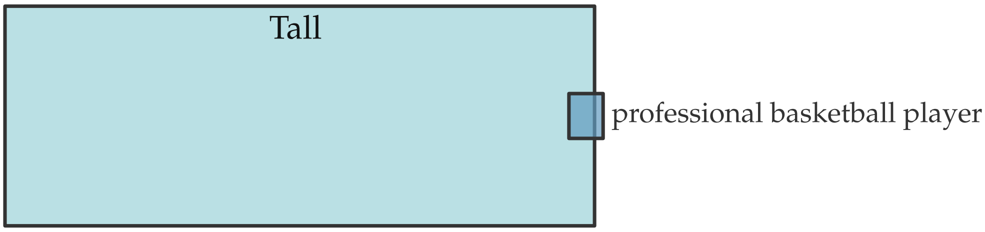
Figure 1.1: Different scenarios with probability judgments. The notation $p(B|A)$ is read "the probability of event B given event A."

As sophisticated as the sensory organs are, their activation by physical stimuli is only the first step in the crucial process of perception. We do not primarily care about the pattern of light wavelengths (colors) and intensities (brightness) entering our eyes, or about the pattern of acoustic energy, varying in amplitude and time, entering our ears. Rather, we care about the interpretation of those sensory inputs. In fact, our quality of life – and often our life itself – depends on our ability to come up with correct interpretations. Does that pattern of light reflect the face of a friend? Is that acoustic waveform the sound of the wind, the howl of a dog, or the voice of our companion? In short, our interest lies not in sensory input per se, but in the information the input provides about the relevant states of the world.

To make the interpretative transition from sensation (the activation of the sensory organs) to perception (a conclusion regarding the state of the world) is a sophisticated task. Broadly speaking, this book is about how the nervous system can optimally accomplish this task. We will examine this issue both at the level of behavior and at the level of neural activity. Our view, based on a large and rapidly growing body of experimental and theoretical work, is that perception is, at least implicitly, an inference process, in which the organism attempts to infer the most probable state of the world, using sensory inputs and all relevant knowledge at its disposal.

1.2 Hypotheses and their probabilities

The transition from sensation to perception requires *conditional probabilities*. A conditional probability is a probability of one event given another: for example, the probability that you are in a good mood given that it is raining outside. We denote conditional probabilities as $p(B|A)$, read "the probability of B given A." Whether people are aware of it or not, we make conditional probability judgments very frequently in daily life (**Fig. 1.1**).



$p(\text{tall}|\text{professional basketball player}) > p(\text{professional basketball player}|\text{tall})$

Figure 1.2: $p(A|B)$ does not in general equal $p(B|A)$. The area of each rectangle represents the probability of the event in the overall relevant set (here, say, all human beings). The overlap of the two rectangles represents $p(A,B)$. Remark: the sizes of the rectangles and their overlap here is conceptual and not calibrated against actual basketball participation data.

Importantly, conditional probabilities are not symmetric. In general $p(A|B) \neq p(B|A)$. For instance, most professional basketball players are tall, but most tall people are not professional basketball players. If A is "being a basketball pro" and B is "being tall", then this example illustrates that $p(A|B) > p(B|A)$.

Exercise 1.1 In each case below, which of the two conditional probabilities is larger and why?

- $p(\text{rain}|\text{clouds})$ or $p(\text{clouds}|\text{rain})$
- $p(\text{speaks French}|\text{born and raised in Paris})$ or $p(\text{born and raised in Paris}|\text{speaks French})$
- $p(\text{unmarried}|\text{college student})$ or $p(\text{college student}|\text{unmarried})$
- $p(\text{you understand Bayes' rule}|\text{you read this book})$ or $p(\text{you read this book}|\text{you understand Bayes' rule})$

■

One way to visualize probabilities is to use areas of rectangles (**Fig. 1.2**). The area of rectangle A is proportional to the probability of event A, denoted $p(A)$ and the area of rectangle B is proportional to the probability of event B, denoted $p(B)$. Returning to the basketball example, there are vastly more tall people than professional basketball players, so the area of the blue rectangle is much greater than the area of the green rectangle. The overlap area occupies nearly all of the green rectangle area, showing that the probability of being tall, given that one is a professional basketball player, is nearly 100%. However, the overlap area is much less than the blue rectangle area, showing that the probability of being a professional basketball player, given that one is tall, is very small. The distinction between $p(A|B)$ and $p(B|A)$ is apparent in a multitude of real-world examples.

In perception, what is given are the sensory inputs or observations that are directly available to the observer, for example the activation pattern of photoreceptors in the retina. Given these inputs (A), the observer would like to infer the current state of the world (B). In order to do so, the observer evaluates the probability that the world is in one possible state or another. Since the observer does not know the true state of the world, B is a hypothesis that the observer is entertaining, and we refer to B as the hypothesized world state. For example, the observer might want to know how probable it is that a floor is wet (hypothesized world state B), given that the floor is shiny (observation A). The conditional probability of interest to the observer is $p(B|A)$, the probability of a hypothesized world state given the sensory observations.

Depending on the situation, the observer may be concerned with evaluating the conditional probabilities of just two hypothesized world states (the floor is wet or dry), multiple distinct world states (the animal on the path ahead is a dog, a cat, a rabbit, a raccoon, or a skunk), or even a

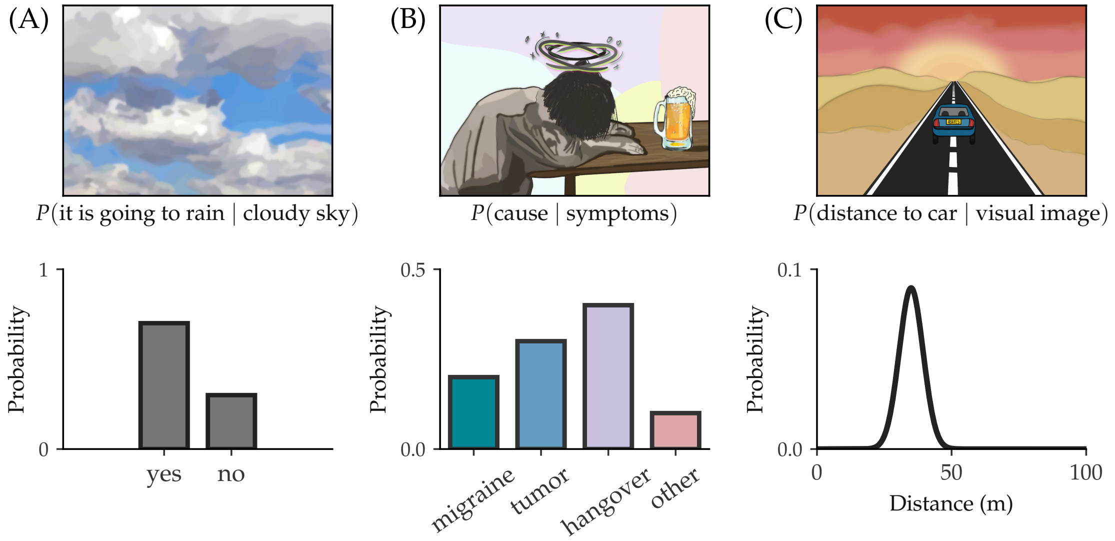


Figure 1.3: Different kinds of probability distributions. **(A)** Two hypotheses. **(B)** Multiple hypotheses. **(C)** Continuous (infinitely many) hypotheses.

continuum (infinite number) of world states. Ultimately, we would like to express the results of our inference by calculating the probability of each world state, given the observation (**Fig. 1.3**). This would allow us to make an informed decision about the world.

Suppose you see a person in the distance who appears to be walking towards you, and you wonder whether they are your friend (**Fig. 1.4A**). Whatever conclusion you reach, you will have some degree of confidence, and your degree of confidence may change over time as you continue to view the scene. We perceive visual scenes with little conscious effort, but scene recognition is in fact a computationally challenging endeavour, and the processing that the brain engages in is remarkably sophisticated.

Scene recognition, like all forms of perception, is challenging because the sensory input captured by the nervous system (the visual image in this case) is typically compatible with multiple interpretations. The visual image could be that of your friend or of another person. The image may provide sufficient information to recognize that the object is, in fact, a person, and it may provide information regarding the approximate shape (height, girth, etc.) of the person. Over time, the moving image may additionally provide information about the person's gait. Nevertheless, the person is far away, and your visual observations are compatible with many possible individuals.

The *likelihood* of a hypothesis is the probability of the observation given the hypothesized world state, $p(\text{observation}|\text{hypothesized world state})$. A plot of the likelihoods for all the hypotheses summarizes how well a visual image allows you to distinguish one possible world state from another (e.g., friend or stranger, **Fig. 1.4**). This plot is known as the *likelihood function*.

In general, many factors affect the likelihood function. In our current example, the person's height (your friend is tall) hair color (brown), and way of holding the head (tilted) would each affect the likelihoods of the hypotheses. We will leave aside at present how we might arrive at the exact form of the likelihood function. For now, it is sufficient to understand that the likelihood function represents the full information content of the image relevant to the question at hand (is that my friend?). Specifically, it represents the probability that your friend would give rise to the visual image you currently sense, compared to the probability that another person would give rise to the same visual image.

Although the likelihood function is a crucial component to our inference process, it alone is

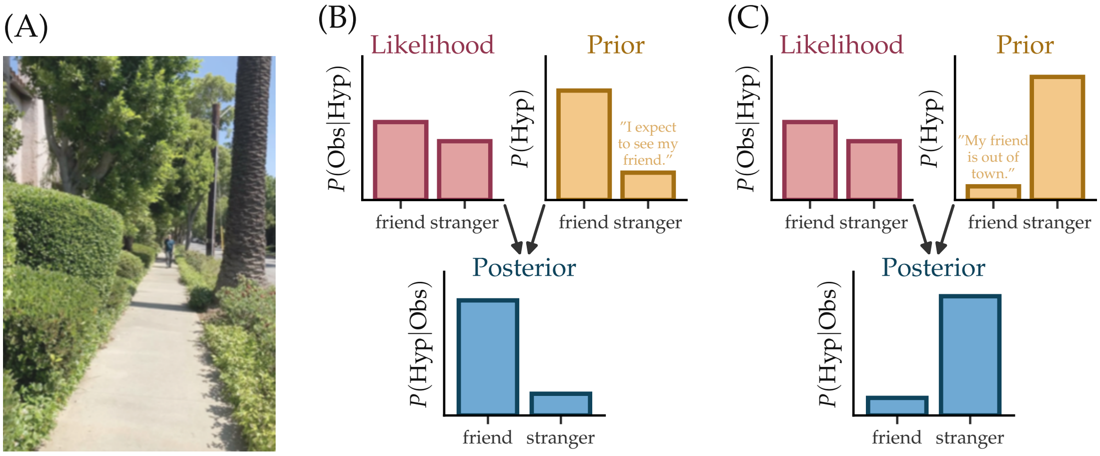


Figure 1.4: Recognizing a friend. **(A)** A visual scene offers a low-resolution view of a person in the distance who resembles your friend. **(B)** You consider the probability that the visual image would result from your friend to be greater than the probability that it would result from a stranger (likelihood function). You expected (prior distribution) to meet your friend at this time and place. Therefore, you believe the person in question is probably your friend (posterior distribution). **(C)** In this alternate scenario, you thought your friend was out of town, so your prior distribution sharply favors the *stranger* hypothesis. Given the same observation (likelihood function), you conclude that the person in question is probably not your friend.

not sufficient to solve the problems we want to solve. The likelihood function plots the probability of the observation given each hypothesized world state: $p(\text{observation}|\text{world state})$. What we want to know is the posterior distribution: the probability of each possible world state, given the observation: $p(\text{world state}|\text{observation})$. To determine the posterior distribution, we combine the likelihood function with a *prior distribution* that plots the *prior probability* of each world state. The prior probability of a hypothesized world state, denoted $p(\text{world state})$, is the probability of the world state based on all knowledge that you have apart from the observation – for instance, your belief that your friend would be present, before you even look up the street.

Let's consider two different scenarios that would cause you to have different prior distributions.

- Scenario 1: You had arranged to meet your friend on the street shown, and at the time shown, when you see the person walking towards you who looks like your friend.
- Scenario 2: When you see the person walking towards you who looks like your friend, you are surprised, because you thought your friend was still away on vacation and not planning to return to town until the following week.

The sensory input is identical in the two scenarios (**Fig. 1.4A**), but your perceptual inference would differ dramatically. Under Scenario 1, $p(\text{Friend})$ was high, and you would conclude that the person walking towards you is probably your friend; under Scenario 2, $p(\text{Friend})$ is low, and you would conclude that the person is probably not your friend. Clearly, your prior probabilities play a crucial role in your perceptual inference process.

Bayes' rule, a fundamental theorem in probability theory, shows how to optimally combine expectation, represented by the prior distribution, with the observation, represented by the likelihood function, in order to calculate a posterior distribution (**Fig. 1.4B-C**). The posterior probability of each world state is your belief in that world state based on all relevant information at your disposal. Bayes' rule states that the posterior probability is proportional to the product of the prior and the

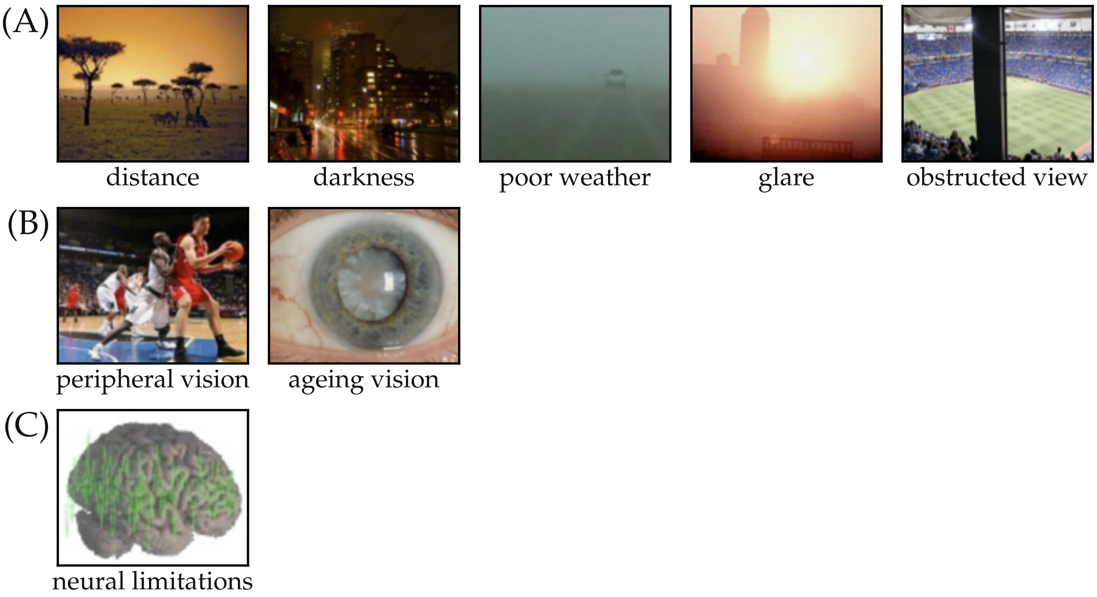


Figure 1.5: Sources of sensory degradation that reduce the quality of visual inputs, causing likelihood functions to flatten. **(A)** Physical features of the environment. **(B)** Limitations of the observer's sensory organs. Most of the factors shown in A and B have analogs in the other senses. For example, in the case of audition, distance, soft speech, ambient noise, and ageing ears all result in low-quality inputs. **(C)** The observer's nervous system. In every sensory system, neural limitations such as faulty background knowledge and neural noise also pose a challenge to perception; *figure inclusion pending permissions*.

likelihood:

$$\text{Posterior} = \text{constant} \cdot \text{Prior} \cdot \text{Likelihood} \tag{1.1}$$

The posterior is based on all knowledge we have (i.e., our current sensory observation and relevant prior knowledge). When we start using actual numbers in Chapter 2, we will discuss the constant in this equation, but for now, this equation captures the relevant intuitions. When both the prior and the likelihood are higher for Hypothesis A than for Hypothesis B, then the posterior will also be higher for Hypothesis A. However, if the prior favors Hypothesis A and the likelihood Hypothesis B, then the posterior could go either way, depending on the exact numbers. For example, if the prior only mildly favors Hypothesis A, but the likelihood strongly favors Hypothesis B, then the posterior will tilt in favour of Hypothesis B. Prior and likelihood can have equivalent influence on the final posterior.

1.3 Sensory noise and perceptual ambiguity

In any perceptual system, the flatter or broader the posterior density, the more ambiguous – i.e., open to multiple interpretations – the observations are. As we've seen, the shape of the posterior density results from the shapes of the likelihood function and prior density. Perceptual ambiguity, therefore, can result from a broad likelihood function (in the absence of a countervailing sharp prior) or vice versa.

Many factors can reduce the quality of sensory data and thereby broaden the likelihood function. These include physical features of the environment, limitations of the observer's sensory organs, and limitations in the nervous system (**Fig. 1.5**). One ubiquitous likelihood function broadening factor

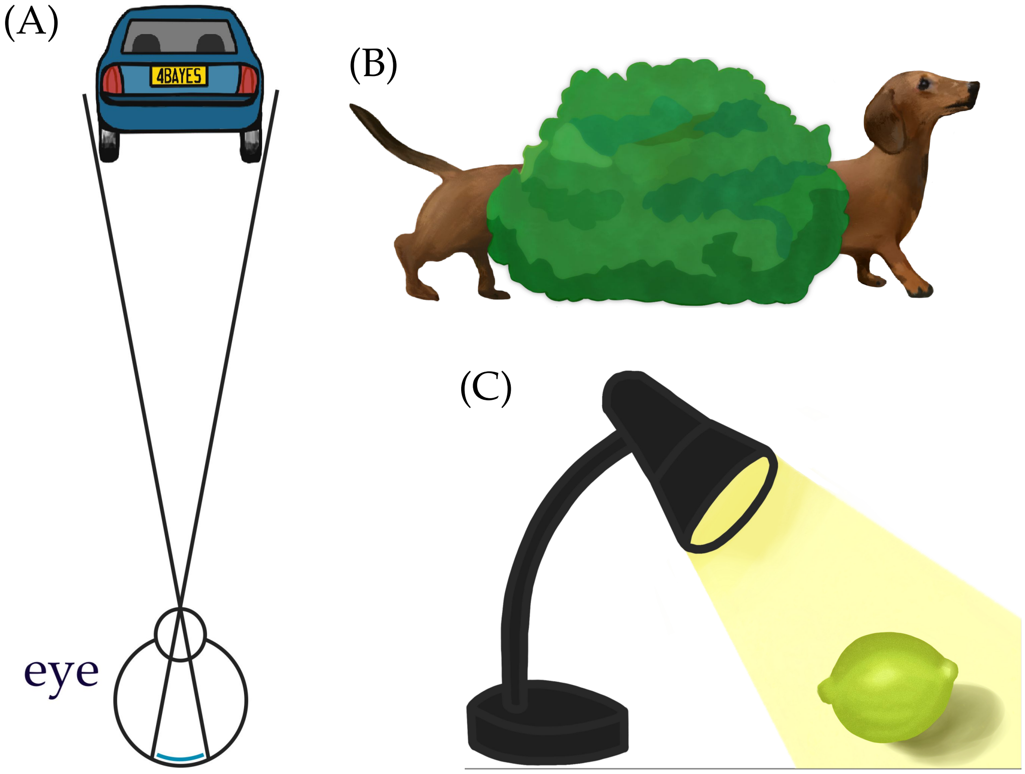


Figure 1.6: Likelihood functions broadening caused by factors other than sensory noise. **(A)** The same retinal image size can result from a small object closer to the observer or a larger object farther away. **(B)** A bush prevents the observer from knowing whether this scene contains two (or even more) dogs or just one longer dog. **(C)** The observer can't be sure of the shade of a surface without also knowing the intensity of the illuminating light.

is sensory noise. By sensory noise, we mean stochastic variability inherent to a physical process that generates a sensory observation. Because of noise, a stimulus that is repeated identically over multiple trials typically produces a somewhat different sensory observation each time. The scattering of light, random variability in ambient sounds, or biophysical variability in the firing rates of sensory neurons are all examples of sensory noise. As these examples indicate, sensory noise can occur both in the external world and within the observer.

While sensory noise is ubiquitous, it is not the only cause of broad likelihood functions. Even if all sensory input could somehow be made noiseless, many visual likelihood functions would be broad for purely geometric and optical reasons. (**Fig. 1.6**). For instance, information is necessarily lost when the three-dimension visual world maps onto a two-dimensional retinal image (or when the three-dimensional auditory world maps onto the two ears). This collapsing across a dimension gives rise to many instances of ambiguity, including size-distance ambiguity. Another common geometric source of broad likelihood functions is occlusion, in which an object partially obstructs the observer's view, such that the scene is compatible with a variety of alternative configurations of objects. As a final example, consider the apparently simple task of perceiving the shade of a grey surface from the intensity of light that reflects off the surface and enters your eyes. A particular light intensity entering your eyes is consistent with multiple combinations of the true shade of the surface and the intensity of the illuminating light source. For example, the same intensity of light entering the eyes can be produced by dark paper in sunlight or white paper in dim light. Therefore, unless the intensity of the illuminating light is known, the likelihood function in this scenario is broad. We consider these and other examples mathematically in later chapters.

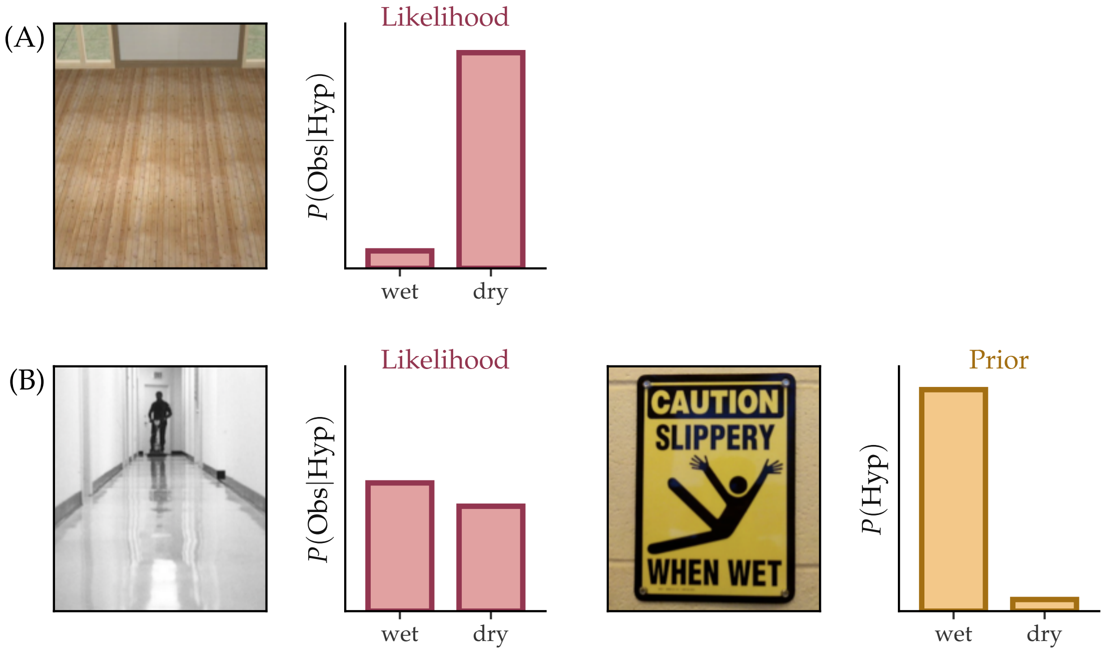


Figure 1.7: Perception of wetness. **(A)**. The likelihood function resulting from the visual image of this wood floor favors the "dry" world state: $p(\text{observation} \mid \text{dry}) \gg (\text{observation} \mid \text{wet})$. **(B)**. The shiny floor results in a likelihood function that favors the "wet" world state: $p(\text{observation} \mid \text{wet}) > p(\text{observation} \mid \text{dry})$. A "caution: slippery when wet" sign would result in a sharper prior in favor of the "wet" world state; *figure inclusion pending permissions*.

The perceptual ambiguity resulting from noise or other likelihood-broadening factors can be lessened or prevented altogether if the observer has a sharp prior distribution. Prior probabilities are based on background knowledge and therefore can evolve over time as an observer acquires new knowledge. Priors also tend to differ from one observer to another. In general, those with greater relevant knowledge have more realistic priors, facilitating accurate perception. Consider **Fig. 1.6B**. What prior knowledge about dogs might help one observer experience less ambiguity in this situation than another observer who lacks that knowledge?

1.4 Bayesian inference in visual perception

As perceptual inference is so important, we want to illustrate it with more examples drawn from everyday life. Our goal is for the reader to develop an intuitive understanding of likelihoods, priors, and posteriors, and an appreciation for the remarkable explanatory power of Bayesian inference as a model of perception. We will see that each example has unique features, yet each is based upon the joining of a likelihood function and a prior distribution via Bayes' rule, in order to generate a posterior perceptual inference. We hope that these examples begin to reveal both the richness of perceptual inference and the wide applicability of the Bayesian perceptual framework.

Slippery when wet

As humans move through the world, we rely on our senses to avoid hazards. In the modern world, hazards come in many forms, for instance an object in our path, a rapidly approaching car, or a downward step such as a curb. Another hazard of modern life is the wet floor. Is the floor wet (**Fig. 1.7**)? If it is – or might be – caution is warranted, and we may want to take small, careful steps. If it is not, we can safely proceed with long, purposeful strides. An important question is how perception can distinguish the two cases.

Many floors are both shiny and wet, whereas many other floors are neither. Observing a shiny surface thus results in a relatively sharp likelihood function. It is important to understand that not only our priors but also our likelihoods depend on background knowledge. In general, to determine likelihoods, the observer needs to have an (implicit) understanding of the process by which different world states generate sensory data. In this case, the observer needs an intuitive understanding of optics, i.e., that a wet floor tends to reflect light to a greater degree. To recognize the dependence of the likelihood on the observer's background knowledge, we sometimes write the likelihood as $p(\text{observation}|\text{world state}, B)$, where B signifies background information obtained through previous experience. This makes explicit that likelihood functions depend on background knowledge.

As explained above, the brain has to combine likelihoods, $p(\text{observation}|\text{hypothesized world state})$, with prior probabilities, $p(\text{hypothesized world state})$, to generate the probabilities it most wants to know: $p(\text{hypothesized world state}|\text{observation})$. These latter probabilities are called *posterior probabilities* to indicate that, unlike prior probabilities, they are formed *after* the observation. The *posterior probability distribution* represents the brain's belief in each possible world state, based on all relevant information (i.e., observation and expectation). We need to calculate the posterior probability of each hypothesized world state, $p(\text{world state}|\text{observation})$. This calculation involves multiplying priors and likelihoods. Recall that the prior probability, $p(\text{wet})$, reflects the observer's expectation regarding the slipperiness of the floor, independently of the visual observation. For instance, before even entering a room, what probability would the observer assign to the hypothesis that the floor will be wet? How does the observer acquire such priors?

The background knowledge that informs priors may have been acquired over a lifetime of previous experience, or very recently. An observer who is entering an unfamiliar building will have a prior that favors dry, because floors in the observer's experience are dry the majority of the time. However, if the observer sees a newly posted *caution: slippery when wet* sign, their prior will change to favor the wet hypothesis. To recognize the dependence of the prior on the observer's background knowledge, we sometimes write the prior as $p(\text{world state}|B)$, where B again signifies information obtained through previous experience.

Since both prior and likelihood depend upon background knowledge, the posterior, too, depends on background knowledge. To recognize this dependency, we sometimes write the posterior probability of each world state as $p(\text{world state}|\text{observation}, B)$.

Camouflage

In the animal kingdom, survival often depends on seeing but not being seen. As noted previously, a sharp likelihood function indicates that an observation is highly informative, whereas a flatter likelihood function provides little information. In general, then, it benefits an animal's survival to have keen senses that produce sharp likelihood functions when it views the world, but at the same time to engage in behaviours or have physical features that produce relatively flat likelihood functions in other species.

Indeed, many species have evolved traits and behaviors that serve to disguise their presence or their identity (**Fig. 1.9**). These diverse examples of camouflage and mimicry in the animal kingdom can be understood as evolved strategies aimed at flattening the likelihood functions of other species. Consider, for instance, the peppered moth caterpillar. Remarkably, individuals of this species assume the color of the tree bark on which they live (**Fig. 1.9A**). By blending in with the background, these caterpillars protect themselves from predatory birds. The visual image observed by a bird provides scant indication of the caterpillar's presence.

Camouflage is not exclusive to prey; predators, too, benefit from it. Consider the image of a lioness as she lies in waiting for her prey. Crouching low in the high golden grass, whose color closely resembles her own, she is nearly invisible until the moment she strikes. Although they can run fast, lions and other large cats lack stamina for long chases. Their success in hunting depends

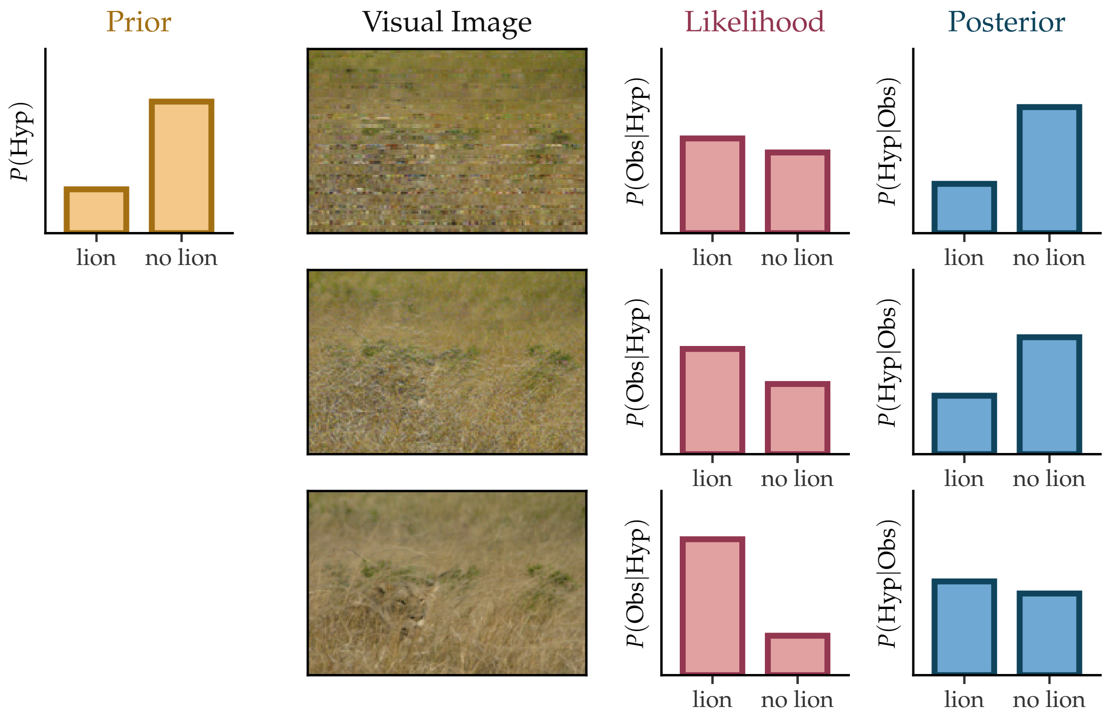


Figure 1.8: Effect of visual acuity on the posterior distribution. Three prey who hold identical (20%) prior expectation for the presence of a lion (upper left) differ in visual acuity, and therefore experience different likelihood functions when confronted with the same visual scene. The flatter the likelihood function, the more the posterior distribution resembles the prior distribution. **Top**: To this animal with poor visual acuity, the visual scene evokes a nearly flat likelihood function. The animal's posterior distribution is therefore similar to its prior distribution; it has learned little from the visual observation. **Middle:** An animal with intermediate visual acuity has a likelihood function that is not flat. This animal's posterior distribution differs slightly from its prior distribution. **Bottom:** For this animal with excellent visual acuity, the scene results in a sharp likelihood function in favor of the lion's presence. The animal's posterior distribution indicates slightly greater than 50% probability that a lion is present; *figure inclusion pending permissions*.

on their ability to approach prey unnoticed. Examples of camouflaged predators and prey abound in the animal kingdom.

As long as an observer's likelihood function is not perfectly flat, the observer will learn something from the sensory input However, when a well-camouflaged animal is viewed, the observer's likelihood function is nearly uninformative. Importantly, the shape of the likelihood function depends not exclusively on the visual scene but also on the sensory acuity and acumen of the observer. A lion that to one observer is nearly perfectly camouflaged may be noticed by another observer who has better visual acuity (**Fig. 1.8**). To an observer who knows from experience that peppered moth caterpillars tend to be slightly wider than the twigs of the tree they inhabit, the same visual scene (**Fig. 1.9A**) will result in a sharper likelihood function than it does for an observer lacking this background knowledge.

Indeed, along with camouflage, evolution has given rise to sophisticated sensory systems – and cognitive abilities – that function to reduce uncertainty about the presence and locations of other animals. In an arms race of sorts, animals have evolved progressively more sophisticated sensory systems to detect their progressively better-hidden opponents. The evolution of mammalian

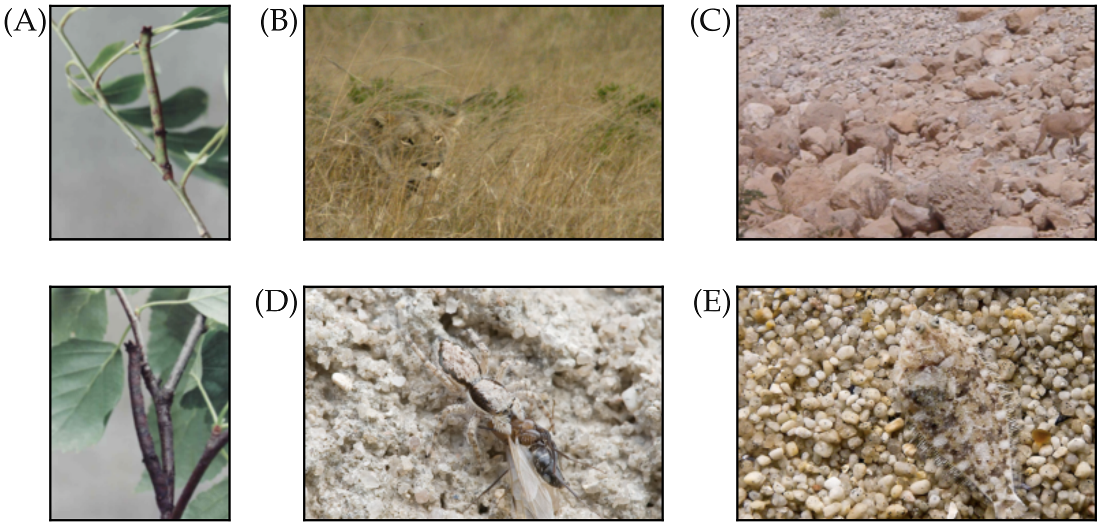


Figure 1.9: Likelihood function-flattening features in the animal kingdom. **(A)** The peppered moth (Biston betularia) caterpillar changes its color to blend in with the background (above: willow; below: birch); from Noor et al., 2008. **(B)** In tall golden grass of Kenya's Masai Mara National Reserve, a well-camouflaged lioness lies in waiting for wildebeest prey. **(C)** Ibex in the Israeli desert. **(D)** A well-camouflaged jumping spider with its captured ant prey (Dar es Salaam, Tanzania). **(E)** A flounder against the sea floor; *figure inclusion pending permissions*.

visual, auditory and olfactory systems are cases in point, as is the evolution of highly specialized detection systems such as the ultrasonic echolocation used by insect-eating bat species. In general, animals benefit if they perceive others sharply while others have difficulty perceiving them. Animals have therefore evolved impressive perceptual systems to achieve sharper likelihood functions for themselves, while at the same time evolving camouflage to force flatter likelihood functions upon others.

1.5 Bayesian inference in auditory perception

So far, we have considered visual examples. However, nothing about inference is specific to vision. In this section and the next, we consider audition. Humans live in an acoustically rich environment: birds chirp, the wind howls, dogs bark, car horns blare, music plays, and, perhaps most importantly, we talk to one another. Whether we are identifying the source of a sound (is that a dog barking?), perceiving its location (where is that barking dog?), or interpreting its meaning (what was that word you just said?), we use perceptual inference, combining likelihoods and priors to generate posterior probabilities.

Birds on a wire

Humans often rely at least in part on our sense of hearing to locate objects. We and other mammals localize sounds sources by using sophisticated yet unconscious calculations, including comparing the intensity and time of arrival of sounds at the two ears. Nevertheless, our ability to localize sounds is not perfect, and consequently we combine prior probabilities with our acoustic likelihoods to reach the most precise perceptual inference we can.

Suppose that you are walking outside on a beautiful sunny morning, when you notice the silhouettes of 5 birds perched on a wire (**Fig. 1.10A**). Suddenly, one of the birds (you cannot see which) bursts into melodious song. Which bird sang? Your auditory system rapidly processes the

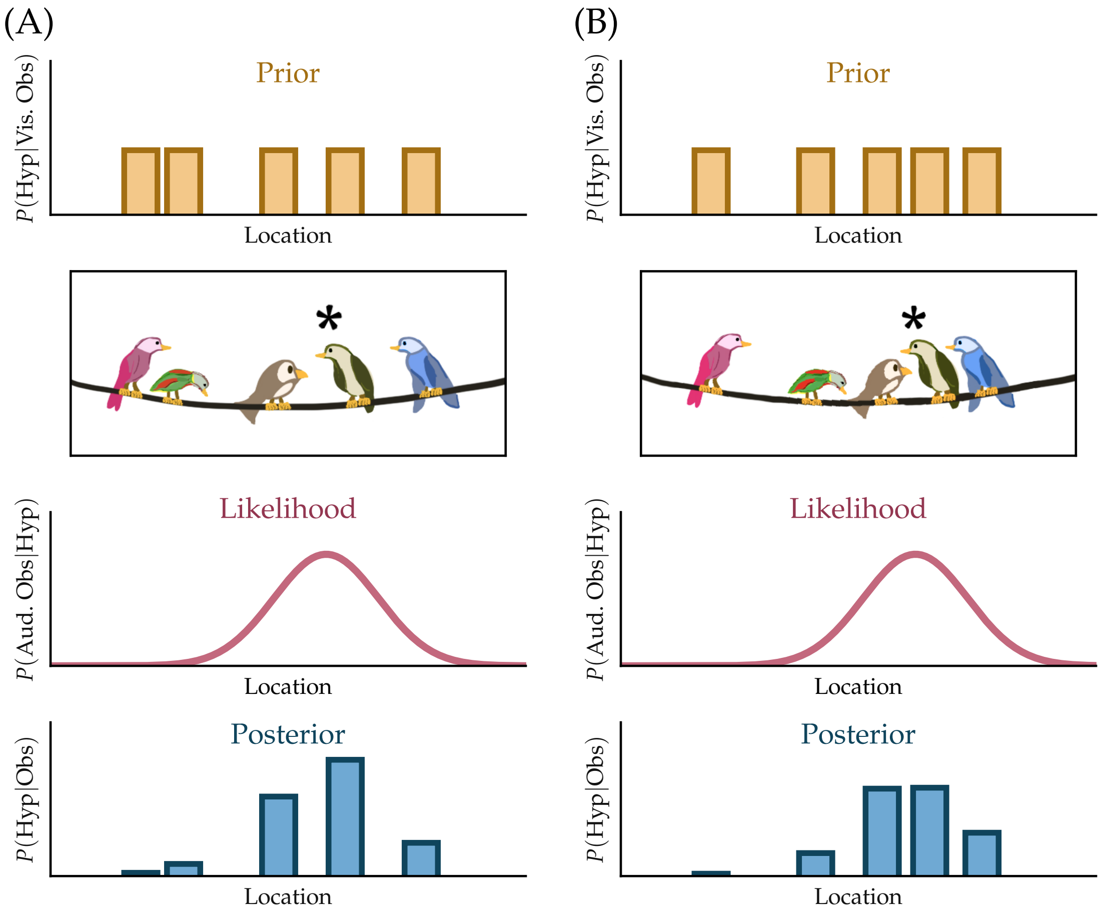


Figure 1.10: Sound source localization. **(A)** The visual image of the birds provides the basis for a prior distribution over sound source location. The broad likelihood function reflects the imprecision of the acoustic observation. The posterior distribution favors the hypothesis that the 4th bird from the left sang (*). **(B)** Perceptual uncertainty increases markedly if the birds crowd closer together.

acoustic observation, yielding a broad likelihood function. This likelihood function is a continuous function over location; that is, the sound you heard is compatible with a source along a continuum of locations. Nevertheless, certain locations are associated with higher likelihoods than others. Interestingly, the location of highest likelihood may not coincide with the exact location of any bird. This situation is common in acoustic perception and can be caused by many factors. For instance, if the bird that chirped was not facing you directly, then the sound it produced may have deflected off nearby objects before reaching your ears. Even if sound waves were able to reach your ears without deflection, stochastic variability in the response of your nervous system can cause the likelihood function to peak at a location that is slightly offset from the source location.

Unlike the likelihood function, the visual information in the present example is not continuous, but rather discrete. You see five individual birds. Your visual observation, which occurred before the bird chirped, provides you with a prior distribution. Thus, the prior distribution is nonzero at five discrete locations (we are assuming that your visual perception is highly accurate for this high-contrast scene). Note that the prior probabilities are taken to be equal across the five birds, and the prior probability that the sound source would occupy an empty location on the wire is zero. This simply means that, prior to hearing the song, you considered it equally probable that any one of the birds would sing. Using Bayes' rule, we can now calculate the posterior distribution for the location of the sound source. For each of the five hypothesized locations, we multiply the likelihood by the prior. The resulting posterior distribution indicates the 4th bird from the left as the most likely source of the pleasing melody.

Intuition suggests that if the birds had been closer together on the wire, our inference would be less certain. This result indeed emerges from the Bayesian inference, as shown in **Fig. 1.10B**. Here, we show the singing bird at the same location, but with three of the other birds closer to it than they were before. Our prior distribution reflects the new positions of the birds, but the acoustic observation and therefore the likelihood function are the same as before. The posterior distribution is now broader and lower, indicating that, although the same bird is the most probable singer, our uncertainty has markedly increased. Indeed, in our judgment the singing bird could nearly equally probably be the 3rd or the 4th bird from the left.

Before leaving this example, we would like to draw the reader's attention to two alternative approaches to solving the problem that would have led to the same answer. In one alternative approach, we could have started, before looking at the wire, with a flat prior over hypothesized bird locations, reflecting the fact that, before looking, we had no idea as to where any birds would be perched. We could then have incorporated the subsequent visual observation into a likelihood function and combined this with our flat prior distribution to produce a posterior distribution over the birds' locations. Indeed, it was this original *posterior* distribution from the visual input that we used here as a *prior* distribution for our analysis of the auditory observation. This illustrates a general important feature of Bayesian inference: it can be done iteratively, the posterior distribution from one inference being used as the prior distribution for the next.

We will learn about a second alternative approach to this problem in Chapter 5, which again would reach the same answer: Starting with a flat prior over position, we could incorporate simultaneously both the visual and the acoustic observations as likelihood functions, in a procedure known as *cue combination*. In this approach, we would not use the visual information to generate a prior distribution for the subsequent auditory observation but would instead combine a visual likelihood function that has five discrete peaks with a continuous acoustic likelihood function. In essence, when we have two or more independent sources of information, we can choose whether to incorporate the different sources sequentially, with the posterior from each observation being used as the prior for the next, or all at once, with all the observations entering through likelihood functions. Thus, there is often a blurring of boundaries between likelihood functions and priors, with the choice of how to incorporate the information left up to the Bayesian modeler. This flexibility is not a problem but a benefit of Bayesian inference. The internal consistently of the rules of Bayesian inference ensures that, as long as all the information is incorporated, the resulting posterior distribution will be the same regardless of the route taken. Within the Bayesian framework, there are often multiple ways of arriving at the same, useful solution.

Mondegreens

Although our brains do it automatically and apparently effortlessly, speech perception requires sophisticated inference on a variety of levels. Most obviously, we must correctly perceive the spoken word. It is easy to misinterpret even a single word spoken in isolation, particularly when one is in the presence of ambient noise (the drone of a car engine, street sounds, chatter from nearby people speaking, and so on). Under such conditions, akin to low-contrast vision, likelihood functions are broad, and one word may be misperceived for another that sounds similar. In fact, misperceived speech is such a common occurrence that humans often pass by these moments without a second thought. We urge the reader to keep a list of such occurrences. The results are both educational and amusing. For instance, in conversations with others, we have misheard *Mongolia* as *magnolia*, *fumaroles* as *funerals*, *hogs* as *hawks*, *census* as *senses*, *a moth* as *I'm off*, *maple leaf* as *make believe*, and *peaches and strawberries too* as *peaches and strawberries stew*. As these examples illustrate, in addition to the similar sounds of different words, a challenge to

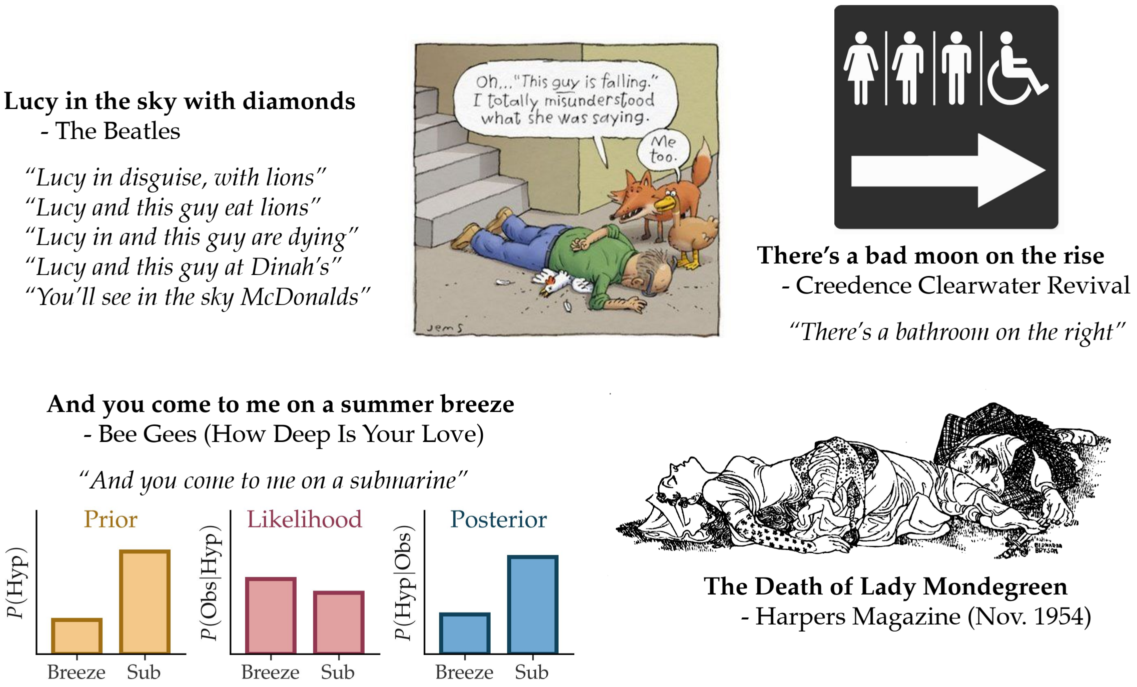

Lucy in the sky with diamonds
- The Beatles
"Lucy in disguise, with lions"
"Lucy and this guy eat lions"
"Lucy in and this guy are dying"
"Lucy and this guy at Dinah's"
"You'll see in the sky McDonalds"
Oh..."This guy is falling." I totally misunderstood what she was saying.
Me too.
Jems
There's a bad moon on the rise
- Creedence Clearwater Revival
"There's a bathroom on the right"
And you come to me on a summer breeze
- Bee Gees (How Deep Is Your Love)
"And you come to me on a submarine"
Prior
Likelihood
Posterior
P(Hyp)
P(Obs|Hyp)
P(Hyp|Obs)
Breeze
Sub
Breeze
Sub
Breeze
Sub
The Death of Lady Mondegreen
- Harpers Magazine (Nov. 1954)

Figure 1.11: Mondegreens result from phonetic ambiguity (broad likelihood functions) coupled with low expectation for the actual phrase that was sung or spoken (prior distribution in favor of the "wrong" hypothesis); *figure inclusion pending permissions.*

speech perception arises because the pauses between spoken words are often no longer than the pauses between syllables within a single word. Consequently, it is by no means a trivial task to infer where one word ends and the next begins. This difficulty can lead to errors in which syllables from different words combine improperly in our perception.

As a child, the author Sylvia Wright enjoyed listening to the popular 17th-century Scottish ballad, The Bonny Earl o'Moray, spoken to her frequently by her mother. She was particularly fond of the sad but beautiful lines describing the murders of the Earl and the love of his life, Lady Mondegreen:

> Ye Highlands and ye Lowlands,
>
> Oh, where hae ye been?
>
> They hae slain the Earl o'Moray,
>
> And Lady Mondegreen.

As impactful as they were, the words heard by the young Sylvia Wright were not those that her mother spoke. In fact, the ballad makes no mention whatsoever of a Lady Mondegreen. The unfortunate dead Earl was placed on the grass, alone; they "laid him on the green." Sylvia Wright's creative but mistaken interpretation of the spoken ballad reflects a perceptual parsing error. She interpreted the sounds "laid hi-" as "lady," and "-m on the green" as "Mondegreen." Sylvia Wright later coined the term "mondegreen" to refer to a misheard word or phrase. Given the inherent phonetic ambiguity of spoken language, examples of mondegreens abound. When Queensland, Australia was inundated by tropical cyclone Tasha, the Morning Bulletin of Rockhampton (Jan 6, 2011) reported the tragic news that, as a result of the flooding, "More than 30,000 pigs have been floating down the Dawson River since last weekend." This startling story, based on an interview between the reporter and the owner of a local piggery, was staggeringly incorrect. The owner had

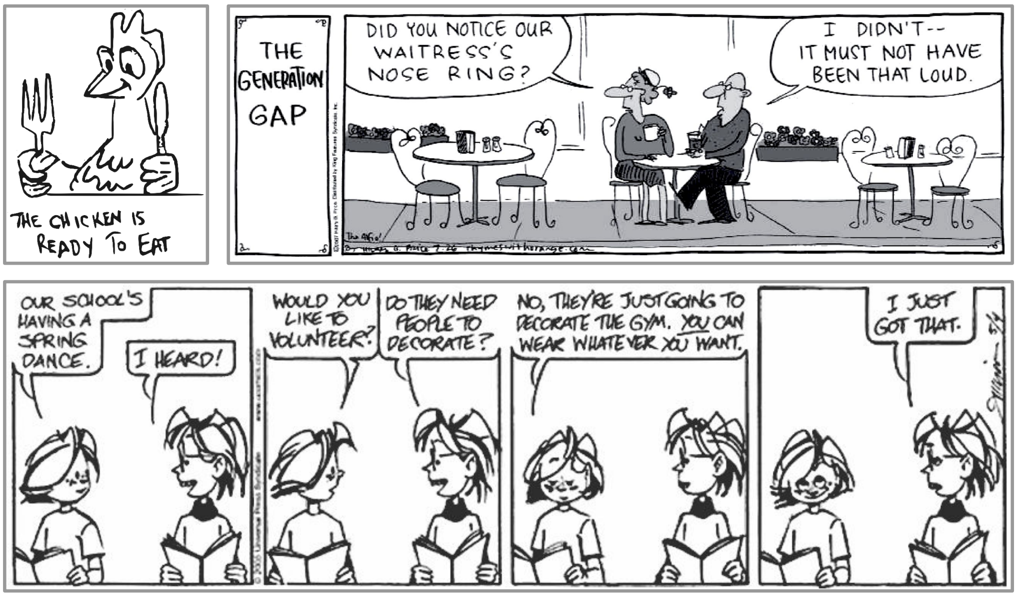

THE CHICKEN IS READY TO EAT
THE GENERATION GAP
DID YOU NOTICE OUR WAITRESS'S NOSE RING?
I DIDN'T-- IT MUST NOT HAVE BEEN THAT LOUD.
OUR SCHOOL'S HAVING A SPRING DANCE.
I HEARD!
WOULD YOU LIKE TO VOLUNTEER?
DO THEY NEED PEOPLE TO DECORATE?
NO, THEY'RE JUST GOING TO DECORATE THE GYM. YOU CAN WEAR WHATEVER YOU WANT.
I JUST GOT THAT.

Figure 1.12: Syntactic ambiguity in language; *figure inclusion pending permissions.*

spoken, not of “30,000 pigs,” but of “30 sows and pigs” swept downstream! The Morning Bulletin published a correction the following day.

Books and many websites are devoted to listing peoples’ favorite mondegreens, particularly those resulting from misheard song lyrics, which we can all enjoy. It is instructive to visit websites on which listeners post their particular misheard versions of the same songs. The many different misheard versions of a line such as “Lucy in the sky with diamonds” presumably reflect both the phonetic ambiguity (broad likelihood function) and improbable content (low prior probability) of the original lyrics. With respect to prior probabilities, “There’s a bathroom on the right” is surely a more common sentence than “There’s a bad moon on the rise”, and “submarine” is arguably more plausible than “summer breeze” as a mode of transport (**Fig. 1.11**).

The occurrence of mondegreens suggests strongly that speech perception, like visual perception, results from the combination of likelihood functions and prior distributions. Humans generally perceive speech accurately, but of course occasional mistakes are inevitable. Indeed, “The more unintelligible the original lyrics, the more likely it is that listeners will hear what they want to hear” (O’Connell, 1998). Rephrased in terms of Bayesian inference, the flatter the likelihood function, the greater will be the influence of the prior distribution on the resulting posterior distribution. Thus, the incorporation into perceptual inference of prior expectation, which in most circumstances improves perceptual accuracy, can backfire to create mondegreens on occasions when we are faced with an unexpected word (low prior) that sounds like (broad likelihood) another, more expected (high prior) word.

Keeping this in mind, it is rather easy to evoke mondegreens in others. Simply select two different words or phrases that sound alike, ensure that your listener has a prior distribution in favor of one of the words or phrases, then speak the other. You may wish to try the following demonstration with a friend. Tell the friend that “You know, humans are very good at speech recognition; in fact, we can understand speech much better than even the best computer programs can. We really know how to wreck a nice beach. Now, what did I just say? We really know how to. . . .?” If you spoke the words “wreck a nice beach” naturally, at your typical speed, and in a typical, not very clearly enunciated fashion, your friend will probably have perceived “recognize speech,” rather than the words you actually spoke. In Bayesian terms, the broad likelihood function experienced by your friend will combine with a sharp prior distribution (given the previous content

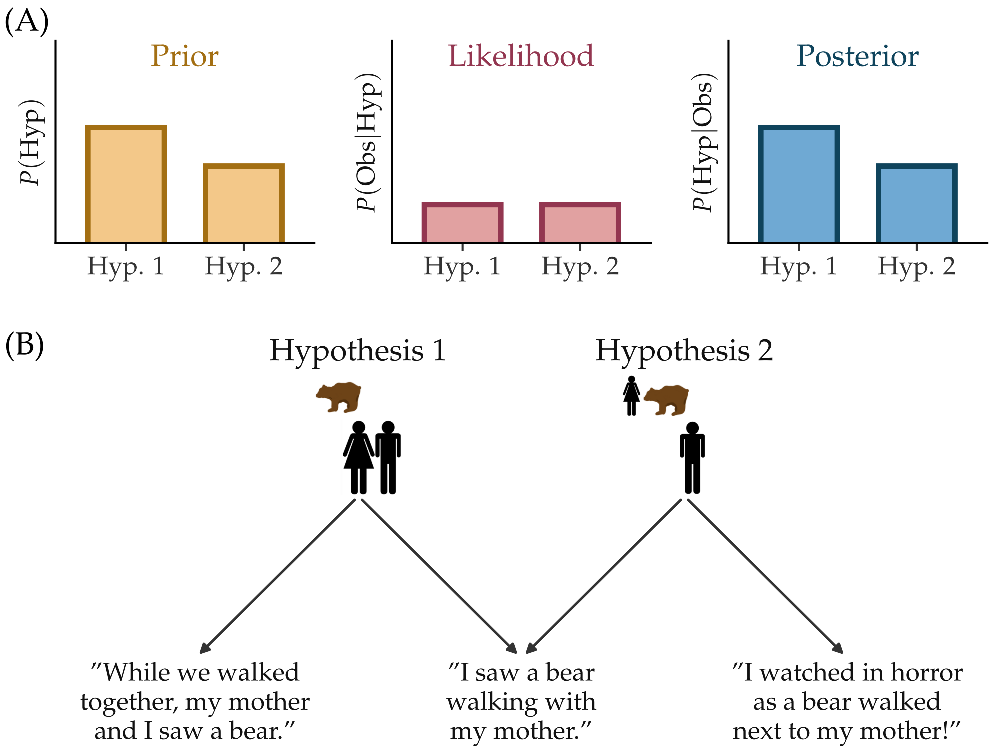

(A)
Prior
$P(\text{Hyp})$
Hyp. 1
Hyp. 2
Likelihood
$P(\text{Obs}|\text{Hyp})$
Hyp. 1
Hyp. 2
Posterior
$P(\text{Hyp}|\text{Obs})$
Hyp. 1
Hyp. 2
(B)
Hypothesis 1
Hypothesis 2
"While we walked together, my mother and I saw a bear."
"I saw a bear walking with my mother."
"I watched in horror as a bear walked next to my mother!"

Figure 1.13: Perceptual inference under syntactic ambiguity. **(A)** Each world state could be described in many different ways, a few of which are shown. The speaker happened to choose an expression that could describes both world states: "I saw a bear walking with my mother." **(B)** Bayesian perceptual inference. Background knowledge suggests that bears are less likely to walk alongside people than to be seen by them at a distance, so the prior distribution favors Hypothesis 1. The likelihood function shows that the spoken sentence has about equal probability under the two hypotheses. The posterior distribution therefore favors Hypothesis 1.

of your discourse) to favor the "recognize speech" hypothesis.

Even when the listener perceives every word correctly, she faces a final crucial challenge: to identify the intended meaning of the string of words. Once again, this often requires evaluating multiple hypotheses. Consider the sentence, "The bridge is being held up by red tape." This sentence, even when perfectly heard, is nevertheless consistent with two interpretations; that is, it evokes a broad likelihood function, due not to phonetic ambiguity but rather to the meaning of "red tape" (semantic ambiguity). In other cases, the sentence structure itself is ambiguous (syntactic ambiguity, **Fig. 1.12**). When we hear or read an ambiguous sentence, we naturally combine the likelihood functions with a prior distribution, and usually reach the correct perception. We are, however, sometimes bemused – and amused – momentarily, as both interpretations cross our minds. This occurred recently to one of the authors when a friend told him of a wilderness trip he took with his parents. "There was wildlife everywhere," he exclaimed, "In fact, I saw a bear walking with my mother" (**Fig. 1.13**). Although this book will not focus on these types of situations, we point them out to illustrate that uncertainty and inference apparently play a role at multiple levels of perceptual and cognitive processing.

Thomas Bayes, 1702 — 1761

Pierre-Simon Laplace, 1749 — 1827
"...the most important questions in life...are indeed, for the most part, only problems in probability. One may even say, strictly speaking, that almost all of our knowledge is only probable." — *Philosophical Essay on Probabilities*

Al Hazen (Ibn al-Haytham), 965 — 1040
"...familiar visible objects are perceived by sight through defining features and through previous knowledge..." — *De aspectibus*

Hermann Ludwig von Helmholtz, 1821 — 1894
"Previous experiences act in conjunction with present sensations to produce a perceptual image." — *Psychological Optics*

$$\underbrace{P(\text{Hyp}|\text{Obs})}_{\text{Posterior}} \propto \underbrace{P(\text{Hyp})}_{\text{Prior}} \cdot \underbrace{P(\text{Obs}|\text{Hyp})}_{\text{Likelihood}}$$

Figure 1.14: Luminaries in the development of Bayesian inference and the view that perception is unconscious inference. Note that any recognition of individuals in the history of science must happen with the understanding that many groups, especially women and people of color, were systemically excluded from the same opportunities; *figure inclusion pending permissions*.

1.6 Historical background: perception as unconscious inference

Bayes' rule is named after the English minister and mathematician Thomas Bayes (1702-1761), who was interested in problems of inverse probability, essentially how to calculate $p(B|A)$ when $p(A)$ and $p(A|B)$ are known. Bayes' *An Essay Towards Solving a Problem in the Doctrine of Chances*, published posthumously in 1763, introduced the foundation for the conditional probability calculus, a field of statistical reasoning now called Bayesian inference. Bayes' rule was later derived independently by the French mathematician and physicist, Pierre Simon Marquis de Laplace (1749-1827). Laplace applied the formula with great effect to problems in a wide range of disciplines. Importantly, Laplace also recognized the pervasiveness of probability, stating that "the most important questions of life…are indeed, for the most part, only problems in probability. One may even say, strictly speaking, that almost all our knowledge is only probable" (Laplace, 1995). Indeed, today Bayesian statistical inference is playing a rapidly growing role in an extraordinarily diverse set of disciplines covering nearly all fields of science and engineering: neuroscience, psychology, evolutionary and molecular biology, geology, astronomy, economics, robotics, and computer science, to name but a few.

The idea that perception is a form of unconscious inference, however, arose independently of Bayes and Laplace. Several scientists contributed to this notion. The early Arab physicist and polymath, Ibn Alhacen (965-c.1040 CE), recognized presciently that "…not everything that is perceived by sight is perceived through brute sensation; instead, many visible characteristics will be perceived through judgment. . . in conjunction with the sensation of the form that is seen." Thus, ". . . familiar visible objects are perceived by sight through defining features and through previous knowledge. . . " (Alhacen, De aspectibus, Book 2, translated by Smith, 2001). Much later, the German physician and physicist Hermann von Helmholtz (1821-1894) again expressed the idea that perception is a form of unconscious inference, stating eloquently that "Previous experiences act in conjunction with present sensations to produce a perceptual image' ' (Physiological Optics, 1867). The ideas of Alhacen and Helmholz fit beautifully with the view that perception is a form of Bayesian inference (**Fig. 1.14**). However, it is difficult to establish whether a particular form of perceptual inference is conscious or unconscious, and we will not comment on that issue in this book.

1.7 Summary and remarks

In this chapter, we have introduced the concept that perception is inherently probabilistic, and as such optimally characterized as a process of Bayesian inference. Regarding Bayesian inference, we have learned the following:

- Conditional probabilities such as $P(A|B)$ represents the probability of A given B. In Bayesian perceptual inference, A and B typically represent a world state and an observation.
- The likelihood function, $P(\text{observation}|\text{world state})$ captures the information content of the sensory observation, relevant to distinguishing one world state from another.
- The flatter the likelihood function, the less we learn from our senses. If the likelihood function is perfectly flat, then the observer has learned nothing from the observation.
- The prior distribution over world states, $P(\text{world state})$, summarizes the information content of our past observations, the background knowledge we have about the world. Perception is not based entirely on sensory observation, but also on expectation grounded in previous experience.
- Flatter prior distributions mean we know less about the potential states of the world.
- Bayes' rule calculates the posterior probability of each possible world state, $P(\text{world state }|$ $\text{observation})$, from the likelihoods and prior probabilities of the world states.
- The procedures of Bayesian inference apply equally to situations in which the hypothesized world states are discrete or in which they are continuous.
- Perceptual situations, whether in vision, audition, or other senses, are subject to various levels of uncertainty.
- Speech perception is fraught with phonetic and syntactic ambiguity, frequently giving rise to flat likelihood functions. The combination of priors and likelihoods can cause misinterpretations such as Mondegreens.

1.8 Suggested readings

Original papers

- FRS Bayes. "An essay towards solving a problem in the doctrine of chances". In: *Biometrika* 45.3-4 (1958), pages 296–315
- Peter Brugger and Susanne Brugger. "The Easter bunny in October: Is it disguised as a duck?" In: *Perceptual and motor skills* 76.2 (1993), pages 577–578
- Wilson S Geisler and Randy L Diehl. "A Bayesian approach to the evolution of perceptual and cognitive systems". In: *Cognitive Science* 27.3 (2003), pages 379–402
- Mohamed AF Noor, Robin S Parnell, and Bruce S Grant. "A reversible color polyphenism in American peppered moth (Biston betularia cognataria) caterpillars". In: *PloS one* 3.9 (2008), e3142

Books

- Gloria Cooper. *Red Tape Holds Up New Bridge, and More Flubs from the Nation's Press*. TarcherPerigee, 1987
- Gary Hatfield. "Perception as unconscious inference". In: *Perception and the physical world: Psychological and philosophical issues in perception*. Citeseer. 2002
- HV Helmholtz and JPC Southall. "Treatise on physiological optics. III. The perceptions of vision." In: (1925)
- Pierre-Simon Laplace. *Pierre-Simon Laplace philosophical essay on probabilities: translated from the fifth french edition of 1825 with notes by the translator*. Volume 13. Springer Science & Business Media, 2012
- Sharon Bertsch McGrayne. *The theory that would not die*. Yale University Press, 2011
- A Mark Smith. *Alhacen's Theory of Visual Perception: A Critical Edition, with English Translation and Commentary, of the First Three Books of Alhacen's De Aspectibus, the Medieval Latin Version of Ibn Al-Haytham's Kitab Al-Manazir*. Volume 1. American Philosophical Society, 2001

Popular press

- Burdon D (Jan. 6, 2011) Pigs float down the Dawson. The Morning Bulletin.
- O'Connell, PL. Sweet Slips Of the Ear: Mondegreens. New York Times (Aug. 9, 1998).
- Smith, R. *Milk drinkers turn to powder and other pun-ishing headlines.* Globe and Mail (Sept. 24, 2009).

1.9 Problems

Problem 1.1 Rephrase in terms of Bayesian perceptual inference the following statement written by Ibn al-Haytham approximately 1,000 years ago: "...when sight perceives a rose-red color among the flowers in some garden, it will immediately perceive that the things in which that color inheres are roses because that color is specific to roses...But this does not happen when sight perceives a myrtle-green color in the garden. For when sight perceives only the myrtle-green in the garden, it will not perceive the myrtle-green to be myrtle simply from the perception of the green, because several plants are green, and, in addition, several plants resemble myrtle in greenness and shape." (*De aspectibus, book 2*, as translated by Smith, 2001).

Problem 1.2 Why is it that we identify ourselves at the very beginning of a phone conversation, even to people we already know, but we do not do this when we meet in person? Express your answer within the framework of Bayesian perceptual inference.

Problem 1.3 When a conversation partner speaks softly, or when a conversation occurs in the presence of significant ambient noise, we sometimes cup our ears and/or look carefully at the speaker's lips. Why, in Bayesian perceptual terms, do we do this?

Problem 1.4 To explore how a noisy environment engenders uncertainty, consider the word "lunch." Suppose that you see this word written (or hear it spoken), with the letter "l" blocked out (making it unknown):_unch (e.g., by ambient auditory noise). List all source words that are compatible with what you see. Now consider the case in which both the l and the n are blocked: _u_ch. In terms of conditional probabilities relevant to perception, what is the effect of blocking out the l, n, and both?

Problem 1.5 The NATO phonetic alphabet, used by many military, maritime, and other organizations during radio communications, represents each letter with a word: A (Alpha), B (Bravo), C (Charlie), D (Delta), E (Echo), F (Foxtrot), and so on. What purpose does this serve in radio communications? Explain with respect to conditional probabilities. In particular, consider a radio communication under conditions of considerable background noise, in which the sender wishes to spell the word "FACE." Compare $p(\text{auditory signal heard by the receiver}|\text{FACE spelled by the sender})$ vs

p(auditory signal heard by the receiver|another word, such as FADE, spelled by the sender) when the sender uses the regular alphabet, and again when the sender uses the NATO phonetic alphabet.

Problem 1.6 English speakers sometimes incorrectly perceive English words when they listen to songs sung in a foreign language with which they are unfamiliar, and listeners also mistakenly perceive words in music that is played backwards. Provide a Bayesian explanation for these phenomena.

Problem 1.7 Suppose you see someone you do not know, getting only a brief look at them from a distance of about 10 meters. If you are interested in estimating this person's age, how would you proceed? What factors, including and in addition to the person's appearance, would affect your estimation? Provide a Bayesian description of your reasoning. As part of your answer, draw examples of your likelihood function, prior distribution, and resulting posterior distribution.

Problem 1.8 A research article entitled "The Easter bunny in October: Is it disguised as a duck?" explained that "Very little is known about the looks of the Easter bunny on his non-working days." To investigate, the authors showed an "ambiguous drawing of a duck/rabbit. . . to. . . 265 subjects on Easter Sunday and to 276 different subjects on a Sunday in October of the same year." The authors report: "Whereas on Easter the drawing was significantly more often recognized as a bunny, in October it was considered a bird by most subjects." The drawing shown by the authors in their study was similar to the following:

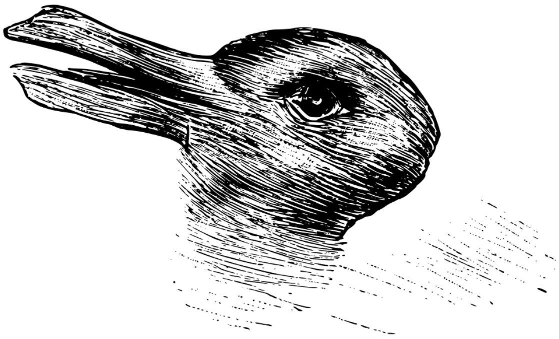


Figure inclusion pending permissions.

Provide a Bayesian perceptual explanation for the authors' results.

Problem 1.9 The images below show a Charlie Chaplin face mask. The left image is a side viewing revealing that the mask is hollow. The middle image is a front view. The right image is a back view of the hollow side of the mask:

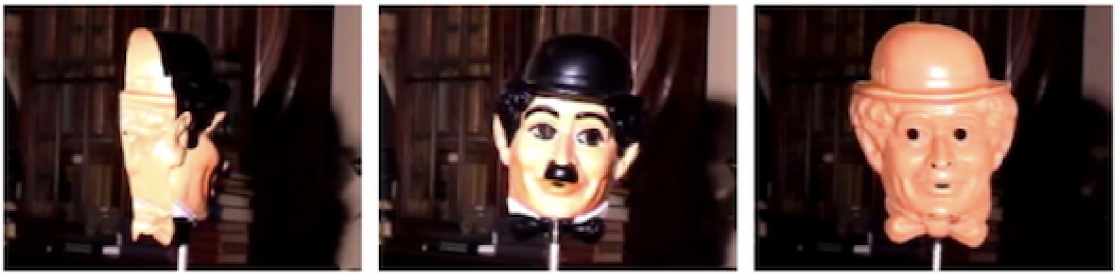


Figure inclusion pending permissions.

Provide a Bayesian explanation for why the right image looks like a normal, convex face, when in reality it is the hollow (concave) side of the mask (images from www.richardgregory.org).

Problem 1.10 Give three daily-life examples (perceptual or cognitive) in which you tried to infer a world state from incomplete or imperfect information. For each example, specify the observation(s), the world state of interest, and the source(s) of uncertainty.

Problem 1.11 Michel Treisman has tried to explain motion sickness in the context of evolution (Michel Treisman (1977), *Motion sickness: an evolutionary hypothesis*, Science 197, 493-495). During the millions of years over which the human brain evolved, accidentally eating toxic food was a real possibility, and that could cause hallucinations. Perhaps, our modern brain still uses prior probabilities genetically passed on from those days; those would not be based on our personal experience, but on our ancestors'! This is a fascinating, though only weakly tested theory. Here, we do not delve into the merits of the theory but try to cast it in Bayesian form. Suppose you are in the windowless room on a ship at sea. Your brain has two sets of sensory observations: visual observations and vestibular observations. Assume that the brain considers three scenarios for what caused these observations:

> Scenario 1: The room is not moving and your motion in the room causes both sets of observations.
>
> Scenario 2: Your motion in the room causes your visual observations whereas your motion in the room and the room's motion in the world together cause the vestibular observations.
>
> Scenario 3: You are hallucinating: your motion in the room and ingested toxins together cause both sets of observations.

Now answer the following four questions about this scenario:

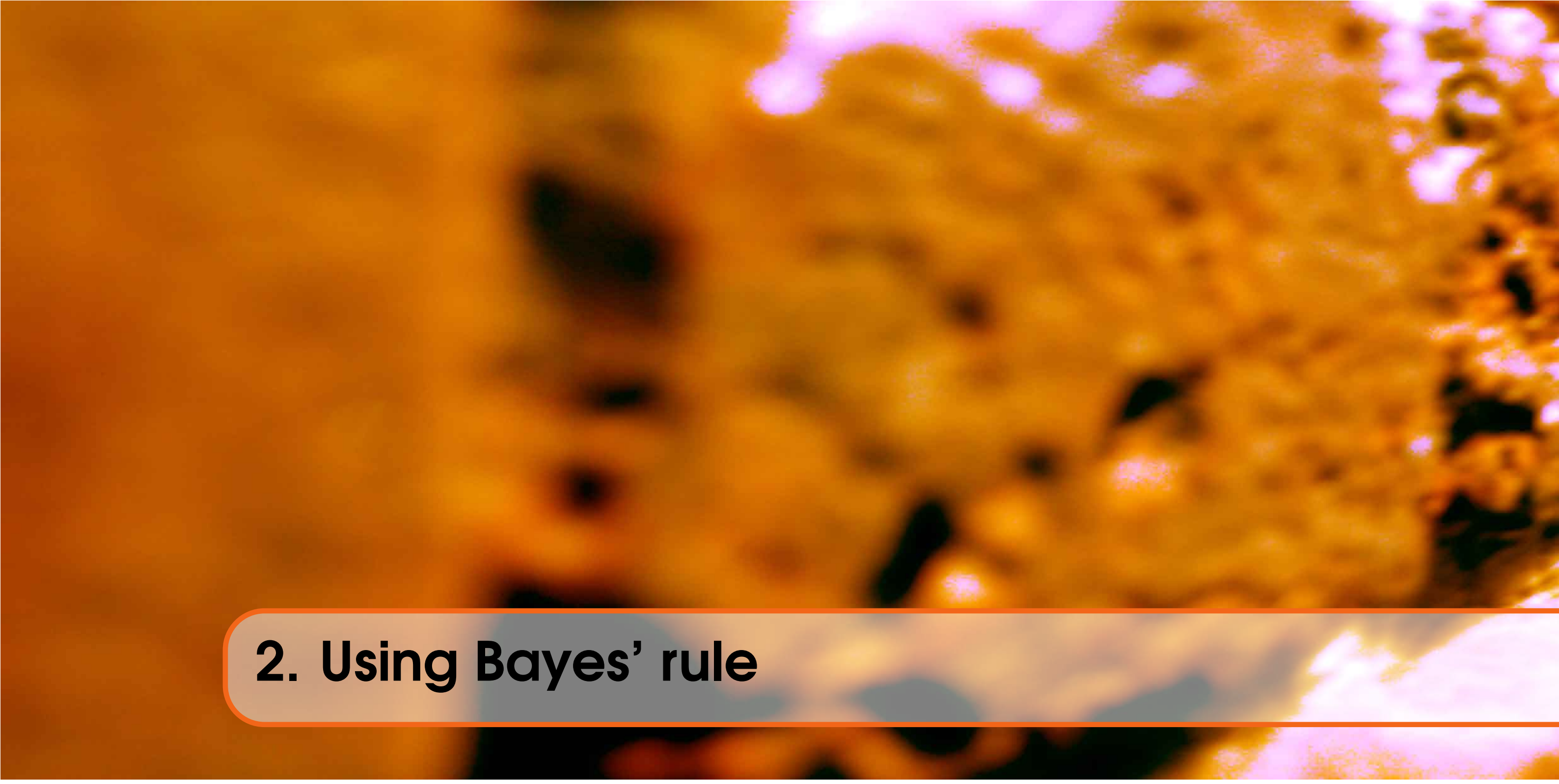
a) In prehistory, surroundings would almost never move. Once in a while, a person might accidentally ingest toxins. Assuming that your innate prior probabilities are based on these prehistoric frequencies of events, draw a bar diagram to represent your prior probabilities of the three scenarios above. No numbers needed.

b) In the windowless room on the ship, there is a big discrepancy between your visual and vestibular observations. Draw a bar diagram that illustrates the likelihoods of the three scenarios in that situation (i.e. how probable these particular sensory observations are under each scenario). No numbers needed.

c) Draw a bar diagram that illustrates the posterior probabilities of the three scenarios. No numbers needed.

d) Explain using the posterior probabilities your "percept" – why you might vomit in this situation.

Problem 1.12 Sometimes, when you press the button to call an elevator, you notice that the elevator car starts moving immediately afterwards (as shown by a display showing the car's current floor). Argue that in those situations, the likelihood that there is nobody inside when it arrives is much higher than the likelihood that there are people inside. (Priors do not play a role in this problem.)



2. Using Bayes’ rule

How do we make quantitative inferences?

In Chapter 1, we described the relevance of Bayes’ rule to understanding perception. In this chapter, we describe how to actually do calculations with Bayes’ rule.

Plan of the chapter

We describe the steps of Bayesian modelling. We present intuitive derivations of Bayes’ rule using mathematical and areal representations. We introduce simple perceptual scenarios with categorical (in particular: binary) variables in order to illustrate how observers perform Bayesian inference.

2.1 Steps of Bayesian modeling

Every Bayesian model consists of a series of steps that must be followed in order. The first step is to formulate the generative model, which represents the statistical structure of the world and the observations. The second step is inference: given a particular observation, how should the observer’s beliefs about the state of the world be updated? The third step is for the observer to reach a conclusion (a read-out) about the probable state of the world.

Step 1: The generative model

Let us return to the example from Chapter 1, that of estimating whether a floor is wet based on visual information (shiny or not). Here we have two possible world states (wet or dry). We have two potential observations (shiny or not). Every Bayesian model starts with a specification of the process by which we believe the observed data are generated; this is called a *generative model*. The generative model describes the statistical structure of the world and the observations; it does so by specifying the probability distributions of the variables. We often visually depict the structure of the generative model in a diagram called a *graphical model* (**Fig. 2.1A**). The generative model specifies probabilities of world states, e.g., that most floors are dry. Suppose that, in our experience, 10% of all floors are wet. Then:

$$p(\text{wet}) = 0.1 \tag{2.1}$$

$$p(\text{wet}) = 0.9 \tag{2.2}$$

The generative model also specifies the probabilities of observations conditioned on world states. Suppose that, in our experience, 80% of wet floors are shiny, but only 40% of dry floors are shiny. These probabilities are *conditional probabilities* and take the form $p(\text{observation} \mid \text{world state})$. These probabilities must be specified for each combination of observation and world state. In the example, we assume that

$$p(\text{shiny}|\text{wet}) = 0.8 \tag{2.3}$$

$$p(\text{not shiny}|\text{wet}) = 0.2 \tag{2.4}$$

$$p(\text{shiny}|\text{dry}) = 0.4 \tag{2.5}$$

$$p(\text{not shiny}|\text{dry}) = 0.6 \tag{2.6}$$

Step 2: Inference using Bayes' rule

If we make the observation that a floor is shiny, we will want to draw a conclusion about whether or not it is wet. This is a form of inference. A reasonable definition of inference is:

Definition 2.1.1 Inference is the process of drawing a probabilistic conclusion about a state of the world based on incomplete or imperfect information.

Inference can be done using Bayes' rule, a central rule of probability calculus:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{2.7}$$

where A and B are any two random variables.

Exercise 2.1 Prove Bayes' rule. Hint: according to the chain rule, the probability for having both A and B is $p(A,B) = p(A)p(B|A)$ ■

In the context of inference, A is a *hypothesized* rather than actual state of the world, and we will often denote it by H. B is an observation or set of observations, which we will denote by "obs". In the shiny floor example, the hypotheses are that the floor is wet and that it is dry.

Bayes' rule. Bayes' rule in the context of inference is therefore

$$p(H|\text{obs}) = \frac{p(\text{obs}|H)p(H)}{p(\text{obs})}. \tag{2.8}$$

Here, $p(H)$ is called the prior probability of H, $p(\text{obs}|H)$ is called the likelihood of H, and $p(H)|\text{obs})$ is called the posterior of H. This equation reflects the intuitive nature of Bayes' rule as it applies to the evaluation of hypotheses: if a hypothesis is more often true in the world, then before we even take the observations into consideration, that hypothesis is more probable (prior), and if the observations are more expected under a hypothesis (likelihood), then that hypothesis is also more probable.

Model mismatch. In this chapter and most of the book, we assume that the observer's understanding of the generative model is correct, i.e., is an accurate description of the true statistics of world states and observations. This means that the observer's priors are numerically identical to the frequencies of the corresponding world states in the generative model, and that the observer's likelihoods correctly represent the probabilities of the observation given the world states. However, it is possible that a mismatch could happen between the observer's assumed generative model (sometimes called the observer's *internal model*) and the actual generative model. For example, for some reason you could believe that 50% of all floors are wet, even though the true percentage is 10%. We will comment on such *model mismatch* in several places later in the book.

Prior. In our example, the prior distribution reflects the believed frequencies of occurrence of the values of the world state of interest, here slipperiness. These values are obtained from the

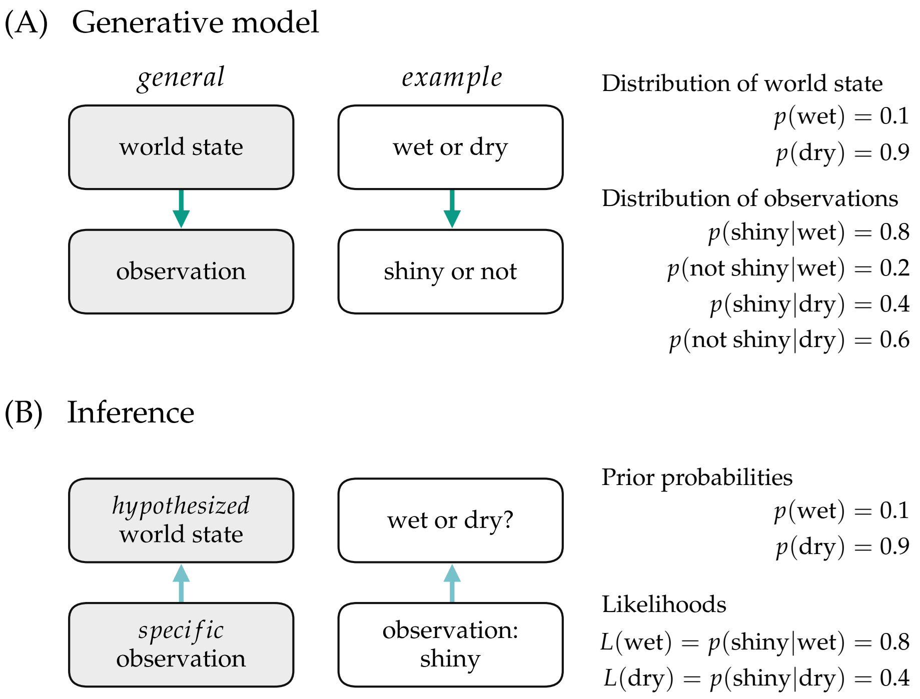

(A) Generative model
general
example
world state
observation
wet or dry
shiny or not
Distribution of world state
$p(\text{wet}) = 0.1$
$p(\text{dry}) = 0.9$
Distribution of observations
$p(\text{shiny}|\text{wet}) = 0.8$
$p(\text{not shiny}|\text{wet}) = 0.2$
$p(\text{shiny}|\text{dry}) = 0.4$
$p(\text{not shiny}|\text{dry}) = 0.6$
(B) Inference
hypothesized world state
specific observation
wet or dry?
observation: shiny
Prior probabilities
$p(\text{wet}) = 0.1$
$p(\text{dry}) = 0.9$
Likelihoods
$L(\text{wet}) = p(\text{shiny}|\text{wet}) = 0.8$
$L(\text{dry}) = p(\text{shiny}|\text{dry}) = 0.4$

Figure 2.1: (A) A generative model can be represented by a diagram with nodes and arrows. Each node is a variable, with the observation(s) always at the bottom. Each node is associated with a probability distribution. Each node with an arrow pointing to it is associated with a conditional probability distribution. Numbers are given for our wet floor example. **(B)** In inference, the observer has a specific observation (in the example: the floor is shiny) and tries to calculate the probabilities of the hypothesized world states. The same numbers are now used as priors and likelihoods. Inference "inverts" the generative model.

frequencies of these world states in the generative model (**Fig. 2.1**).

$$p(\text{wet}) = 0.1 \tag{2.9}$$

$$p(\text{wet}) = 0.9 \tag{2.10}$$

Likelihoods. Next we consider the likelihoods. Suppose you make the observation that the floor is shiny (see **Fig 2.2**). The likelihood of "wet" is then the probability that the floor is shiny if it is wet, whereas the likelihood of "dry" is the probability that the floor is shiny if it is dry. These values are obtained from the generative model, but only using those probabilities that pertain to the actual observation, "shiny". The likelihoods are

$$\text{Likelihood}(\text{wet}) = p(\text{shiny}|\text{wet}) = 0.8 \tag{2.11}$$

$$\text{Likelihood}(\text{dry}) = p(\text{shiny}|\text{dry}) = 0.4 \tag{2.12}$$

It is important to understand that likelihoods *do not need to sum to 1*. They are not probability distributions over the world state of interest. How probable the observations are under one scenario about the world is independent of how probable those same observations are under another scenario about the world.

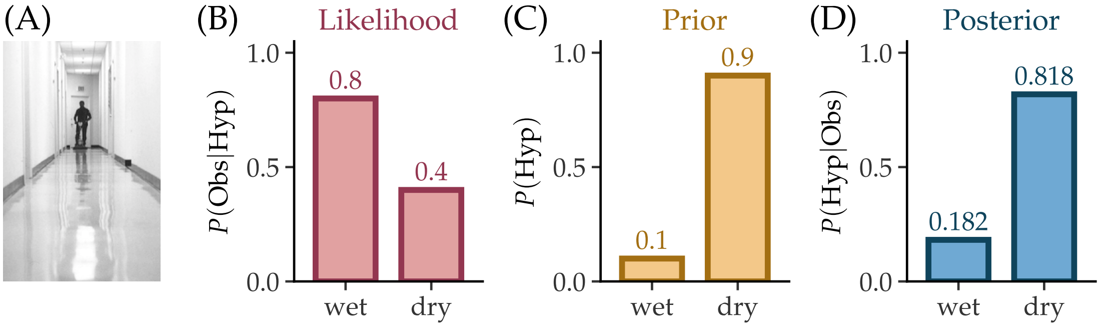


Figure 2.2: A shining example of perceptual inference. **(A)** Is this floor wet? **(B)** A wet floor is more likely than a dry floor to be shiny. **(C)** Most floors are dry. **(D)** The posterior distribution favors the hypothesis that the floor is dry; *figure inclusion pending permissions*.

Terminology note: Never say "the likelihood of the observation". The likelihood is always a function of the hypothesized state of the world. It is equal to the *probability* of the observation(s) under the hypothesized state of the world. To emphasize the dependence on the hypothesis, we sometimes write the likelihood as $\mathcal{L}(H)$ to denote $p(\text{obs}|H)$.

Protoposterior. The numerator in Bayes' rule, the product of likelihood and prior, does not have an official name, but we will call it the *protoposterior* of a hypothesis G:

$$\text{Protoposterior}(H) = p(\text{obs}|H)p(H) = p(H,\text{obs}) \tag{2.13}$$

In our example, the protoposteriors of the two hypotheses are

$$\text{Protoposterior}(\text{wet}) = \text{Prior}(\text{wet}) \cdot \text{Likelihood}(\text{wet}) = 0.1 \cdot 0.8 = 0.08 \tag{2.14}$$

$$\text{Protoposterior}(\text{dry}) = \text{Prior}(\text{dry}) \cdot \text{Likelihood}(\text{dry}) = 0.9 \cdot 0.4 = 0.36 \tag{2.15}$$

The protoposterior is the probability of the hypothesis and the observation. In our example, 8% of floors are wet and shiny, and 36% of floors are dry and shiny. Like likelihoods, protoposteriors do not need to sum to 1.

Normalization. Bayes' rule, Eq. 2.8 , tells us to divide the protoposteriors by $p(\text{obs})$, the probability of the observations. As it turns out, $p(\text{obs})$ is equal to the sum of the protoposteriors over all hypotheses:

$$p(\text{obs}) = \sum_H \text{Protoposterior}(H) \tag{2.16}$$

$$= \sum_H p(\text{obs}|H)p(H) \tag{2.17}$$

Exercise 2.2 Prove this. Hint: think of all the ways in which a given observation can be reached. ■

The probability $p(\text{obs})$ is the probability of the observations regardless of what world state they were produced by. In our example,

$$p(\text{shiny}) = \text{Protoposterior}(\text{wet}) + \text{Protoposterior}(\text{dry}) \tag{2.18}$$

$$= 0.08 + 0.36 = 0.44. \tag{2.19}$$

This calculation indicates that 44% of all floors are shiny.

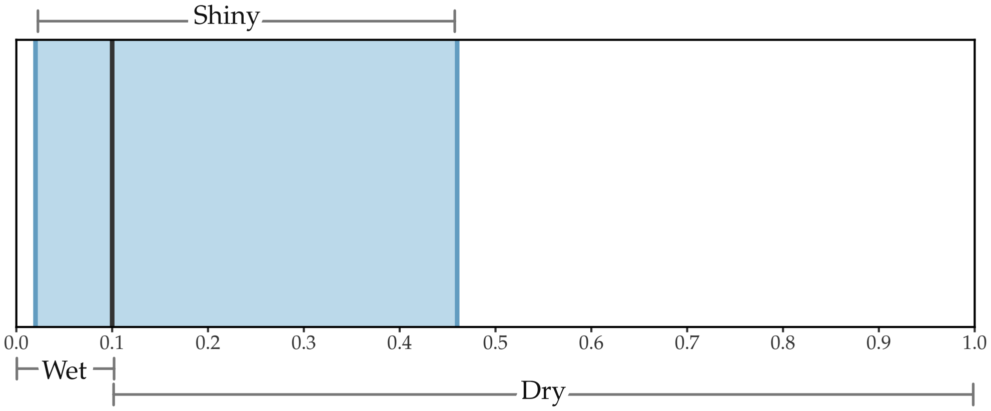


Figure 2.3: Areal representation of Bayesian inference. The posterior probability that the floor is wet, given that it is shiny, is the blue area within the left rectangle divided by the total blue area. The posterior probability that the floor is dry, given that it is shiny, is the blue area within the right rectangle divided by the total blue area.

Posteriors. As noted above, the protoposteriors do not sum to 1. In order to obtain posterior probabilities that sum to 1, we normalize the protoposteriors, i.e., divide each by their sum, $p(\text{obs})$. In this sense, $p(\text{obs})$ is a *normalization factor* in Bayes' rule. In our example,

$$\text{Posterior}(\text{wet}) = p(\text{wet}|\text{shiny}) = \frac{\text{Protoposterior}(\text{wet})}{p(\text{shiny})} = \frac{0.08}{0.44} = 0.182 \tag{2.20}$$

$$\text{Posterior}(\text{dry}) = p(\text{dry}|\text{shiny}) = \frac{\text{Protoposterior}(\text{dry})}{p(\text{shiny})} = \frac{0.36}{0.44} = 0.818. \tag{2.21}$$

Thus, $p(\text{obs})$ guarantees that unlike likelihoods or protoposteriors, posterior probabilities always sum to 1. The protoposterior is not normalized, but the posterior is. We could combine Eqs. (2.8) and (2.17) to obtain an alternative form of Bayes' rule:

$$p(H|\text{obs}) = \frac{p(\text{obs}|H)p(H)}{\sum_H p(\text{obs}|H)p(H)} \tag{2.22}$$

As a result of our observation that the floor is shiny, we have updated our beliefs in the hypotheses. Our belief that the floor is dry has reduced from our prior, 90%, to our posterior, 82%. The likelihood function favors the wet hypothesis (i.e., wet floors tend more often than dry ones to be shiny), so we have lost some confidence in the hypothesis that the floor is dry. Nevertheless, we still favor this hypothesis, because the strength of the evidence in favor of wet, as indicated by the ratio of likelihoods for the hypotheses, $\frac{\text{likelihood}(\text{wet})}{\text{likelihood}(\text{dry})} = 2$, is not sufficiently large to offset the ratio of priors, which favors the hypothesis that the floor is dry: $\frac{p(\text{dry})}{p(\text{wet})} = 9$.

Exercise 2.3 Suppose the floor is not shiny. What is the posterior probability that it is dry? ■

2.2 Areal representation

In order to deepen our understanding of Bayes' rule, let's recast the shiny floor problem by representing the relevant probabilities as areas of rectangles (**Fig. 2.3**). The large outer rectangle, of area 1, represents the universe of possibilities, namely that the floor is either wet or dry. The prior probability that the floor is wet, $p(\text{wet})$, is represented by the area of the left rectangle; the prior probability that the floor is dry, $p(\text{dry})$, is represented by the area of the right rectangle. The probability of a shiny floor, p(shiny), is represented by the shaded blue area. Blue fills 80% of the left rectangle and 40% of the right rectangle; these probabilities represent $p(\text{shiny} \mid \text{wet})$ and $p(\text{shiny} \mid \text{dry})$, respectively. The probability that the floor is both wet and shiny, $p(\text{wet, shiny})$, is the blue area within the left rectangle. The probability that the floor is both dry and shiny, $p(\text{dry, shiny})$, is the blue area within the right rectangle.

Now, let's suppose that you throw a dart that has uniform probability of landing anywhere within the large outer rectangle of area 1. Suppose your dart happens to land somewhere in the blue region (i.e., you have sampled a shiny floor). What is the probability that it is also within the left (wet) rectangle? Upon reflection, it should be clear that this probability is equal to the proportion of blue area that is also "wet," i.e.:

$$p(\text{wet}|\text{shiny}) = \frac{p(\text{wet}, \text{shiny})}{p(\text{shiny})} \tag{2.23}$$

Similar reasoning reveals that

$$p(\text{dry}|\text{shiny}) = \frac{p(\text{dry}, \text{shiny})}{p(\text{shiny})} \tag{2.24}$$

To drive this point home, imagine throwing 100 darts that land uniformly randomly within the large outer rectangle of area 1. Whenever a dart happens to land in the blue region, you record whether it is within the Wet or the Dry rectangle. On average, you would find that 44 of your 100 darts land in the blue, and that, of those 44 darts, 8 land in the Wet rectangle and 36 in the Dry rectangle. Therefore, the probability of wet, given shiny, is 8/44, and the probability of dry, given shiny, is 36/44, as we found above.

Step 3: Read-out of the posterior

The output of Bayes' rule is a posterior distribution over the hypothesized state of the world, $p(H|\text{obs})$. However, when we obtain the posterior distribution, we are not yet done, because usually the goal of a model is to predict what an observer would *decide* or *perceive.* A decision is an answer to a question such as "which hypothesis is true?" A percept is what the observer perceives. It could be the answer to the question "What do you see?" (or hear, smell, feel, etc.), but it is not necessarily tied to a query – it could be an unprompted experience.

Terminology note: Do not confuse the terms "percept" and "observation", even they might appear interchangeable. An observation is the input to the inference process, a percept is the output.

A *read-out* of the posterior is a mapping from the posterior distribution to a report of a hypothesis. The most obvious read-out is to pick the maximum, the hypothesis with the highest posterior probability. This read-out is also called *maximum-a-posteriori* (MAP) estimation. In our example, the MAP percept would be that the floor is dry. Moreover, the observer would be moderately confident about that conclusion, since the corresponding posterior probability is 81.8%.

Another difference between decision and percept is that the former can incorporate external rewards and costs, whereas the latter does not. For example, it can be much more costly to mistake a wet floor for dry than the other way round. As a consequence, even with a posterior probability of 81.8% that the floor is dry, it might still be a good decision to walk more carefully. We will say more about combining inference with utility in Chapter **??**.

We will also later put two more caveats on the universality of MAP estimation: one, decision noise might be present and cause deviations from MAP; two, for a continuous variable, picking the maximum is not always the best read-out.

2.3 The prosecutor's fallacy

We now consider several interesting examples that will serve to consolidate the reader's understanding of the steps outlined above, and that also hint at the vast range of scenarios to which Bayesian inference applies. The first of these is the prosecutor's fallacy.

As we have seen, conditional probabilities are not symmetric. In general $p(A|B) \neq p(B|A)$. For instance, in the shiny floor example, we found that $p(\text{shiny}|\text{wet}) = 0.8$ whereas $p(\text{wet}|\text{shiny}) = 0.182$. This asymmetry is apparent in the areal diagram (**Fig. 2.3**), where 80% of the Wet rectangle is filled with blue, but only 18.2% of the blue region falls within the Wet rectangle. Unfortunately, people who are untrained in the fine art of probabilistic thinking sometimes make the mistake of equating $p(A|B)$ and $p(B|A)$. This fallacy is called the *prosecutor's fallacy* or the *conditional probability fallacy*. The prosecutor's fallacy takes its name from the false argument, sometimes put forth in courts of law, that $p(\text{defendant is innocent}|\text{evidence}) = p(\text{evidence}|defendantisinnocent)$. For example, suppose that a partial, smudged fingerprint is found on a weapon left at a crime scene. A fingerprint database search reveals that a man who lives in the same city has a fingerprint that matches the one left on the weapon. A forensic expert testifies that only 1 in 1000 randomly selected people would provide such a match. The prosecutor argues that, based on the forensic expert's testimony, the probability that the defendant is innocent is only 1 in 1000. The prosecutor is confusing $p(\text{observation}|\text{innocent})$ – the testimony of the forensic expert – with $p(\text{innocent}|\text{observation})$.

Bayes' rule permits the correct calculation of $p(A|B)$ from $p(B|A)$ and other relevant probabilities and has been used for this purpose in some courts (Fenton, 2011). Let's suppose that the city has 1,000,001 (1 million plus 1) adult inhabitants. Given only that the defendant lives in the city, his prior probabilities of being innocent (H_1) or guilty (H_2) are therefore:

$$\text{Prior}(H_1) = \frac{1000000}{1000001} \tag{2.25}$$

$$\text{Prior}(H_2) = \frac{1}{1000001} \tag{2.26}$$

$$\tag{2.27}$$

The observation that the defendant's fingerprint matches that at the crime scene results in the likelihoods:

$$\text{Likelihood}(H_1) = \frac{1}{1000} \tag{2.28}$$

$$\text{Likelihood}(H_2) = 1 \tag{2.29}$$

$$\tag{2.30}$$

The protoposteriors are:

$$\text{Protoposterior}(H_1) = \text{Prior}(H_1) \cdot \text{Likelihood}(H_1) = \frac{1000000}{1000001} \cdot \frac{1}{1000} = \frac{1000}{1000001} \tag{2.31}$$

$$\text{Protoposterior}(H_2) = \text{Prior}(H_2) \cdot \text{Likelihood}(H_2) = \frac{1}{1000001} \cdot 1 = \frac{1}{1000001} \tag{2.32}$$

$$\tag{2.33}$$

The normalization is

$$\text{Normalization} = \text{Protoposterior}(H_1) + \text{Protoposterior}(H_1) \tag{2.34}$$

$$= \frac{1000}{1000001} + \frac{1}{1000001} = \frac{1001}{1000001} \tag{2.35}$$

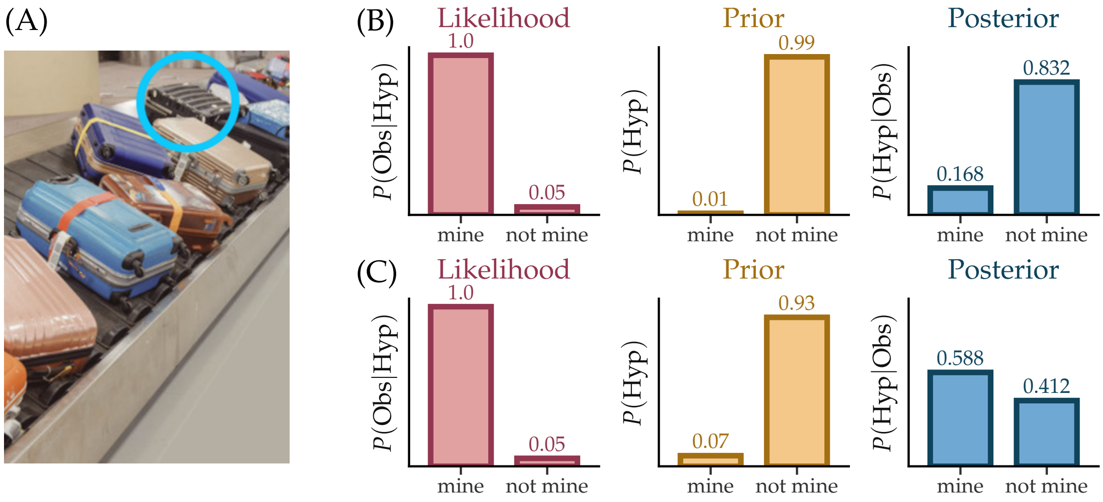


Figure 2.4: Expectation influences perception. **(A)** The first bag and the 86th bag both match yours in shape, size, and color. **(B)** Likelihood function, prior probability distribution, and posterior probability distribution upon viewing the 1st bag. Your posterior distribution indicates that the bag is probably not yours. **(C)** Likelihood function, prior distribution, and posterior distribution upon viewing the 86th bag. The same likelihood as in **(A)** combined with a different prior expectation produces a posterior distribution that favors the hypothesis the bag is yours. In this and all subsequent figures in the book, likelihood functions are drawn in red, prior distributions in yellow, and posterior distributions in blue; *figure inclusion pending permissions*.

Thus,

$$\text{Posterior}(H_1) = \frac{\text{Protoposterior}(H_1)}{\text{Normalization}} = \frac{1000}{1001} \tag{2.36}$$

$$\text{Posterior}(H_2) = \frac{\text{Protoposterior}(H_2)}{\text{Normalization}} = \frac{1}{1001} \tag{2.37}$$

The defendant is almost surely innocent, despite the prosecutor's argument! Another way to explain this: The city contains 1000001 people, 1000000 of whom are innocent and 1 of whom is guilty. Consequently, if we had the fingerprints of everyone in the city, we would expect 1001 matches, only 1 of which is from the guilty citizen. The probability of guilt given a fingerprint match is therefore $\frac{1}{1001}$.

2.4 A changing prior: luggage carousel example

We now move to a slightly more complicated example, in which the prior changes over time. Many air travelers have waited expectantly in an airport baggage claim area, watching for their bags to drop down the chute into the circulating luggage carousel (**Fig. 2.4A**). Let us suppose that you are engaged in this ritual of modern-day air travel along with 99 other passengers from your flight, each of whom, like you, checked one item of luggage. A recording piped through the speakers reminds you that "Many bags look alike. Please check your bag carefully before exiting the terminal." Indeed, your bag is one of the most popular models on the market, a black rectangular case used by 5% of all travelers. Of course, if you look at your bag close-up, you will notice individual markings — a name tag, a piece of string you have attached to the handle, etc. — that allow you to unambiguously identify your bag. But at the distance you are standing from the luggage chute, you cannot tell your bag from the 5% of bags in general that have the same shape, size and color. Now let's suppose that the first bag from your flight to enter the luggage carousel indeed has the same shape, size, and color as your bag. Is it your bag?

This question cannot be answered with a definitive "yes" or "no." Rather, the question demands a probabilistic judgment. You may consider it more or less likely that the bag is yours, but cannot yet be sure. In lieu of certainty, perception is most often characterized by varying degrees of confidence, which can be expressed as probabilities ranging from impossible to certain, occupying some particular place along the stretch of numbers between 0 and 1 (0% to 100%). As you view the bag in the luggage carousel, you will have an intuitive sense of the probability that it is your bag, $p(\text{this bag is mine} \mid \text{shape, size, color})$. But how could you arrive at this probability estimate?

Likelihoods. At the root of perceptual uncertainty is the fact that different world states can generate the same sensory observation. Not only do "many bags look alike," but many objects, people, and events produce nearly identical observations of one kind or another (sights, sounds, etc.). Thus, the information provided by the senses is typically imprecise, open to multiple interpretations.

What information is contained in your observation? If the bag you are viewing is in fact your own, it will have the same shape, size, and color. Thus, $p(\text{observed shape, size, color} \mid \text{my bag}) = 1$. But even if the bag you are viewing is not your own, it has some chance of matching the shape, size, and color of your bag. Since your bag is the model used by 5% of travelers, $p(\text{observed shape, size, color} \mid \text{not my bag}) = 0.05$. These two conditional probabilities are known as *likelihoods*. The likelihood of a hypothesis is the probability of the sensory observations if the hypothesis were true, or in other words, how expected the observations are if the hypothesis were true. As we have seen, a plot of the likelihood of every possible world state, known as the *likelihood function*, summarizes the degree to which the observation favors one world state interpretation over the other (**Fig. 2.4**). The less informative the observation, the "broader" or "flatter" will be the likelihood function; the more informative the observation, the "narrower" or "sharper" will be the likelihood function.

Prior. Importantly, the likelihood function, while a crucial component of the inference process, is not exactly what the observer wants to know. The likelihood function plots the probability of the observation given each hypothesized world state: $p(\text{observation} \mid \text{hypothesized world state})$. What the observer wants to know, however, is the probability of each possible world state, given the observation: $p(\text{world state} \mid \text{observation})$. To make this distinction clear, and to discover how to move from $p(\text{observation} \mid \text{world state})$ to $p(\text{world state}|\text{observation})$, let's consider how your perceptual inference will change over time as you wait at the baggage claim carousel.

When the very first bag from your flight enters the luggage carousel, and you notice the resemblance to your own bag, you will be hopeful but at the same time probably somewhat doubtful, that the bag in question is your own. Your skepticism is justified because not only do 5% of bags look like yours, but the probability that your bag would emerge as the first off the flight is just 1 in 100. After all, your flight carried 100 passengers, each of whom checked one bag. Now let's suppose that you have waited expectantly at the carousel, viewing each bag that emerges and checking more closely those that resembled your own, only to find yourself, 10 minutes and 85 bags later, still without having encountered your bag. At this point, let's suppose that the 86th bag emerges, and it again resembles your own. This time, you will be more confident than before that the bag is yours, despite the fact that the observation, and therefore the likelihood function, is identical for the first and the 86th bags. This illustrates that your perception, $p(\text{world state}|\text{observation})$ is not the same as the likelihood, $p(\text{observation} \mid \text{world state})$.

In short, perceptual inference is based not just on the observation (as reflected in the likelihood function), but also on expectation. As explained previously, we represent expectation by *prior probability*. The prior probability of a world state is based on all relevant information except the current observation. In the present example, your experience of waiting patiently as 85 bags emerged onto the carousel, together with your background knowledge that 100 bags were present on your flight, has informed you that the prior probability that your bag will emerge next is 1 in 15 (i.e., 6.7%), which is greater than the 1% that it was for the first bag. Although prior probabilities are conditioned on experience and background knowledge, in the interest of brevity we usually omit the conditioning symbol (|) and write prior probabilities simply as p(hypothesized world state), e.g., p(the bag is mine) and p(the bag is not mine). We plot the prior probability of each hypothesized world state as a *prior probability distribution* (**Fig. 2.4B-C**).

Posterior. Let's calculate the posterior probability that the first bag that you see emerge onto the luggage chute is your own. We first enumerate the possible world states or hypotheses: H_1 (the bag is mine), and H_2 (the bag is not mine). Next, we write down the prior probabilities of each hypothesis, given our knowledge that this is the first bag to appear:

$$p(H_1) = 0.01 \tag{2.38}$$

$$p(H_2) = 0.99. \tag{2.39}$$

We then write the likelihoods that express the probability of the sensory observation, (shape, size, and color of the luggage seen) given each hypothesis

$$p(\text{observation}|H_1) = 1 \tag{2.40}$$

$$p(\text{observation}|H_2) = 0.05. \tag{2.41}$$

Since the prior probability is 1% that the first bag is yours, it is 99% that the first bag is not yours. Note that, since the visual image shows a bag that matched yours in shape, size, and color, we set the likelihood to 1 for H_1. This is logical, since if it were your bag, the visual image will surely match the shape, size, and color of your bag. Finally, we enter the prior probabilities and likelihoods into Bayes' rule, to calculate the posterior probabilities of the hypotheses:

$$p(H_1|\text{observation}) = \frac{1 \cdot 0.01}{1 \cdot 0.01 + 0.05 \cdot 0.99} = 0.168 \tag{2.42}$$

$$p(H_2|\text{observation}) = \frac{0.05 \cdot 0.99}{1 \cdot 0.01 + 0.05 \cdot 0.99} = 0.832. \tag{2.43}$$

There are several important considerations to appreciate at this point:

1. First and foremost, it is important to realize that we have learned from the observation, updating our prior probability for H_1 (0.01) to a posterior probability that is much greater (0.168). Our posterior probably for H_1 has increased because the observation was more consistent with H_1 than with H_2. In general, the more strongly the observation favors one hypothesis over the other, the more we will learn.
2. Nevertheless, we are still more confident that the bag is not ours (83.2%) than that it is ours (16.8%). Despite the favorable observation, we believe that the bag is most probably not ours, because we started with such a low prior probability for H_1. In essence, the observation of a bag that looks like ours does not sufficiently favor H_1 to overcome our well-justified prior bias against H_1.
3. Another important point is that the posterior probability, $p(H_1|\text{observation}) = 16.8\%$, does not equal the likelihood, $p(\text{observation}|H_1) = 100\%$. As explained above, in general, $p(A|B) \neq p(B|A)$.
4. Finally, note that in this example the hypothesis with the maximum likelihood (known as the maximum likelihood estimate, or MLE) – H_1 – is not the hypothesis with the maximum posterior probability (the maximum a posteriori estimate, MAP) – H_2. This situation is not uncommon in perceptual inference. Sometimes the MLE and the MAP are the same, but often they are not.

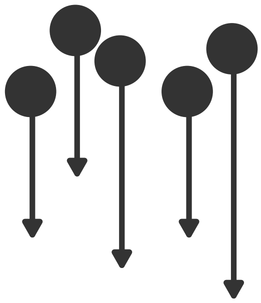


Figure 2.5: Five dots that all move downward.

Now suppose that we continue to wait for our bag to appear, failing to see it among the first 85 bags to enter the carousel. To calculate the posterior probability that the 86th bag, which also matches ours in shape, size, and color, is our own, we follow the same procedure, but with new prior probabilities of $\frac{1}{15}$ for H_1 and $\frac{14}{15}$ for H_2 (**Fig. 2.4C**).

Exercise 2.4 : Verify that the posterior probabilities when evaluating the 86th bag will be roughly $p(H_1|\text{observation}) = 0.588$ and $p(H_2|\text{observation}) = 0.412$. ■

Thus, the probability that the bag we are viewing is our own has now increased dramatically, from 16.8% (first bag seen) to 58.8% (86th bag seen), despite the fact that in the two cases the observation, and therefore the likelihood functions, are the same. The posterior distribution depends not only on the sensory data but also on the prior distribution.

2.5 A flat prior: Gestalt perception example

A misperception about Bayesian perceptual inference is that unequal priors must always play a role. In fact, important results can be obtained from scenarios that involve flat prior distributions, as the following example shows. Suppose that you observe five dots all moving downward, as indicated by the arrows in **Fig. 2.5**. Most people would perceive such a stimulus as a single group or object moving downward. The traditional account of the percept is that the brain has a tendency to group the dots together because of their common motion. This is captured by the Gestalt principle of "common fate". Gestalt principles, however, are merely narrative summaries of the phenomenology. A Bayesian model can provide a deeper *explanation* of the percept and in some cases even make quantitative predictions.

Box 2.1 **Myth:** Bayesian inference is all about priors.
Truth: In many interesting Bayesian models, the prior is flat or mostly irrelevant. In the moving dots cases study, the "work is done" by the likelihoods. Besides the moving dots case study, we will see another example in Chapter 5, on cue combination. ■

Let's take a Bayesian approach to modelling this scenario.

Sensory observations. The retinal image of each dot serves as a sensory observation. We will denote these five retinal images by I_1, I_2, I_3, I_4, and I_5.

Generative model. The first step in Bayesian modeling is to formulate a generative model: a graphical or mathematical description of the scenarios that could have produced the sensory observations. Let's say that the brain considers only two scenarios:

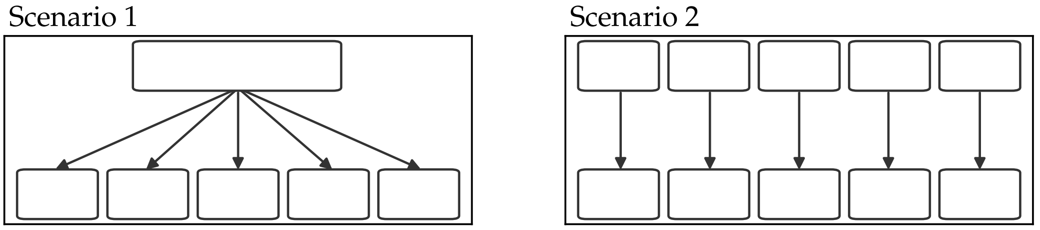
Scenario 1: All dots are part of the same object, and they therefore always move together. They move together either up or down, each with probability 0.5.

Scenario 2: Each dot is an object by itself. Each dot independently moves either up or down, each with probability 0.5.

(We are assuming that dots are only allowed to move up and down, and speed and position do not play a role in this problem.)

The generative model diagram below shows each scenario in a big box. Inside each box, the bubbles contain the variables and the arrows represent dependencies between variables. In other words, an arrow can be understood to represent the influence of one variable on another; it can be read as "produces" or "generates" or "gives rise to". The sensory observations should always be at the *bottom* of the diagram.

Exercise 2.5 Put the following variable names in the correct boxes below, for Scenario 1 (left) and Scenario 2 (right): retinal images I_1, I_2, I_3, I_4, and I_5, and motion directions s (a single motion direction), or s_1, s_2, s_3, s_4, and s_5. The same variable might appear more than once.



■

Inference. Consider now the specific observed configuration {down, down, down, down, down}, i.e. all five dots are moving down, which we will denote by I_{obs}. The brain's challenge is to determine based on I_{obs} which of the two Scenarios (1 or 2) is the right one. In inference, the two scenarios become *hypothesized* scenarios and we will denote them by H_1 and H_2, respectively. Inference involves likelihoods and priors. The *likelihood* of a scenario is the probability of the sensory observations under (i.e., given) the scenario. By the information provided in the problem, the two likelihoods are

$$\text{Likelihood}(H_1) = p(I_{\text{obs}}|H_1) = 0.5 \tag{2.44}$$

$$\text{Likelihood}(H_1) = p(I_{\text{obs}}|H_2) = 0.5^5 = \frac{1}{32} \tag{2.45}$$

Note, as we have seen previously, that the two likelihoods do not sum to 1. Note also that the second likelihood is much lower than the first one. This is because, under H_2, quite a coincidence is required for all the dots to happen to be moving downward.

Priors. Let's say that Scenario 1 occurs as often in the world as Scenario 2. The observer can use these frequencies of occurrence as prior probabilities, reflecting expectations in the absence of specific sensory observations. The prior probabilities of Scenarios 1 and 2 are then

$$\text{Prior}(H_1) = 0.5 \tag{2.46}$$

$$\text{Prior}(H_2) = 0.5 \tag{2.47}$$

$$\tag{2.48}$$

Protoposteriors. The protoposterior of the two scenarios are

$$\text{Protoposterior}(H_1) = \text{Prior}(H_1) \cdot \text{Likelihood}(H_1) = 0.5 \cdot 0.5 = \frac{1}{4} \tag{2.49}$$

$$\text{Protoposterior}(H_2) = \text{Prior}(H_2) \cdot \text{Likelihood}(H_2) = 0.5 \cdot 0.5^5 = \frac{1}{64} \tag{2.50}$$

Normalization. The normalization is the sum of the protoposteriors:

$$\text{Normalization} = \text{Protoposterior}(H_1) + \text{Protoposterior}(H_2) = \frac{1}{4} + \frac{1}{64} = \frac{17}{64}. \tag{2.51}$$

Posteriors. The posterior probabilities are obtained by dividing each protoposterior by the normalization:

$$\text{Posterior}(H_1) = p(H_1|I_{\text{obs}}) = \frac{\text{Protoposterior}(H_1)}{\text{Normalization}} = \frac{\frac{1}{4}}{\frac{17}{64}} = \frac{16}{17} \tag{2.52}$$

$$\text{Posterior}(H_2) = p(H_2|I_{\text{obs}}) = \frac{\text{Protoposterior}(H_2)}{\text{Normalization}} = \frac{\frac{1}{64}}{\frac{17}{64}} = \frac{1}{17} \tag{2.53}$$

Percept. In this case, the MAP percept is H1, i.e. that the dots move down together. This is consistent with the law of common fate.

2.6 Optimality, evolution, and motivations for doing Bayesian modeling

The notion of Bayesian inference is closely tied to that of *optimal* behavior. Optimality is defined only in terms of an objective function – a function that specifies how good or bad a response is, given a true state of the world. In perception, the objective function is usually accuracy – correctly reporting the true state of the world. In other decisions, external rewards and costs play a role. An observer could be optimal *with respect to* a specific objective function. For instance, considering further an example from Chapter 1, an observer might calculate a posterior probability of only 0.4 that a lion is hiding in the grass; nevertheless, it may be optimal to leave the area because the cost of remaining could be extremely high if the lion is present. Given a generative model and an objective function, computing the posterior is always a part of deriving the optimal solution. This is because the posterior represents whatever there is to know about the world state of interest given the observations. We will come back to the notion of optimality in Chapters 4 (Response distribution) and 13 (combining inference with utility).

A caveat on the link between Bayesian inference and optimality is that if the observer or agent uses an incorrect generative model in inference, their behavior will not be optimal. Using an incorrect generative model in inference is a form of *model mismatch*, which we will discuss in Chapter 3.

The link between Bayesian inference and optimality serves as a potential motivation for building Bayesian models. The argument would be that through evolution, some common and important brain functions might to a large extent have been optimized. Therefore, it is likely that in some tasks, behavior close to optimal will be found. This argument is more plausible for evolutionarily old functions – perception and movement – than for more recent functions, such as higher cognitive functions. However, one does not need to accept this argument in order to motivate Bayesian modeling. One could also view a Bayesian model as a starting point for building alternative, suboptimal models.

The fact that the posterior distribution is a part of *deriving* the optimal strategy should, however, not be taken to mean that optimal behavior implies that the brain represents posterior distributions. It could simply mean that the mapping from observations to response is the same as the one obtained using the Bayesian recipe. This is sometimes called "as-if" Bayesian behavior. Demonstrating that the brain represents posterior distributions (or priors, or likelihoods) is more involved.

2.7 Summary and remarks

In this chapter, we have introduced the precise formulation of Bayes rule and applied it to a range of discrete estimation problems. We have seen how Bayes rule makes concrete meaningful statements about probabilities possible. Regarding the use of Bayes rule, we have learned the following:

- All Bayesian modeling effectively starts with a model of the statistical structure of the world and the observations: the generative model.
- Conditional probabilities are not symmetrical. In general, $p(A|B) \neq p(B|A)$.
- Bayes rule calculates the probabilities of hypotheses conditioned on the observation – the posteriors, $p(H|\text{obs})$ – from the probabilities of the observation conditioned on the hypotheses – the likelihoods, $p(\text{obs}|H)$ – and the prior probabilities, $p(H)$.
- $p(\text{obs})$ in Bayes' rule normalizes the probability and thus implements the fact that we believe there to be exactly one correct hypothesis. $p(\text{obs})$ can be rewritten as $p(\text{obs}) = \sum_H p(\text{obs}|H)p(H)$.
- Priors and likelihoods can be equally important. In some problems, e.g. the Gestalt example, the prior is flat and the likelihood dominates. In other cases, e.g. the birds on a wire example in Chapter 1, the prior is most important.
- In some cases, e.g. the baggage claim example, the prior changes over time.
- In some cases, e.g. the "is that my friend?" example, the likelihood changes over time.
- Bayesian inference is a component of deriving the optimal strategy in any inference task. However, not all Bayesian inference is optimal, and experimentally finding near-optimal behavior does not necessarily mean that the components of the Bayesian model are internally represented in the brain.

2.8 Suggested readings

Original papers

- Gar Ming Chan. "Bayes' theorem, COVID19, and screening tests". In: *The American Journal of Emergency Medicine* 38.10 (2020), pages 2011–2013
- Norman Fenton. "Improve statistics in court". In: *Nature* 479.7371 (2011), pages 36–37
- Wilson S Geisler and Jeffrey S Perry. "Contour statistics in natural images: Grouping across occlusions". In: *Visual neuroscience* 26.1 (2009), pages 109–121
- Thomas L Griffiths and Joshua B Tenenbaum. "Optimal predictions in everyday cognition". In: *Psychological science* 17.9 (2006), pages 767–773

2.9 Problems

Problem 2.1 Think of three examples of random variables A and B for which intuitively, $p(A|B) \neq p(B|A)$. In each case, state which probability is greater, and explain why.

Problem 2.2 In early July, 2021, following large gatherings associated with Independence Day festivities, hundreds of people in Provincetown, MA, USA, became infected with the virus that causes COVID-19. According to an article published in the Washington Post (C.Y. Johnson, July 30, 2021), "A sobering scientific analysis published Friday found that three-quarters of the people infected...were fully vaccinated." The article does go on to emphasize that infected fully vaccinated individuals are very unlikely to suffer severe illness. Nevertheless, the quoted statement understandably alarmed many readers as it suggested that the vaccines against COVID-19 were ineffective at preventing infection. A crucial piece of information, not provided in the article, was that a large majority of people in Provincetown were fully vaccinated against COVID-19. Interpret these data, and the alarm readers may have felt, in light of the prosecutor's fallacy.

Problem 2.3 Imagine you have collected data about reported sightings of the dodo throughout history. We will call these data S. Suppose you are interested in the time the dodo went extinct, denoted E. Then the likelihood function of interest to you[1] is

[1] Incidentally, a paper has calculated this likelihood: Roberts DL, Solow AR (2003), *Flightless birds: when did the*

a) $p(E|S)$ as a function of S
b) $p(E|S)$ as a function of E
c) $p(S|E)$ as a function of S
d) $p(S|E)$ as a function of E

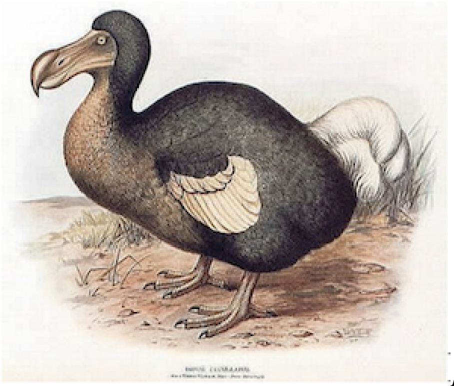


Figure inclusion pending permissions.

Problem 2.4 At a particular university, 15% of all students are in humanities, 55% of all students are undergrads, and 18% of undergrads are in humanities. What is the probability that a random humanities student is an undergrad?

Problem 2.5 1% of the population suffers from disease D. A diagnostic test for D is being piloted. The probability that someone without D tests positive (false-alarm rate) is 2%. The probability that someone with D tests negative (miss rate) is 3%.

a) Make a quick guess of the probability that someone who tests positive actually has D.
b) Calculate this probability. If it is very different from your answer to a), what went wrong in your intuition?
c) (*) (Due to Huihui Zhang, Beijing University) Suppose now that there is an extra variable we have ignored, namely whether someone goes to the doctor to have a diagnostic test done. This probability is higher if someone has the disease (because there will likely be symptoms) than if someone does not have the disease. Assume a 5-to-1 probability ratio for this. Now recalculate the probability that someone who tests positive actually has D. Is it closer to your original intuition?

Problem 2.6 Explain intuitively why likelihoods do not need to sum to 1, whereas priors and posteriors do.

Problem 2.7 Prove using Bayes' rule that when the likelihood function is perfectly flat (has the same value for all hypotheses), the posterior distribution is identical to the prior distribution.

Problem 2.8 (See Section 2.4.) Prove using Bayes' rule that if you see a bag on the luggage carousel that does not match yours (for instance, a small red bag, when yours is large and black), the posterior probability that it is yours is zero.

Problem 2.9 * (See Section 2.4.) You are one of 100 passengers waiting for your bag at an airport luggage carousel. Your bag looks the same as 5% of all bags. Derive a general expression for the probability that the bag you are viewing (which matches your bag visually) is your own, as a function of the number of bags you have viewed so far. How many bags must you view (without finding your own) before the posterior probability that the bag you are viewing (which matches your own visually) is greater than 70%?

Problem 2.10 This problem reveals a central feature of Bayesian inference; namely, that an observation increases or decreases support for a hypothesis not through the hypothesis' likelihood but rather through the ratios of likelihoods among hypotheses. As a consequence, we can omit irrelevant details (i.e., details that scale the likelihoods of all hypotheses equally) from our definition of the observation.

dodo become extinct? Nature, 426 (6964): 245.

In the luggage carousel example (Section 2.4.), we defined the visual observation as the shape, size, and color of the bag seen, and we therefore took $p(\text{observation}|H_1)$ to equal 1 when the observation matched the shape, size, and color of your bag. But of course, the exact "look" of a bag on a luggage carousel involves much more than just its shape, size, and color. For instance, as the bag enters the carousel, it may come to rest at any one of many different orientations. Suppose we were to expand our definition of the observation to include the bag's *orientation* as well as the other three features. To keep things simple, let's assume that there are 360 possible angles (one for each degree around the circle) and two possible sides (right-side up or upside down), for a total of 720 possible orientations with which a bag may come to rest on the carousel. If we further assume that each orientation is equally probable, then the probability of the observation given hypothesis 1 is no longer 1, but rather $\frac{1}{720}$. Similarly, the probably of the observation given hypothesis 2 would no longer be 0.05, but rather $\frac{0.05}{720}$. Since the likelihoods have changed, must not the posterior distribution change as well? Explain why or why not.

Now suppose that the 720 orientations do not all have probability 1/720 but that every orientation does have the same probability under each hypothesis. For instance, every bag (whether yours or not) comes to rest upright and aligned parallel to the edge of the carousel with probability 0.2. Again, would your inference be affected by the orientation of the observed bag?

Problem 2.11 Suppose you are waiting to catch a particular bus in a city that has just 10 bus routes; the route followed by each bus is indicated by an integer in the corner of its front display. You see the bus below from a distance, and naturally wonder whether this is the bus you are waiting for.

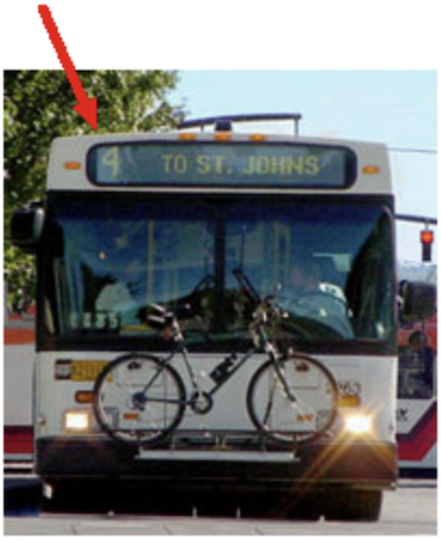


Figure inclusion pending permissions.

a) Based on the visual image of the difficult-to-discern bus route number (see arrow), and your intuitive understanding of how different route numbers might appear, construct a plausible likelihood function that plots $p(\text{visual observation} \mid \text{hypothesized bus route})$, for all numbers from 1 to 10.

b) As it turns out, you happen to know that only buses 3, 4, 5, and 6 travel down the street you are on. Furthermore, you know that buses 3 and 4 come twice as frequently as buses 5 and 6. Based on this background knowledge, construct your prior distribution for the bus number.

c) Use Bayes' rule to calculate your posterior distribution for the number of the bus.

Problem 2.12 [2] You are a student in a math class. The professor writes a symbol on the board that looks like a "u" or a "v" and you try to determine whether it is a "u" or a "v". The sensory observations consist of the retinal image of the handwritten letter, which we denote by I_{obs}. We make the following assumptions:

- There are no other possible letters.

[2] Thanks to Jonathan Gornet, then-undergraduate at New York University, for suggesting this problem.

- There is no relevant context.
- “u” occurs 1.5 times as often as “v”.
- The probability that a “u” produces I_{obs} is 0.0008, whereas the probability that a “v” produces I_{obs} is 0.0010.

Problem 2.13 In the birds-on-a-wire example from section 1.5, suppose the five birds sing with frequencies p_1 through p_5. You get 1 point for each time you correctly guess which bird is singing. Show that MAP estimation (under the correct generative model, which means incorporating knowledge of p_1 through p_5) maximizes the number of points you receive over many trials.

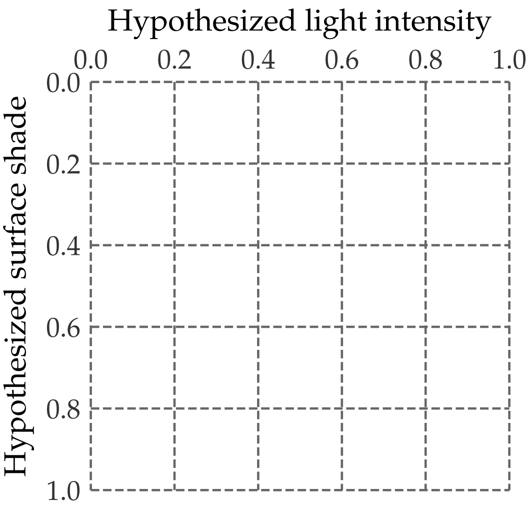
a) Explain why these probabilities are so low compared to the examples in this chapter.
b) Calculate the posterior probabilities $p(\text{“u”}|I_{\text{obs}})$ and $p(\text{“v”}|I_{\text{obs}})$.
c) If there were other possible symbols, would they typically have low likelihood, low prior, or both? Explain.

Problem 2.14 Inferring the shade of a grayscale surface. We use the formula

$$\text{Retinal intensity} = \text{surface shade} \cdot \text{light intensity} \tag{2.54}$$

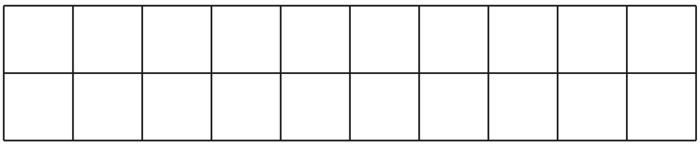
Let’s take each of these three variables to be between 0 (representing black) and 1 (representing white). For example, if the surface shade is 0.5 (mid-level gray) and the light intensity is 0.2 (very dim light), then the retinal intensity is $0.5 \cdot 0.2 = 0.1$.

a) Suppose your observed retinal intensity is 0.3. Connect in the diagram below all combinations of hypothesized surface shade and hypothesized light intensity that could have produced this retinal intensity. You should get a curved, not a straight line.



b) Explain the statement: “The curve that we just drew represents the combinations of surface shade and light intensity that have a high likelihood.”

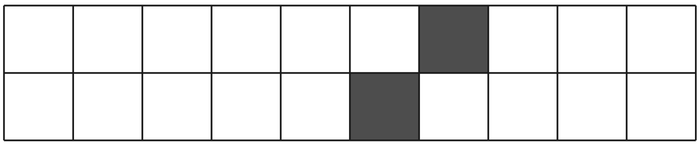
Problem 2.15 Imagine you live in a very boring world consisting of a 2 x 10 grid of squares:



Only two things ever happen in this world: *Scenario 1 (“Bar”):* With a probability of 40%, a vertical bar will appear in this world, consisting of one black square right above another, chosen so that each possible column is equally probable.

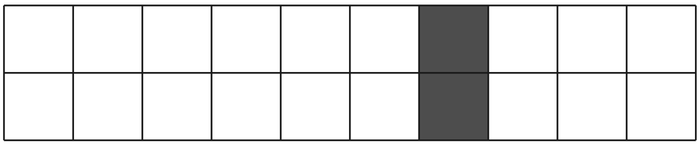
Scenario 2 (“Blocks”): With a probability of 60%, one black square will appear in a random position in the top row, and another black square will appear in a random position in the bottom row.

a) Suppose you have the following retinal image:



Calculate that based on this retinal image, the posterior probability of Scenario 1 is 0%. Write out all steps in your reasoning, with a brief explanation accompanying each step.

b) Suppose you have the following retinal image:



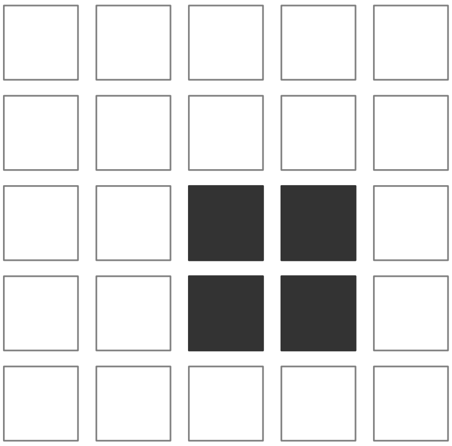
Calculate that based on this retinal image, the posterior probability of Scenario 1 is roughly 87%. Write out all steps in your reasoning, with a brief explanation accompanying each step.

c) Explain in fewer than 100 words how this imaginary world and specifically part (b) are related to Gestalt perception. Correctly use the terms "likelihood" and "percept".

Problem 2.16 Imagine you live in a different very boring world consisting of a 5 x 5 grid of squares. Only two things ever happen in this world:

Scenario 1 ("Big box"): With a probability of 50%, a big box consisting of 2 by 2 small black squares will appear in a random position.

Scenario 2 ("Small boxes"): With a probability of 50%, 4 independent black boxes will appear, each occupying 1 square and again in random places. The boxes will appear one by one, and no box can occupy a location of a box that is already present.



a) Draw a generative model diagram. Explain why you drew the diagram the way you did.

b) We will first consider Scenario 1. How many possibilities exist within this 5 x 5 grid to place a 2 x 2 object? Assuming equal probabilities of possible locations of appearance, what is the probability that it appears exactly in the observed location?

c) Is the answer to part (b) the likelihood, prior, or posterior of the hypothesis that there is a single object? Explain.

We will now consider Scenario 2.

d) Fill in: to place one of these 4 small boxes, there are ____ out of ____ locations on the grid that are consistent with the observations. This gives a probability of ______. To then place the second small box, there are then ____ out of ____ remaining locations on the grid that are consistent with the observations. This gives a probability of ______.

e) Repeat for the third and fourth small box

f) Multiply the four probabilities that you found in (d) and (e).

g) Is the answer to part (f) a likelihood, a prior, or a posterior of the hypothesis? Explain.

h) How many times bigger is the probability you found in (c) than the one you found in (f)?

i) How does your answer to (h) explain why human observers tend to perceive the set of 4 black boxes below as a single object.

Problem 2.17 This is a continuation of Problem 1.11 in Chapter 1.

a) Draw a diagram of the generative model. It should contain all of the italicized variables in the paragraph that starts with "Suppose", and a box for each scenario. Some variables might appear more than once.

b) Make up some numbers to illustrate the calculations you did there.

3. Bayesian inference under measurement noise

How does a Bayesian observer infer the state of the world from a noisy measurement?

In Chapter 1, we introduced the concept of inference by discussing a variety of daily-life examples to. In Chapter 2, we calculated posterior distributions for categorical variables. The use of categorical variables makes Bayesian calculations particularly convenient, but in practice, many variables are not categorical but rather are continuous. In perception, examples of continuous variables that the brain may want to infer include the orientation of a line segment, the location of a sound source, the speed of a moving object, the color of a surface, or the time elapsed between two events. Some continuous variables are strictly positive, such as speed, length, and duration. Other continuous variables are circular, such as orientation or motion direction.

Plan of the chapter

In this chapter, we consider real-valued variables, which can take values from negative infinity to infinity. We consider an observer who infers such variables from noisy sensory observations, such as might occur in a laboratory psychophysics experiment. We will introduce the concept of a measurement, which is an abstraction of the sensory observations that is suitable for continuous world state variables. Our focus will be on sensory noise as a source of uncertainty.

3.1 The steps of Bayesian modeling

Psychophysics is the study of how controlled stimuli are perceived or acted upon by organisms. For example, an experimenter might show you two lines on a computer screen and ask you which one is longer. When the lines are very similar in length, this is a difficult task and you will make mistakes. These mistakes can tell the experimenter about the way you are solving the task. Researchers have used psychophysics for more than a century to probe the nature of perceptual processing. The psychophysical task we will use as the leading example in this chapter is an auditory localization task. Imagine that you are facing a projection screen that displays a horizontal line stretching across the width of the screen. Located behind the screen, at the same elevation as the line, is a densely spaced array of many tiny loudspeakers. A tone will originate from one of these speakers. Your task is to report with a cursor the location from which you perceived the tone to emanate.

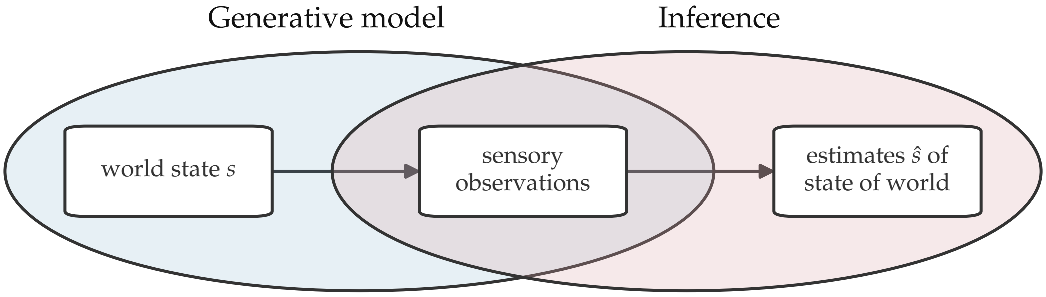


Figure 3.1: Schematic of a Bayesian model. The probabilities of world states and of sensory observations given world states constitute the generative model. Specifying these distributions is Step 1. On each trial, the observer performs inference to obtain an estimate of the world state. Specifying an expression for this estimate is Step 2. Across many trials, the estimate itself follows a distribution for a given true s. Specifying this distribution is Step 3.

This task is repeated many times; each repetition is called a trial. In this experiment, you are estimating a continuous quantity, namely the position of a sound source along a line. You have sensory observations, but possibly also prior knowledge. The steps involved in building a Bayesian model for this simple task provide a complete recipe for building and applying any Bayesian model. The observer's task is to infer the value of a world state of interest (here sound location) from a given sensory observation or multiple sensory observations (here the sound waves impinging on your eardrum).

Step 1 is to define the *generative model*: the probabilities of world states and of sensory observations given world states. The generative model (also called forward model) describes the statistics of the task-relevant variables, including the observer's sensory observations. It is a full statistical description of what is happening in a task. Variables in the generative model always include the world state of interest and the observer's sensory observations, but could also include other variables. The generative model specifies the probability distributions over all variables in the task. In the slipperiness example of Chapter 2, the generative model specifies the probability that the floor is slippery, the probability that the floor is shiny given that it slippery, and the probability that the floor is shiny given that it is not slippery. Many of the distributions in the generative model are specified by the experimental design. For instance, the probability of a sound occurring at a particular location is specified in the experimental design. However, we need to make an assumption about the distribution of the sensory observations. Many Bayesian models allow for the possibility that sensory observations are noisy. “Noise” has a diverse set of meanings across distinct scientific fields, but in this book, we mean that the same stimulus does not always produce the same internal representation in the brain. Noise thus implies random variability of the sensory observations from trial to trial. Noise can be due to a wide variety of factors, both external (in the world) and internal (in the brain).

Step 2 is to derive how the observer performs inference, i.e. how they estimate the state of the world based on the sensory observations and prior expectations. This is also called the *recognition model*. This step consists of two substeps. The first substep is to compute the observer's posterior distribution, i.e. the observer's probability distribution over the world state of interest, given the sensory observations. The observer's inference process “inverts” the generative model, in order to reach a conclusion about the world state in light of the sensory observations. In the slipperiness example, the inference process consisted of computing the probability that the floor is wet from the sensory observations and prior information. The generative model as understood by the observer, along with the sensory observations, completely defines the posterior distribution; no additional information is needed. The second substep is to specify how the observer obtains an estimate of the world state from the posterior distribution. This could for example be the mode or the mean of the posterior. The inference process is typically, but not always, a deterministic function of the sensory observations: for given sensory observations, the estimate of the world state is always the same.

The generative model describes the input into the decision-making process (**Fig. 3.1**). The inference process describes the observer's calculation of a posterior probability distribution over world states, and selection of a world state estimate; in the end, the inference process is summarized as an input-output relationship between the sensory observations and the world state estimate. The estimate distribution, how frequently the subject will exhibit each possible behavior, is obtained by combining this input-output relationship with the distribution of the sensory observations. In the following sections, we undertake the full Bayesian modeling procedure for the example auditory localization task described above. Despite its simplicity, this example illustrates the three steps and captures many of the subtleties of Bayesian modeling.

3.2 Step 1: The generative model

The generative model is a description of the statistical structure of the task. In the auditory localization task, the world state the observer tries to infer is a single feature of the stimulus, namely its horizontal position along a continuum; the sound's loudness, frequency, or other characteristics are not of interest in this task. We will often call the task-relevant feature of a stimulus, denoted s, simply "the stimulus". The sensory observations generated by the sound location consist of a complex pattern of auditory neural activity. For the purpose of our model, and reflecting common practice in the modeling of psychophysical data, we reduce the sensory observations to a single scalar, namely a noisy *measurement* x. The measurement "lives" in the same space as the stimulus itself and has the same units as the stimulus. We will now elaborate on this concept.

3.2.1 The measurement: an abstracted sensory representation

The brain is a noisy place. A physical stimulus elicits activity in the nervous system. This activity will vary randomly from trial to trial even when the physical stimulus itself is identical each time. Such variability originates from many sources. Our sensors are subject to random variability due to intrinsic stochastic processes. For instance, thermal noise affects the responses of hair cells in the inner ear that sense sound waves. The transduction process by which the nervous system captures physical energy and converts it into an electrical response is also stochastic. For instance, the absorption of photons by photoreceptors is a stochastic process; only sometimes is there a response to a single photon. At the subcellular level, neurotransmitter release and ion channel opening and closing are stochastic processes. Behavioral consequences of noise. In some cases, this noise can be easily illustrated. For example, if we place the index finger of our right hand on top of a table and try to place the index finger of our left hand at the matching location underneath the table, we often observe quite a difference (typical variability is about 2 cm in this task). This indicates noise in our internal proprioceptive representations of limb location. Similarly, it is difficult to estimate whether one object is heavier than another based on our sense of force because the internal measurement of force is noisy – necessitating the use of scales to compare weights. These examples suggest that the relationship between stimulus and sensor response is stochastic.

In this book, as in most behavioral models, we do not directly model the neural representation of a stimulus, because it is both unnecessary and underconstrained (many additional assumptions would be needed). Instead, we define a measurement as an abstraction or reduction of the neural representation in stimulus space itself. For example, if the true location s of the sound is 3° to the right of straight ahead, then its measurement x could be 2.7° or 3.1°. The terminology "measurement" stems from the analogy with making physical measurements. If a stick is 89.0

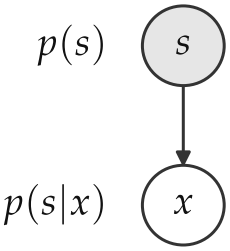


Figure 3.2: Schematic of a Bayesian model. The first step in Bayesian modeling is to define the generative model. This diagram is a graphical representation of the generative model discussed in this chapter. Each node represents a random variable, each arrow an influence. Here, s is the true stimulus and x the noisy measurement of the stimulus.

cm long, you might measure its length to be 89.5, 88.1, 88.9 cm, or so on. We say that the measurement "lives" in the same space as the stimulus, because it has the same units as the stimulus. A measurement of temperature is itself a temperature, a measurement of a color is itself a color, etc.

In our case study, the common space of stimulus and measurement is the real line, but it could be many other things, such as the positive real line (for a stimulus such as length or weight), the circle (example: motion direction), or a high-dimensional space.

Definition 3.2.1 — Measurement. A measurement of a stimulus is an abstraction of the noisy internal representation of that stimulus. The measurement lives in the same space as the stimulus itself and is typically modeled as the stimulus value plus a generic form of noise, such as Gaussian noise.

3.2.2 Graphical model

Our case study contains two variables: the stimulus (true sound location, s) and the observer's measurement of the stimulus, x. These two variables appear in the generative model, which is depicted in **Fig. 3.2**. A diagram like this is called a *graphical model*; it consists of nodes that contain the random variables and arrows that represent stochastic dependencies between variables. Each node is associated with a probability distribution. The variable at the end of an arrow has a probability distribution that depends on the value of the variable or variables at the origin of the arrow. In other words, an arrow can be understood to represents the influence one variable has on another. The arrow can be read as "produces" or "generates" or "gives rise to", e.g. "the sound location s gives rise to a measurement x". No arrow points to s, and therefore the distribution of s is a regular distribution $p_s(s)$. This distribution represents the overall frequency of occurrence of each possible value of the stimulus. The arrow pointing from s to x indicates that the distribution of x depends on the value of s. Mathematically, this is expressed as a conditional probability distribution $p_{x|s}(x|s)$. Conditional probability distributions are formally defined in Appendix A, Section 10. We will now describe the components of the generative model in detail.

R If you are not familiar with probability distributions, or if you have not worked with them recently, this would be a good moment to read Appendix B.

Box 3.1 — Notation for probability distributions. Strictly speaking, a probability should be labeled by both the random variable and its value. In other words, $p_s(2)$ would denote the probability of the random variable s evaluated at the value 2. The value is often general, which leads to somewhat redundant notation, such as $p_s(s)$. Therefore, we typically leave out the subscript indicating the name of the random variable and instead assign this name to the value. This shorthand notation is virtually always unambiguous. Occasionally, it is necessary to include the subscript, for example when a specific value gets substituted and one has to keep track of which distribution is being considered. In this chapter, we will keep the subscripts for clarity. ■

3.2.3 The stimulus distribution

The distribution associated with the stimulus s is denoted $p_s(s)$. This world state distribution or in our current example, stimulus distribution, reflects how often each possible value of s occurs. In models of cognition, $p_s(s)$ would be called the base rate of s, but this terminology is less common in perceptual modeling. In our case study, the experimenter has programmed a computer to draw the stimulus on each trial from a Gaussian or normal distribution with a mean μ_s and variance σ^2. This distribution is defined by the following probability density function (see Box):

$$p_s(s) = \frac{1}{\sqrt{2\pi\sigma_s^2}} e^{-\frac{(s-\mu_s)^2}{2\sigma_s^2}}. \tag{3.1}$$

This density is depicted in **Fig. 3.3A** for $\mu_s = 0$ and $\sigma_s = 3$. This Gaussian shape with mean zero implies that the experimenter more often presents the tone straight ahead than at any other location.

$$p_s(s) = \frac{1}{\sqrt{2\pi\sigma_s^2}} e^{-\frac{(s-\mu_s)^2}{2\sigma_s^2}}. \tag{3.2}$$

Box 3.2 — The Gaussian (normal) distribution. The most frequently used continuous probability distribution is the normal or Gaussian distribution. Its density function is

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}. \tag{3.3}$$

This is the famous "bell-shaped" distribution (or bell curve). We will sometimes use the notation $p(y) = \text{Normal}(y; \mu, \sigma^2)$ as short-hand for Eq. (3.3). The parameters μ and σ are the mean and variance of the random variable, respectively. The factor is needed so that the total probability – the integral of $p(y)$ – is equal to 1; such a factor is called a *normalization constant*. The exponent, $-\frac{(y-\mu)^2}{2\sigma^2}$, has a maximum value of zero at $y = \mu$, which is therefore the maximum of the Gaussian distribution. From this maximum outward, the exponent decays. It will be -1 once the difference between y and μ has reached $\sigma\sqrt{2}$. There, the Gaussian will have decreased by a factor of e. Gaussian distributions result when many randomly occurring fluctuations can affect the variable of interest. The more formal version of this statement is called the *Central Limit Theorem*. A typical example is the height of people, which follows a roughly Gaussian distribution, presumably because many factors contribute to height. Suppose that the average height for females is 165 cm, with a standard deviation of 10 cm. In this case we would find many females between 155 and 175 cm (within one standard deviation from the mean), fewer between 145 and 155 and between 175 and 185 cm, and very few above 185 cm or below 145 cm (more than two standard deviations away from the mean).

In this book, Gaussian distributions appear in many places. We will generally model the measurement distribution (see below) as Gaussian, which is motivated by the idea that many sources of random fluctuations contribute to the noisy measurement. However, it should be kept in mind that this is still an assumption. Gaussian distributions are convenient for analytical

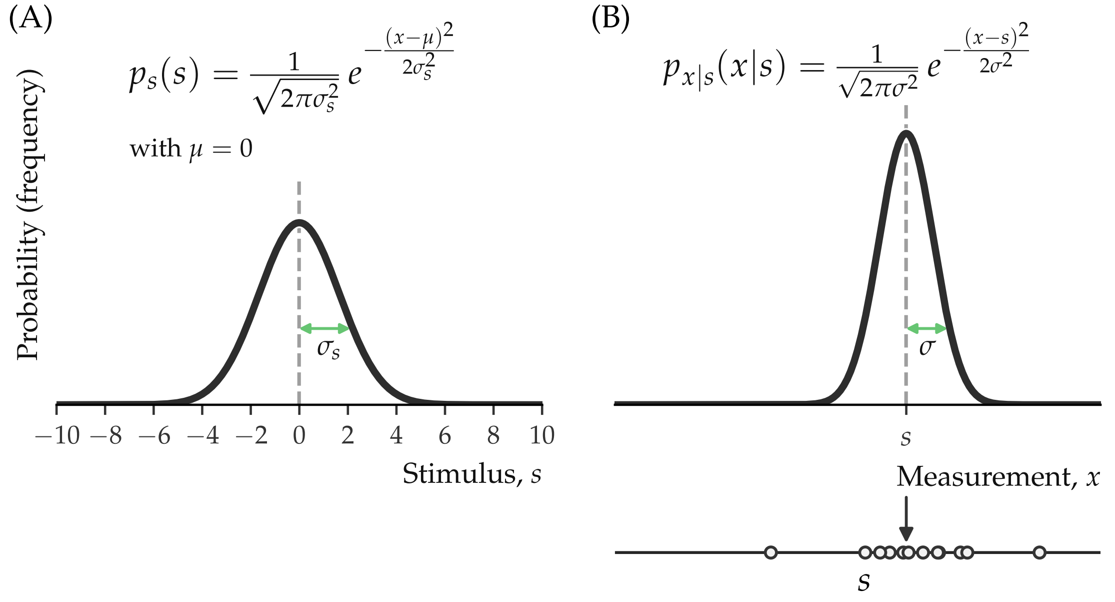


Figure 3.3: The probability distributions that belong to the two random variables in the generative model. **(A)** A Gaussian distribution over the stimulus, $p_s(s)$, reflecting the frequency of occurrence of each stimulus value in the world. We are not using units in our case study, but if you prefer to be concrete, you can think of the unit being cm, inches, or degrees of visual angle. In many plots, we will leave out numerical values altogether. **(B)** Suppose we now fix a particular value of s. Then we assume that the measurement x follows a Gaussian distribution around that s with variance σ^2. The diagram at the bottom shows a few samples of x.

calculations; for example, as we will see later, multiplying two Gaussians produces another Gaussian. Gaussian distributions are also convenient for simulations; for example, to draw samples from a Gaussian distribution with mean μ and variance σ^2, one can draw samples from one with mean 0 and standard deviation 1 (a standard normal distribution), multiply them by σ, and add μ. (Why not the other way round?) ■

Exercise 3.1 If one substitutes $y = \mu$ and $\sigma = 0.1$ in Eq. (3.3), one finds $p(y) = 3.99$. How can a probability be larger than 1? If the answer to this question is not immediately clear, read Section B.5.3, on the difference between probability mass and density functions. ■

3.2.4 The measurement distribution

The measurement distribution is the distribution of the measurement x for a given stimulus value s. This conditional distribution, $p_{x|s}(x|s)$, describes the frequency of occurrence of each value of the measurement when the same stimulus value s is repeated many times. If many sources contribute to the variability of the measurement, we will end up with a measurement distribution that is roughly Gaussian. This assertion is – loosely – a consequence of the Central Limit Theorem (see Box 3.2). While the Gaussian form of the stimulus distribution, Eq. (3.1), is often chosen simply because it facilitates calculations, the Gaussian form of the measurement distribution is quite fundamental, independent of the experimental design, and common to most Bayesian models we discuss in this book. Thus, the equation for the measurement distribution is

$$p_{x|s}(x|s) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-s)^2}{2\sigma^2}}, \tag{3.4}$$

where σ^2 is the variance of the noise in the measurement, also called *measurement noise level* or *sensory noise level*. This Gaussian distribution is shown in **Fig. 3.3B** for one value of s. The higher σ, the noisier the measurement and the wider its distribution. Lowering the light level in a room, increasing the distance to an object, decreasing the presentation time, or removing your corrective eyewear are all ways to increase the variance of the measurement noise of a visual stimulus. For an auditory or tactile stimulus, the same is achieved by introducing background noise, decreasing the intensity of a stimulus, or decreasing the presentation time. The inverse of the variance of the measurement distribution, $\frac{1}{\sigma^2}$, is sometimes called the reliability or the precision of the measurement x.

Box 3.3 — Noise and ambiguity. As explained in Chapter 1, there are many sources of uncertainty in perception. In this chapter, we consider uncertainty that arises from noise in the observer's sensory measurement. Because our sensory systems are universally subject to measurement noise, this form of uncertainty is always present to some degree in perception. However, uncertainty can additionally arises from ambiguity in the stimulus itself: different world states can produce the same sensory stimulus. An example of an ambiguous image is a shiny floor, which may or may not be slippery (Section 2.X). Later in the book, we will encounter further examples of ambiguous images (e.g., size-distance ambiguity). Whether uncertainty is caused only by sensory noise, or also by stimulus ambiguity, the end result is that the observation has nonzero probability given more than one hypothesized world state. ■

3.2.5 Joint distribution

Together, the two distributions $p_s(s)$ and $p_{x|s}(x|s)$ completely specify the generative model. One could combine them into a single, joint distribution which expresses the frequency of occurrence of every combination of s and x:

$$p_{s,x}(s,x) = p_s(s)p_{x|s}(x|s). \tag{3.5}$$

This mathematical identity specifies the joint distribution of all variables in the task.

3.2.6 Heteroskedasticity

In this book, we focus on the case where the measurement noise level σ in Eq. (3.4) does not depend on the stimulus. In practice, however, it often does. For example, noise in the measurement of visual position increases with retinal eccentricity, and noise in the measurement of auditory position increases with angle away from straight ahead. When the variance of a random variable depends on the mean, we have an example of *heteroskedasticity*. Heteroskedasticity does not prevent us from formulating a Bayesian model. However, heteroskedasticity is often an non-essential complication, which in particular limits the ability to derive analytical equations. Therefore, we will examine heteroskedasticity only in Problem 3.7.

3.3 Step 2: Inference

Organisms do not have direct knowledge of world states. The observer's brain has to infer the value of a world state of interest based on the observations. In our case study, that means inferring the value of s from an observed measurement, which we will denote by x_{obs}. First, the observer entertains different hypotheses about what s could be; we denote the hypothesized stimulus by s.

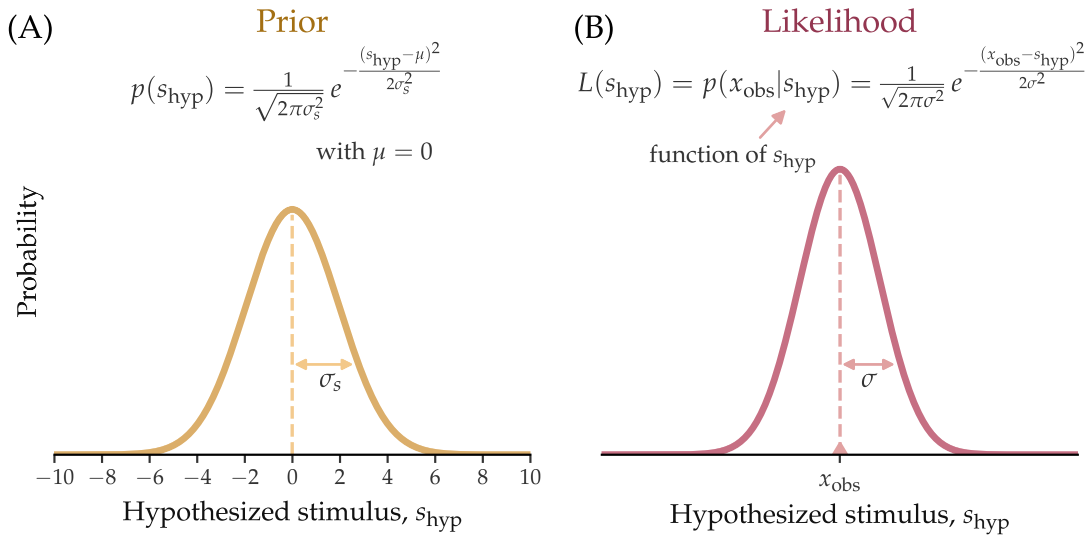


Figure 3.4: Consider a single trial on which the measurement is x_{obs}. The observer is trying to infer which stimulus s produced this measurement. The two functions that play a role in the observer's inference process (on a single trial) are the prior and the likelihood. The argument of both the prior and the likelihood function is s, the hypothesized stimulus. **(A)** Prior distribution. This distribution reflects the observer's beliefs about different possible values the stimulus can take. **(B)** The likelihood function over the stimulus based on the measurement x_{obs}. The likelihood function is centered at x_{obs}.

It is important to distinguish this from the true stimulus s. On a given trial, there is only one true stimulus, corresponding to the fact that there is only a single objective reality[1]; across many trials, s is a random variable. By contrast, s takes on a range of values even on a given trials; these values represent the different hypotheses that the observer entertains about the stimulus. Thus, s is neither a single value nor a random variable; it is a regular variable. Then, the observer computes the prior distribution, the likelihood function, and the posterior over s. Finally, the observer would "read out" the posterior distribution to obtain an estimate of the stimulus. Table 3.1 summarizes differences between Step 1 and Step 2 of the modeling process.

	Step 1: Generative model	**Step 2: Inference (decision model)**
Point of view of:	the experimenter	the observer (decision-maker)
Is realized:	across trials	on any single trial
Prior, likelihood, and posterior	play no role	are central
Nature of distributions (see Box 3.4)	objective (frequencies)	subjective (degrees of belief)
Stochastic?	always	by default not
At the end, the modeler has:	expressions for the distributions of all variables, including the observations	a rule to map observations to a decision (estimate of world state)

[1] Quantum mechanics plays no role in this book and should not in any form of cognitive modeling.

3.3.1 The prior distribution

In Section 3.2, we introduced the stimulus distribution $p_s(s)$, which reflects how often each stimulus value (auditory location) occurs in the experiment. Suppose that the observer has learned this distribution through extensive training on this experiment. Then, the observer will already have an expectation about the stimulus before it even appears, namely that $s=\mu$ will be most probable, and that the probability falls off according to the learned Gaussian curve. The expectation that the observer holds about the stimulus without having received any evidence on the given trial constitutes prior knowledge. The prior probability of a hypothesized stimulus value s is obtained by substituting that value in the stimulus distribution p_s; thus, the prior probability is equal to $p_s(s)$. In the inference process, $p_s(s)$ is referred to as the *prior distribution* (**Fig. 3.4A**). Unlike the stimulus distribution, the prior distribution exists on an individual trial: it reflects the observer's beliefs on that trial. It is therefore an example of a subjective distribution (Box 3.4): probability is interpreted as *degree of belief* rather than frequency of occurrence.

Box 3.4 — Objective and subjective probabilities. A distinction is sometimes made between objective and subjective probability distributions. Objective probabilities reflect frequencies of occurrence, while subjective probabilities are tied to an observer and reflect degrees of belief. Of the three steps in Bayesian modeling, the second one (inference) deals with subjective probability distributions, because all distributions in that step represent the beliefs the observer holds about world states on a given trial. The first and the third steps deal with objective probability distributions, since sensory observations and estimates can (in principle) be counted. The distinction between objective and subjective probability is discussed further in Section sec:B:objective-and-subjective-probability. This distinction is not important for calculations, only for interpretation. ■

3.3.2 The likelihood function

Intuition. The likelihood function represents the observer's belief about a variable given the measurements only – absent any prior knowledge. The likelihood function contains all information that can objectively be obtained from the measurement: no more information can be obtained, and any different information would be incorrect.

Definition. When the measurement distribution is known, so is the likelihood function. In our current example, we know the measurement distribution of x given s, $p(x|s)$. This means that we know the likelihood function over s, which we denote $\mathscr{L}(s;x)$:

$$\mathscr{L}(s;x) \equiv p_{x|s}(x_{\text{obs}}|s). \tag{3.6}$$

The slightly involved notation reflect the fact that we are substituting a specific observation, x_{obs}, and a specific hypothesized state of the world, s, into the measurement distribution that we have since Step 1, $p_x(x|s)$. Specifically, for the measurement distribution given by Eq. (3.4), the likelihood function is

$$\mathscr{L}(s;x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_{\text{obs}}-s)^2}{2\sigma^2}}. \tag{3.7}$$

Interpretation. At first sight, our definition of the likelihood seems strange: why would we define $p(x|s)$ under a new name? The key point lies in the fact that the likelihood function is a function of s, not of x_{obs}. The interpretation of the likelihood function is in terms of hypotheses. When an observer is faced with a particular measurement, what is the probability of that measurement when s takes on a certain value? Each possible value of the world state is a hypothesis, and the likelihood of that hypothesis is the observer's belief that the measurement would arise under that hypothesis. Thus, a fundamental difference between the measurement distribution and the likelihood function

	Prior ignored	**Prior incorporated**
Not normalized	Likelihood function	Protoposterior
Normalized	Normalized likelihood function	Posterior distribution

Table 3.1: Relationships between likelihood, normalized likelihood, protoposterior, and posterior.

is that the former is an objective probability distribution, while the latter represents the subjective beliefs of an observer (Box 3.4). This is analogous to the distinction between the world state distribution and the prior distribution. In the context of our auditory localization task, the likelihood function in **Fig. 3.4B** reflects the observer's belief that the measurement would arise from each hypothesized sound location.

Box 3.5 — The likelihood of what?. A likelihood function is numerically equal to a conditional probability, but is always a function of the variable after the "|" sign (for us the world state). It is common but incorrect to say "the likelihood of the measurements" or "the likelihood of the observations". The correct terminology is "the probability of the measurements (given a world state)" and "the likelihood of the world state (given the measurements)". ■

The likelihood function is not necessarily normalized, i.e., does not generally integrate to 1. The reason is that it is a function of the variable after the "given" sign, not of the one before it. This is why the likelihood function is called a function and not a distribution (a distribution is always normalized). This likelihood function, shown in **Fig. 3.4B**, happens to be normalized (over s), since in a Gaussian distribution, argument and mean can be interchanged without changing the distribution [2]. However, we already know from Chapter 2 that the likelihood does not need to be normalized.

The maximum-likelihood estimate. Using the likelihood function in Eq. (3.7), one could make a best guess of the value of the stimulus in the world. This is called the maximum-likelihood estimate of s, and we denoted it $\hat{s}$; the hat is common notation for an estimate [3]. In our case study, the maximum-likelihood estimate is simply equal to the measurement, x_{obs}. This means that the location of a sound source that would with highest probability produce the measurement x_{obs} is x_{obs} itself.

The width of the likelihood function. The width of the likelihood function is interpreted as the observer's level of uncertainty. A narrow likelihood means that the observer is very certain, a wide likelihood that the observer is very uncertain. Although it follows from Eq. (3.7) that the width of the likelihood function is identical to the width of the measurement distribution, these widths have different interpretations. The latter quantifies the spread of the measurements, the former the level of uncertainty based on a single measurement.

3.3.3 The posterior distribution

An optimal Bayesian observer computes a posterior distribution over a world state from measurements, using knowledge of the generative model. In the example central to this chapter, the relevant posterior distribution is $p(s|x_{\text{obs}})$ the probability density function over hypothesized stimulus s

[2] Because $\mathscr{L}(s;x)$ happens to be normalized, we can talk about the variance of this likelihood function, which is σ^2. If $\mathscr{L}(s;x)$ had not been normalized, it would strictly speaking not have been appropriate to talk about variance, but it is still common to do so. The precise but cumbersome verbiage would be "variance of the normalized likelihood function".

[3] Formally, we can write the definition of the maximum-likelihood estimate as $\hat{s} = \underset{s}{\text{argmax}}\mathscr{L}(s;x)$. "Argmax" stands for "the argument of the maximum": the value of the variable written below it for which the function following it takes its largest value.

given a measurement x_{obs}. Bayes' rule takes the form

$$p(s|x_{\text{obs}}) = \frac{p(x_{\text{obs}}|s)p(s)}{p(x_{\text{obs}})} \tag{3.8}$$

It is also commonly written as

$$p(s|x_{\text{obs}}) \propto p(x_{\text{obs}}|s)p(s) \tag{3.9}$$

or, using Eq. 3.6, as

$$p(s|x_{\text{obs}}) \propto \mathscr{L}(s;x)p(s). \tag{3.10}$$

This states that the posterior is proportional to the product of the likelihood and the prior. In the last two expressions, we used the proportionality sign ($\propto$) to stand in for the factor $\frac{1}{x_{\text{obs}}}$, since this factor simply acts as a normalization constant. If we do not know the normalization factor, we still know the full shape of the posterior probability distribution. We refer to Box 3.6 for a more detailed explanation.

Box 3.6 — Why that proportionality sign?. It is very common to see the form of Bayes' rule in Eq. (3.10), with a proportionality sign. How is this particular form justified? It is because the denominator of Bayes' rule, here $p(\text{obs})$, does not depend on the argument of the posterior, here the world state of interest. Thus, it simply acts as a multiplicative constant. A multiplicative constant does not change the shape of a function or where that function is maximal. Of course, the multiplicative constant is not an arbitrary number. It has to be such that the total integrated probability equals 1. For this reason, $\frac{1}{p(x_{\text{obs}})}$ is also called a *normalization constant*. One can write $p_x(x_{\text{obs}})$ as the sum or integral of the numerator over all possible values of the world state:

$$p(x_{\text{obs}}) = \int p(x_{\text{obs}}|s)p(s)ds. \tag{3.11}$$

This is in analogy to Eq. (2.17). However, there, the world state variable was discrete and therefore the normalization consisted of a sum rather than an integral. The common way of dealing with the normalization constant is to first calculate the numerator, $p_{x|s}(x_{\text{obs}}|s_{\text{hyp}})p_s(s_{\text{hyp}})$, and then normalize at the end if desired. There is nothing wrong with explicitly writing $\frac{1}{p_x(x_{\text{obs}})}$. However, this factor would just stand there until the end of the computation, unless you choose to write it in the integral form during the computation and evaluate the inside of the integral along with evaluating the numerator. This would be cumbersome, however, since you would have to write down the same expression twice, once in the numerator, and once inside the integral in the denominator. The effect of working with the proportionality sign is that you first evaluate the entire numerator, and then in the end evaluate the denominator by plugging the final expression of the numerator into the integral. Sometimes that final evaluation of the integral is not even needed, because the integral has a standard known form (e.g. Gaussian). ■

Case study. We will now compute the posterior under the assumptions we made in the previous section about the stimulus distribution and the measurement distribution. Upon substituting the expressions for $\mathscr{L}(s_{\text{hyp}};x)$ and $p_s(s_{\text{hyp}})$ into Eq. (3.10), we see that in order to compute the posterior, we need to compute the product of two Gaussian functions. Multiplying two Gaussian functions over the same variable is a common occurrence in Bayesian models of perception. The result is, after normalization, a new Gaussian distribution (Box 3.7)

Box 3.7 — Multiplying two Gaussians. Here, we discuss the product of two Gaussian functions over the same random variable Y. One has mean μ_1 and variance σ_1^2, and the other one has mean μ_2 and variance σ_2^2:

$$p_1(y) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(y-\mu_1)^2}{2\sigma_1^2}} \tag{3.12}$$

$$p_2(y) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}. \tag{3.13}$$

We multiply these distributions just like we would multiply regular functions, and normalize the result (since the product is not automatically normalized). The resulting probability distribution is another normal distribution, now with mean

$$\frac{\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}. \tag{3.14}$$

and variance

$$\frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}. \tag{3.15}$$

Since it is a normal distribution, it should get the standard normalization constant of the normal distribution, namely 1 divided by the square root of 2π times the variance. These results are derived in a Problem.

Life-saving notation: Expressions arising from multiplying Gaussians involve many σ^2 factors, and they appear either in nested fractions or in expressions that are hard to interpret. This can make people miserable. Fortunately, there exists a useful simplification of notation that facilitates interpretation. This is to rewrite all expressions in terms of *precision* quantities, of the form

$$J \equiv \frac{1}{\sigma^2} \tag{3.16}$$

with subscripts as needed. The precision J is the inverse of variance. In the new notation, the product of the two Gaussians above has a mean equal to $\frac{J_1\mu_1+J_2\mu_2}{J_1+J_2}$ and a precision equal to J_1+J_2. These expressions are substantially simpler. We encourage you to get in the habit of using precision notation whenever Gaussian distributions get multiplied. ■

Applied to our problem, we find from Eq. (3.10) that the posterior is a new Gaussian distribution

$$p(s|x_{\text{obs}}) = \frac{1}{\sqrt{2\pi\sigma_{\text{post}}^2}} e^{-\frac{(s-\mu_{\text{post}})^2}{2\sigma_{\text{post}}^2}} \tag{3.17}$$

where the mean of the posterior is

$$\mu_{\text{post}} = \frac{\frac{x_{\text{obs}}}{\sigma^2} + \frac{\mu}{\sigma_s^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}}. \tag{3.18}$$

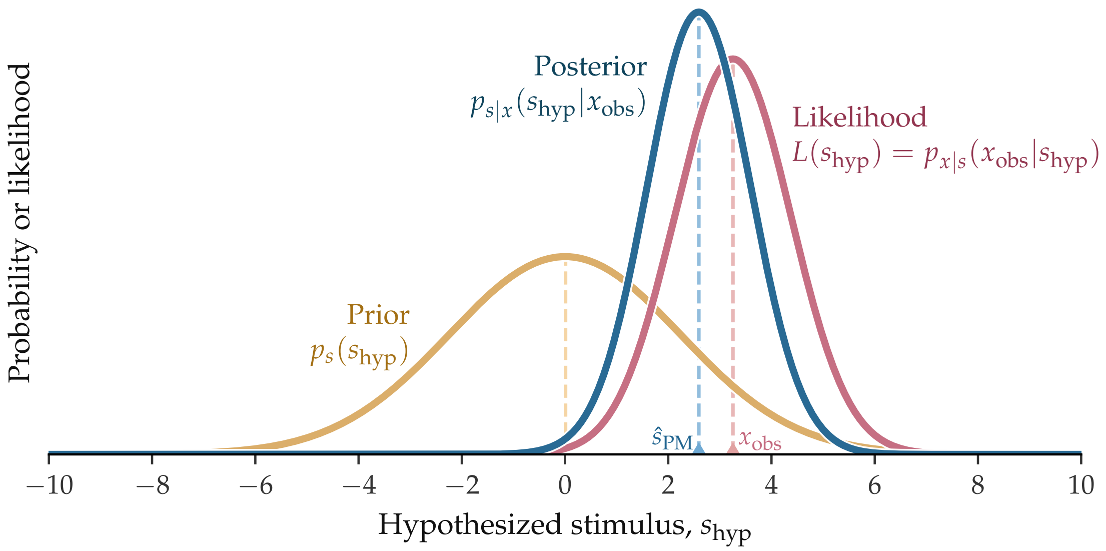


Figure 3.5: The posterior distribution is obtained by multiplying the prior distribution with the likelihood function, and normalizing the resulting function (the protoposterior). The hypothesized stimulus value with the highest posterior probability is the observer's posterior mean estimate of the stimulus, $\hat{s}_{\rm PM}$.

and its variance is

$$\sigma^2_{\rm post} = \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}}. \tag{3.19}$$

From here on, we will use a more convenient notation by introducing *precision* (inverse variance) variables (see Box 3.7):

$$J_s \equiv \frac{1}{\sigma_s^2} \tag{3.20}$$

$$J \equiv \frac{1}{\sigma^2} \tag{3.21}$$

$$J_{\rm post} \equiv \frac{1}{\sigma^2_{\rm post}}. \tag{3.22}$$

In this notation, Eqs. (3.18) and (3.19) can be written as

$$\mu_{\rm post} = \frac{Jx_{\rm obs} + J_s\mu}{J + J_s}. \tag{3.23}$$

and

$$J_{\rm post} = J + J_s. \tag{3.24}$$

This posterior is drawn in **Fig. 3.5**.

3.3.4 The posterior mean as a weighted average.

We now look more closely at the mean of the posterior, as given by Eq. (3.18) or (3.23). The expression of the mean is of the form $wx_{\rm obs} + (1-w)\mu$, where w is defined as

$$w \equiv \frac{J}{J + J_s}. \tag{3.25}$$

Since J and J_s are both non-negative, w is a number between 0 and 1. In other words, the posterior mean is a *weighted average* of the observed measurement, x_{obs} and the mean of the prior, μ. This weighted average will always lie somewhere in between x_{obs} and μ.

Exercise 3.2 Prove that statement mathematically. ■

Box 3.8 — Weighted averages. Suppose a student takes a midterm and a final exam and gets grades μ (for *M*idterm) and x (for e*x*am). However, the midterm is less important than the final. Therefore, the teacher weights the final exam grade by a factor $w = 0.7$ and the midterm grade by a factor $1-w = 0.3$. Then the student's overall grade in the class is the weighted average $wx + (1-w)\mu$. It will lie in between μ and x. ■

Where exactly μ_{post} lies is determined by the weights w and $1-w$. The weights are normalized versions of the inverse variances of the likelihood function and the prior distribution. If the variance of the likelihood is lower than that of the prior distribution, the inverse variance of the likelihood (i.e., the reliability of the measurement) is higher than that of the prior. As a consequence, the weight to x_{obs} is higher than to μ, causing the mean of the posterior to lie closer to x_{obs} than to μ. Of course, the reverse also holds: if the variance of the likelihood is larger than that of the prior, the mean of the posterior will lie closer to the mean of the prior than to the measurement. To Bayes' rule, the priors and the likelihood are just two pieces of information that need to be combined. Each piece of information has an influence that corresponds to the quality of the information.

Exercise 3.3 In the special case that $\sigma = \sigma_s$, compute the mean of the posterior. ■

The intuition behind the weighted average in Eq. (3.18) is that the prior "pulls the posterior away" from the measurement and towards its own mean, but its ability to pull depends on how narrow it is compared to the likelihood function. If the likelihood function is narrow – which happens when the noise level is low – then the posterior will not budge much: it will be centered close to the mean of the likelihood function. This intuition is often still valid even if the likelihood function and the prior are not Gaussian.

3.3.5 Width of the posterior

So far, we have discussed only the mean of the posterior. The variance of the posterior is given by Eq. (3.19); the corresponding variance is a measure of the width of the posterior. It is interpreted as the overall level of uncertainty the observer has about the stimulus after combining the measurement with the prior. It is different from both the variance of the likelihood function and the variance of the prior distribution.

Exercise 3.4
a) Show that the variance of the posterior can also be written as $\frac{\sigma^2\sigma_s^2}{\sigma^2+\sigma_s^2}$. Note that this is not our favorite way of writing it, as it is harder to interpret than Eq. (3.19) and in particular than Eq. (3.24).
b) Show that the variance of the posterior is smaller than both the variance of the likelihood function and the variance of the prior distribution.

■

The significance of the posterior variance being smaller than the individual likelihood and prior variances is that combining a measurement with prior knowledge makes an observer less uncertain about the stimulus, compared to when the observer has only the measurement or only prior knowledge.

Exercise 3.5 What is the variance of the posterior in the special case that $\sigma = \sigma_s$? What are the mean and the variance of the posterior when $\frac{\sigma}{\sigma_s}$ is very large or very small? Interpret. ■

3.3.6 The posterior mean estimate

Motivation. Given a posterior distribution, the Bayesian observer has use it to obtain an estimate of the world state of interest, here the stimulus s. The Bayesian observer does this by minimizing the expected value of some quantity. For continuous variables such as s here, a common assumption is that the observer minimizes expected squared error. As we will prove in Chapter 13, this is equivalent to choosing the *mean* of the posterior distribution. This is called *posterior mean* (PM) estimation[4].

Definition 3.3.1 — Posterior mean estimate. In general, the posterior mean estimate, denoted by $\hat{s}_{\text{PM}}$, is defined as

$$\hat{s}_{\text{PM}} \equiv \int sp(s|x_{\text{obs}})ds, \tag{3.26}$$

or in other words, as the expected value of s under the posterior distribution $p(s|x_{\text{obs}})$[a].

[a] For a review of expected values, see Section B.6.

Case study. In our example of combining a measurement with a Gaussian prior, we already calculated the posterior mean in Eq. (3.23). Thus, we have

$$\hat{s}_{\text{PM}} = \frac{Jx_{\text{obs}} + J_s\mu}{J + J_s}. \tag{3.27}$$

Thus, the PM estimate is a weighted average of the observed measurement x_{obs} and the prior mean μ, weighted by the inverse variances of likelihood function and prior, respectively.

Response. Recall that the observer's task is to report an estimate of the stimulus on a continuum. According to the model of the task adopted here, the PM estimate is what the Bayesian observer should report. A stronger statement, with different philosophical implications, is that the estimate is also what is being *perceived*, i.e. that the observer sees, hears, etc. the PM estimate of the stimulus.

3.4 Uncertainty and confidence

Uncertainty and confidence are common terms in daily life, but can be made specific in the context of Bayesian modeling.

3.4.1 Uncertainty

Every belief function in the inference stage (Step 2) is associated with a notion of uncertainty: prior uncertainty reflects the quality of the observer's knowledge about the state of the world *before* making any observations, likelihood uncertainty or sensory uncertainty reflects the quality of the observer's knowledge about the state of the world solely based on the observations, and posterior uncertainty reflects the quality of the observer's knowledge about the state of the world *after* making the observations.

In all cases, uncertainty is tied to the observer and therefore a *subjective* quantity (**Box 3.4**). It would be incorrect to talk about uncertainty in the context of the generative model (Step 1). In particularly, the phrase "the uncertainty in the measurement" reflects confusion; correct usage would be "the uncertainty about the state of the world based on the measurement".

[4] Also called Bayes-least-squares estimation. An alternative read-out is to minimize expected *absolute* error, instead of expected *squared* error; this would lead to a *posterior median* read-out (see Problem 13.2. For Gaussian posteriors, the mean and the median are the same.

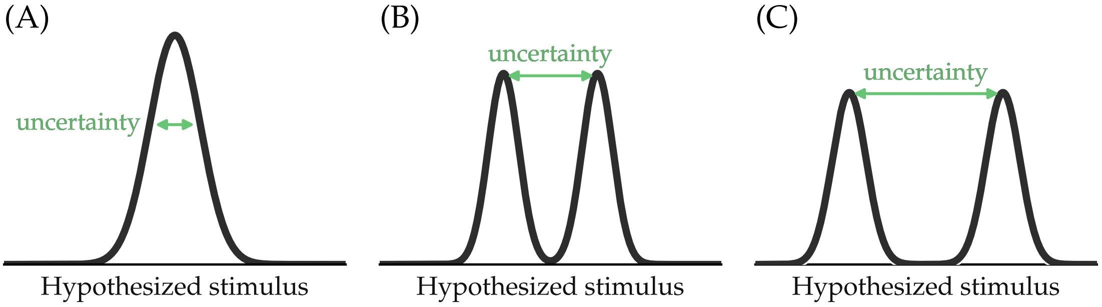


Figure 3.6: Uncertainty when defined as the standard deviation of a random variable. Each of the distributions shown could be the prior, the (normalized) likelihood, or the posterior. The double arrow is two standard deviations long. **(B)** and **(C)** show bimodal (two-peaked) distributions. Uncertainty in **(C)** is higher solely due to the greater separation between the peaks.

How, then, is uncertainty computed from a probability distribution or likelihood function? There is no generally agreed-upon definition, but the intuitive use of the term is captured if we define uncertainty as the variance of the probability distribution or likelihood function. For example, a narrow posterior distribution $p(s|x)$ means low posterior uncertainty, and a wide posterior distribution means high posterior uncertainty.

Definition 3.4.1 — Uncertainty about a state of a world. We define uncertainty measures as any monotonic function of the standard deviation of a probability distribution or likelihood function over that state of the world.

The part "monotonic function" refers to the fact that any monotonically increasing function of standard deviation (for example, $3\sigma+1$, σ^2, $\log\sigma$, or e^{σ}) would also be a legitimate definition of uncertainty and behave qualitatively in the same way. Here, we simply use the standard deviation itself.

Case study. In our case study, all distributions are Gaussian and the uncertainties are:

- Prior uncertainty: σ_s
- Likelihood or sensory uncertainty: σ
- Posterior uncertainty: $\sigma_{\text{post}} = \frac{1}{\sqrt{J+J_s}}$, which can also be written as $\frac{\sigma\sigma_s}{\sqrt{\sigma^2+\sigma_s^2}}$.

We see that likelihood uncertainty happens to have the same numerical value as sensory noise level, but in more complicated examples, that is not necessarily the case.

For Gaussian priors and likelihoods, posterior uncertainty is always smaller than both prior uncertainty and likelihood uncertainty (see Exercise 3.5). The definition of uncertainty extends to non-Gaussian distributions (see **Fig. 3.6**). However, for non-Gaussian priors and/or likelihoods, it is not necessarily the case that posterior uncertainty is always smaller than both prior uncertainty and likelihood uncertainty (see Problem).

Box 3.9 — Terminology: noise, uncertainty, variability. In this book, the term "noise" is reserved for the process by which the observations are generated, i.e. it describes the trial-to-trial variability of observations or measurements. Noise is thus part of the generative model. "Uncertainty", on the other hand, reflects the observer's knowledge, or lack of knowledge, about variables in the world. The width of the posterior distribution is a measure of uncertainty, not of noise. Uncertainty is part of the inference process and is *subjective* in the sense of Box 3.4. Noise is one possible cause of uncertainty, but not the only one. For example, when an object is partially occluded and there is no direct information about the part of the object behind the occluder, the observer has uncertainty without having noise.

Variability is an encompassing term for anything that varies from trial to trial. Noise is a form of variability; this can be called the variability of the measurement. The stimulus estimate is also variable from to trial. This can be called "behavioral variability". Uncertainty is *not* a form of variability. ■

3.4.2 Bayesian confidence

In daily life, decisions are made with greater or lesser confidence – the strength of the sense that you are right. You might be confident that you can cross the road before a car reaches you, that it is your friend who is approaching you, or that someone's accent is Italian. Confidence naturally fits into a Bayesian framework and is related to the posterior distribution. In the present chapter, the decision has been the estimate $\hat{s}$ that the observer makes of the stimulus. We can then define *Bayesian confidence* as any monotonic function F of the posterior probability density of that estimate:

$$\text{Bayesian Confidence} \equiv p_{s|x}(\hat{s}|x_{\text{obs}}). \tag{3.28}$$

Definition 3.4.2 — Bayesian confidence (about an estimate of a world state). (Any monotonic function of) the posterior probability distribution evaluated at that estimate.

The phrase "a monotonic function of" is inserted for the same reason as in our definition of uncertainty in Section 3.4.1. Note that our definition of Bayesian confidence does not mean that human confidence ratings in a behavioral experiment necessarily follow this definition.

Gaussian posteriors. The "estimate" in the definition of confidence might or might not be the PM estimate. We now apply Eq. (3.28) to a Gaussian posterior combined with a PM estimate. The maximum of a Gaussian probability density function is equal to the factor preceding the exponential (1 divided by the square root of 2π times the variance). Thus, for a Gaussian posterior, confidence from Eq. (3.28) would be $\sqrt{\frac{J_{\text{post}}}{2\pi}}$ or $\frac{1}{\sigma_{\text{post}}\sqrt{2\pi}}$. We further work this out in a Problem.

Empirical descriptions of confidence. Here, we have *defined* confidence as the posterior probability of the chosen value. However, one can also choose a more empirical approach of asking what model well describes human confidence ratings. There is evidence that human confidence ratings do not simply follow Eq. (3.28) and that they might not even be Bayesian at all.

3.5 Model mismatch in inference

So far, we have only discussed optimal Bayesian inference. However, that discussion was predicated upon the assumption that the observer possesses complete and correct knowledge of the generative model (Step 1), and fully utilizes this knowledge during inference (Step 2). However, it is possible that an observer uses a different, assumed generative model to perform inference. This is called *model mismatch* and could have many causes:

- Learning of the generative model has not completed.
- The generative model is too complex to learn and the observer approximates it.
- The generative model of an experiment is different from that in the natural environment, and the observer uses the latter for inference.
- The observer holds wrong beliefs about the generative model. A situation in which this might happen is when a subject is neither instructed about nor trained on the task distributions.

Bayesian observers are optimal when they possess and correctly incorporate full knowledge of all distributions in the generative model, but an observer with a mismatched likelihood or prior will in many cases be suboptimal. In our case study, that would mean having a higher expected squared error than the optimal observer. We will return to this in Chapter 4.

Definition 3.5.1 — Model mismatch. During inference, the observer uses a generative model that is incorrect for the task. The observer would still be Bayesian but potentially suboptimal.

True generative model	Generative model used in inference	Result
$p(x,s)$	$p(x,s)$	Optimal
	$q(x,s) \neq p(x,s)$	Potentially suboptimal

If an observer performs Bayesian inference using a generative model that differs from the true generative model, then the resulting behavior might be suboptimal.

When is the true generative model relevant? In natural conditions, researchers won't usually know the true generative model. Central to inference is the assumed generative model. The true generative model is only relevant when it plausibly serves as the basis for the assumed generative model. During learning, the assumed generative model will generally go closer and closer to the true generative model.

3.5.1 Prior mismatch

A special case of model mismatch is prior mismatch (another case would be "likelihood mismatch", where the likelihood function is derived from the wrong measurement distribution). The prior distribution at value s, $p(s)$, can be thought of as our belief that the stimulus was s before we have received any sensory information. Within the context of a psychophysics experiment, one cannot blindly assume that subjects learn the world state distribution and use it as a prior. In general, the observer's prior might differ from the world state distribution. Subjects might be using a prior that they come into the experiment with, for example one that is based on the world state distribution in the natural world. Well-established examples are a prior favoring low speeds and a prior for light coming from above. Such "natural" priors, acquired over a lifetime of sensory experience, might be hard to override during the relatively short duration of an experiment. Extensive training might be needed to override a natural prior. Of course, subjects might be using a prior intermediate between the "natural world state distribution" and the "experimental world state distribution". The prior might also change over time, as the observer is exposed to more stimuli during the experiment.

In Step 2, under prior mismatch, the observer would be computing the PM estimate using a different generative model than the correct one:

$$q(s|x_{\text{obs}}) \propto q(s)p(x_{\text{obs}}|s). \tag{3.29}$$

For example, if the assumed stimulus distribution $q(s)$ has mean μ_{assumed} and variance $\sigma^2_{s,\text{assumed}}$, then the PM estimate when the measurement is x_{obs} is

$$\hat{s}_{\text{PM, mismatched}} = \frac{Jx_{\text{obs}} + J_{s,\text{assumed}}\mu_{\text{assumed}}}{J + J_{s,\text{assumed}}}. \tag{3.30}$$

Thus, this observer would be making different trial-to-trial responses than the optimal Bayesian observer, and overall perform would be worse.

Box 3.10 — The improper prior. We considered the case of a flat prior, $p(s) = $ constant. An immediate question is then what value this constant takes. If s were limited to an interval, say $[a,b]$, the answer would be clear: $p(s) = \frac{1}{b-a}$, so that the area under the prior is 1 (**Fig. 3.7**). In this chapter, however, the domain of s is the entire real line. A uniform distribution is not properly defined on the entire real line, since the line has infinite length, so the uniform distribution would have value 0. In any practical task, s can of course not grow arbitrarily large in either direction. Therefore, it would be reasonable to cut off its domain at some large value. Choosing this value – and with it, the value of the constant prior – would, however, be arbitrary. Fortunately, this conundrum does not need to be solved, since it turns out that in the inference, the value of the constant prior does not play a role. Namely, if this value is c, then the posterior distribution is

$$p(s|x_{\text{obs}}) \propto cp(x_{\text{obs}}|s) \propto p(x_{\text{obs}}|s). \tag{3.31}$$

In other words, since the prior is constant, it gets absorbed into the proportionality constant. After normalizing the posterior distribution, c would drop out. We could even pretend that, for whatever value of c, $p(s) = c$ on the entire real line (**Fig. 3.7** yellow), and the posterior would still be well-defined in spite of the prior distribution not being normalized. Such an unnormalized prior distribution (i.e., one whose integral is infinite rather than 1) is called an *improper prior.*

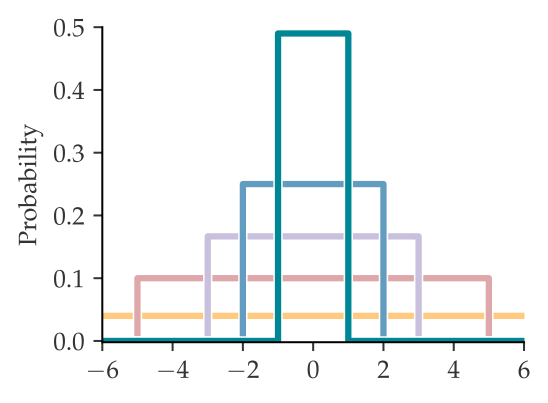


Figure 3.7: Various uniform priors on bounded intervals

■

3.5.2 The case of a flat prior

A special case of prior mismatch is if the observer uses a constant prior on the real line, $q(s) =$ constant. Such a prior can never reflect the true stimulus distribution, since a stimulus distribution is an objective distribution and has to be properly normalized. For inference, however, a prior that is not normalizable (*improper*) is okay as long as the posterior is normalizable (Box 3.10). Using a flat prior simplifies inference, because the posterior is then simply the normalized likelihood function.

3.6 Magnitude variables

We have assumed that the domain of the stimulus is the entire real line, from $-\infty$ to ∞. There are many world state variables that have a different domain. In Appendix B.7.6, we discuss variables that are periodic, such as angles. Here, we consider variables that can only take positive values and thus have a domain from 0 to ∞, including line length, depth, weight, speed, loudness, duration, and surface lightness. These can be called magnitude variables, and none of them can ever take negative values. For magnitude variables, a common choice is a lognormal distribution. Since the domain of s is $[0, \infty)$, the domain of the logarithm of s is the entire real line. Thus, it is possible to define a Gaussian distribution on $\log s$:

$$p(\log s) = \frac{1}{\sqrt{2\pi\sigma_s^2}} e^{-\frac{(\log s - \mu)^2}{2\sigma_s^2}} \tag{3.32}$$

Transforming this to the original variable s, we obtain

$$p(s) = \frac{1}{s\sqrt{2\pi\sigma_s^2}} e^{-\frac{(\log s - \mu)^2}{2\sigma_s^2}} \tag{3.33}$$

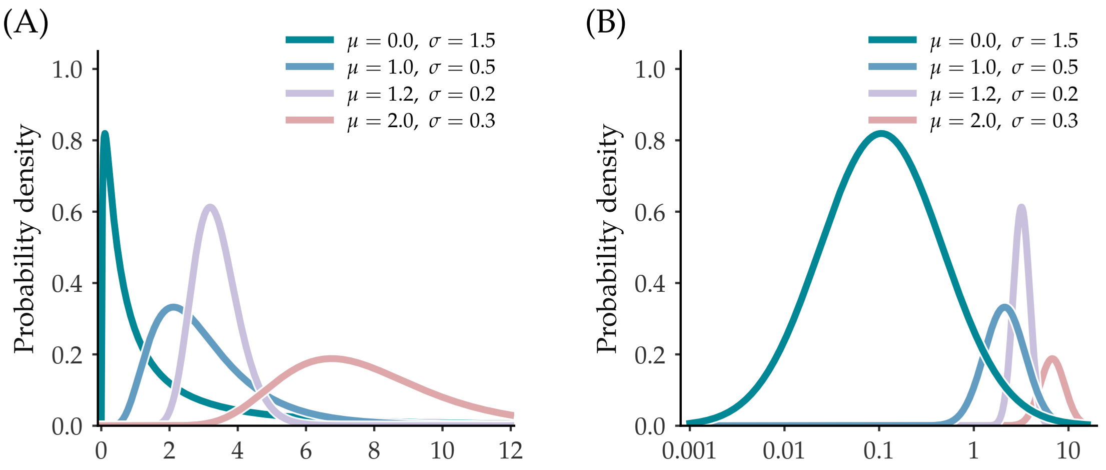


Figure 3.8: Example of lognormal distributions, and their corresponding equivalents on a logarithmic horizontal axis. The lognormal distribution is defined on the positive real line.

(Read Section B.12.1 if you don't know how to transform probability distributions. Note the $\frac{1}{s}$ factor!) This is called the *lognormal distribution* with parameters μ and σ_s^2, also written as Lognormal$(s;\mu,\sigma_s^2)$. Examples are shown in **Fig. 3.8**. Importantly, the parameters do not correspond directly to mean and variance. The mean of the lognormally distributed variable s is $e^{\mu+\frac{1}{2}\sigma_s^2}$ and its variance is a rather complicated expression that we will not use. The median of s under the lognormal distribution is e^{μ} and its mode is $e^{\mu-\sigma_s^2}$.

A lognormal distribution can also be used for the measurement distribution:

$$p(\log x|\log s) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(\log x-\log s)^2}{2\sigma^2}} \tag{3.34}$$

This use has special significance, because an important property of the lognormal distribution is that the standard deviation is proportional to the mean. It turns out that this is empirically found to be a good description of human magnitude judgments – a relation called the *Weber-Fechner law*. For example, the level of noise in the observer's measurement of line length is proportional to the length itself: telling apart a distance of 1.02 m from a distance of 1 m is about as hard as telling apart 10.2 cm from 10 cm. We will extensively use lognormal distributions in Chapter 9.

3.7 Applications

The Bayesian model described in this chapter has been applied to a wide variety of scenarios within the domains of visual perception, somatosensory perception, auditory perception, and motor control. Here, we mention just a few of the studies from this vast literature.

In visual perception, low-contrast stimuli (e.g., a gray object moving against a background of a different shade of gray) appear to move slower than high-contrast stimuli (for example, a white object moving against the same background) that are in reality moving a the same speed (Stone and Thompson, 1992). Weiss, Simoncelli, and Adelson (2002) explained this and many other puzzling visual illusions as resulting from the brain's use of a prior distribution over speed that is higher for lower speeds (a so-called low-speed prior). A lower-contrast object provides less reliable information, which means that the likelihood function is wider. This results in a posterior estimate that is shifted more towards the mean of the prior. If the prior peaks at 0, then the perceived speed will be shifted towards a lower value when contrast is lower. In later work, Stocker and Simoncelli (2006) refined the Bayesian model for visual motion perception by estimating the shape of the prior.

In the realm of somatosensory perception, Goldreich (2007) showed that a variety of tactile spatiotemporal illusions could be explained using the model in this chapter, again with a prior expectation for low-speed movement. The proposed model provides a unified explanation for several startling tactile illusions in which taps delivered in rapid succession to distinct points on the skin surface are perceived to occur closer together than they actually are. Remarkably, the perceived distance between taps shrinks as the time between tap is reduced, a phenomenon known as perceptual length contraction (Goldreich, 2007; Goldreich & Tong, 2013; Tong et al., 2016). Tong et al. (2016) confirmed a prediction of the model; namely, that perceptual length contraction is more pronounced for weaker taps, which evoke broader likelihood functions.

In the realm of auditory perception, Fischer and Peña (2011) put forth a Bayesian model to explain the observation that owls are able to process acoustic stimuli remarkably accurately in order to localize a sound source that occurs in the region straight ahead of the owl, yet they consistently underestimate the position of sound sources that occur in eccentric regions. The authors model this behaviour as resulting from the owl's use of a centrality prior, i.e., the expectation that sound sources tend to occur more frequently in central locations.

Bayesian models have been applied not only to spatial perception but to temporal perception. For instance, the model of Goldreich (2007) provides an explanation for a time perception illusion in which the perceived time between spatially separated stimuli expands as the distance between the stimuli is increased. This time dilation phenomenon (aka kappa effect) occurs in both the tactile and visual domains (Chen et al., 2016). Using visual stimuli, Jazayeri and Shadlen (2010), showed that, when presented with stimuli whose durations were drawn from different distributions, participants' duration estimates were consistent with those of a Bayesian observer model that incorporated the duration distributions as prior probability distributions.

In the domain of movement, Kording and Wolpert (2004) studied finger reaching movements to a visual target. In a virtual-reality set-up, the target position was laterally displaced by a distance randomly drawn from a Gaussian distribution with a mean displacement of 1 cm. This acted as the prior distribution. Observers were trained until they had learned this distribution. The finger's movement was not visible to the subject; a noisy visual measurement of the finger's position was provided by showing a cloud of dots halfway through the movement. Importantly, the reliability of the visual measurement was manipulated on a trial-by-trial basis. Observers were found to combine the noisy measurement with their learned knowledge of the prior distribution in the manner described by Eq. 3.27. In a second experiment, the Gaussian prior was replaced by a bimodal (two-peaked) prior; even this prior was incorporated in good approximation, although this required extensive training.

3.8 Summary and remarks

- We considered inference of a continuous variable based on noisy sensory observations.
- We defined two steps of a Bayesian model: the generative model and the inference process.
- Generative model:
  - Noise in sensory observations has many sources.
  - The measurement serves as an abstraction of the sensory observations.
  - The measurement is noisy and we modeled this noise as Gaussian. This was justified by the central limit theorem.
  - We also assumed a Gaussian stimulus distribution.
- Inference:
  - In the inference, the observer uses the stimulus and measurement distributions from the generative model to form a prior distribution and a likelihood function, respectively.
  - The posterior is the normalized product of prior and likelihood. Under our assumptions, the posterior was also Gaussian.

- The posterior mean is then a weighted average between the observed measurement and the prior mean. In addition, the posterior is narrower than both prior and likelihood, reflecting decreased uncertainty.
- To minimize expected squared error, the Bayesian observer would estimate the stimulus as the posterior mean.
- Confidence can be defined as the value of the posterior at the posterior mean.
- An observer can be Bayesian but use the "wrong" generative model. This is called model mismatch.

3.9 Suggested readings

- Wendy J Adams, Erich W Graf, and Marc O Ernst. "Experience can change the'light-from-above'prior". In: *Nature neuroscience* 7.10 (2004), pages 1057–1058
- Brian J Fischer and José Luis Peña. "Owl's behavior and neural representation predicted by Bayesian inference". In: *Nature neuroscience* 14.8 (2011), pages 1061–1066
- Daniel Goldreich and Jonathan Tong. "Prediction, postdiction, and perceptual length contraction: a Bayesian low-speed prior captures the cutaneous rabbit and related illusions". In: *Frontiers in psychology* 4 (2013), page 221
- Mehrdad Jazayeri and Michael N Shadlen. "Temporal context calibrates interval timing". In: *Nature neuroscience* 13.8 (2010), pages 1020–1026
- Konrad P Körding and Daniel M Wolpert. "Bayesian integration in sensorimotor learning". In: *Nature* 427.6971 (2004), pages 244–247
- Xue-Xin Wei and Alan A Stocker. "A Bayesian observer model constrained by efficient coding can explain'anti-Bayesian'percepts". In: *Nature neuroscience* 18.10 (2015), pages 1509–1517
- Yair Weiss, Eero P Simoncelli, and Edward H Adelson. "Motion illusions as optimal percepts". In: *Nature neuroscience* 5.6 (2002), pages 598–604

3.10 Problems

Problem 3.1 Let s be the stimulus of interest, x the measurement, $p_s(s)$ the stimulus distribution, and $p_{x|s}(x|s)$ the measurement distribution.

a) Write down the posterior distribution over hypothesized stimulus s, given an observed measurement x_{obs}.
b) Which of the terms in your expression is called the likelihood function?
c) What is the difference between the likelihood function and the measurement distribution?

Problem 3.2 Let us numerically calculate the posterior (1). Suppose the stimulus distribution $p(s)$ is Gaussian with mean 20 and standard deviation 4. The measurement distribution $p(x|s)$ is Gaussian with standard deviation $\sigma = 5$. A Bayesian observer infers s from an observed measurement $x_{\text{obs}} = 30$. We are now going to calculate the posterior pdf using numerical methods.

a) Define a vector of hypothesized stimulus values s: $[0, 0.2, 0.4, \dots, 40]$.
b) Compute the likelihood function and the prior on this vector of s values.
c) Multiply the likelihood and the prior (pointwise).
d) Divide this product by its sum over all s (normalization step).
e) Convert this posterior probability mass function into a probability density function by dividing by the step size you used in your vector of s values (e.g., 0.2).
f) Plot the likelihood, prior, and posterior in the same plot.
g) Is the posterior wider or narrower than likelihood and prior? Do you expect this based on the equations we discussed?

h) Change the standard deviation of the measurement distribution to a large value. What happens to the posterior? Can you explain this?

i) Change the standard deviation of the measurement distribution to a small value. What happens to the posterior? Can you explain this?

Problem 3.3 Let us look at multiplying two Gaussians. The stimulus distribution $p(s)$ is Gaussian with mean μ and variance σ_s^2. The measurement distribution $p(x|s)$ is Gaussian with mean s and variance σ^2.

a) Write down the equations for $p(x|s)$ and $p(s)$.

b) A Bayesian observer infers s from a measurement x_{obs}. Use Bayes' rule to write down the equation for the posterior, $p(s|x_{\text{obs}}$. Substitute the expressions for $p(x_{\text{obs}}|s)$ and $p(s)$, but do not simplify yet.

The numerator is a product of two Gaussians. As we discussed in Section 3.3.3, the denominator, $p(x_{\text{obs}})$, is a normalization factor that ensures that the integral equals 1. For now, we will ignore it and focus on the numerator.

c) Apply the rule $e^A e^B = e^{A+B}$ to simplify the numerator.

d) Expand the two quadratic terms in the exponent.

e) Rewrite the exponent to the form $as^2 + bs + c$, with a, b, and c constants. Importantly, since c is just leading to a constant scaling, no need to calculate it.

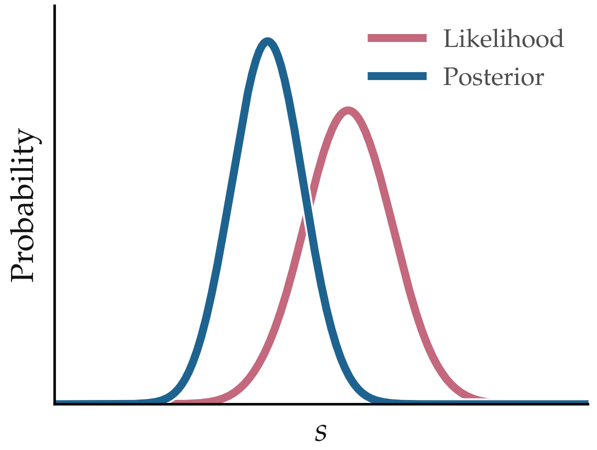
f) Rewrite your expression obtained in (e) in a simpler form $exp(c_1(s+c_2)^2 + const)$. Feel free to use the fact ('completing the square') that any quadratic function of the form $as^2 + bs + c$ can be written as $a\left(s + \frac{b}{2a}\right)^2 + c - \frac{b^2}{4a}$.

g) Now rewrite your expression into the form $e^Z e^{-\frac{(s-\mu_{\text{combined}})^2}{2\sigma_{\text{combined}}^2}}$. Express μ_{combined} and σ_{combined} in terms of x, σ, μ, and σ_s.

h) Recall that $p(s|x_{\text{obs}})$ is a distribution and that its integral should therefore be equal to 1. However, the expression that you obtained in (h) is not properly normalized because we ignored $p(x_{\text{obs}})$. Modify the expression such that it is properly normalized, but without explicitly calculating $p(x_{\text{obs}})$ (Hint: does e^Z depend on s?)

Many Bayesian inference problems involve a product of two or more Gaussians, so this derivation will come in handy later!

Problem 3.4 The figure below shows a likelihood function and a posterior distribution. Both are Gaussian, with $\sigma_{\text{posterior}} = 1.2$ and $\sigma_{\text{likelihood}} = 1.5$.



Assume that the prior is also Gaussian. Which of the following statements is true? Explain. You will need both the plot and the given numbers.

a) The prior is centered to the left of the likelihood function and is narrower.

b) The prior is centered to the left of the likelihood function and is wider than it.

c) The prior is centered to the right of the likelihood function and is narrower than it.

d) The prior is centered to the right of the likelihood function and is wider than it.

Problem 3.5 There are stimulus distribution that do not look like Gaussians but for which we can still calculate the posterior. The stimulus distribution $p(s)$ is 0 for $s < 0$ and an exponential distribution, $p(s) \propto e^{-\lambda s}$ with $\lambda > 0$, for $s \geq 0$. The measurement distribution $p(x|s)$ is Gaussian with mean s and variance σ^2. A Bayesian observer infers s from a measurement x_{obs}. Derive an equation for the posterior mean estimate.

Problem 3.6 Let us numerically calculate the posterior (2). Start with the code from the above numerical example with two Gaussians. Suppose the stimulus distribution $p(s)$ is uniform on the interval $[-15, 25]$ and 0 outside this interval. The measurement distribution $p(x|s)$ is Gaussian with standard deviation $\sigma = 5$. A Bayesian observer infers s from an observed measurement $x_{\text{obs}} = 30$. We are now going to calculate the posterior pdf numerically.

a) What is the value of $p(s)$ on the interval $[-15, 25]$?

b) Repeat parts (a)-(f) from the previous numerical (Gaussians) problem for this new stimulus distribution.

c) Is the posterior Gaussian?

d) Numerically calculate the mean of the posterior.

e) Numerically calculate the variance of the posterior.

Problem 3.7 In this chapter, we assumed that σ is independent of the stimulus s is often violated in real cases, leading to *heteroskedasticity*, where the standard deviations are non-constant. Assume the measurement distribution

$$p(x|s) = \frac{1}{\sqrt{2\pi\sigma(s)^2}} e - \frac{(x-s)^2}{2\sigma(s)^2}, \tag{3.35}$$

with σ a particular function of s: $\sigma(s) = 1 + s$.

a) For $s = 0, 1, 2$, plot the measurement distribution (three curves in one plot, color-coded). All three should look Gaussian.

b) For $x_{\text{obs}} = 0, 1, 2$, plot the likelihood function over hypothesized s (three curves in one plot, color-coded). None of them should look Gaussian.

c) Explain how it is possible that the measurement distributions are all Gaussian but the likelihoods are not.

d) If the prior were Gaussian in this case, would the posterior be? Demonstrate this mathematically.

Problem 3.8 This problem builds on Section 3.4.2.

a) Show that in our case study, Bayesian confidence associated with the PM estimate is $\sqrt{\frac{1}{2\pi}\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}\right)}$.

b) Why would an equally legitimate definition of Bayesian confidence in this case be $\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}$?

c) Does the dependence of either expression on σ and σ_s make sense? Explain.

Problem 3.9 What is confidence if we have a prior mismatch? When the observer uses a wrong prior to compute the posterior (prior mismatch), can decision confidence be higher than when using the correct prior? If so, give an example. If not, prove it.

Problem 3.10 In Section 3.4.1, we defined uncertainty as the standard deviation of a belief distribution. An alternative is as the *entropy* of the distribution. Look up the definition of entropy.

a) Show that for a Gaussian distribution, the definitions are equivalent.

b) Construct an example of two belief distributions with different standard deviations but the same entropy.

c) What are the advantages and disadvantages of either definition?

Problem 3.11 Some stimulus variables that take values on the circle, such as motion direction, which takes values between (for instance) $-\pi$ and π. In B.7.6, we discuss the Von Mises distribution, which is suitable for such variables. Suppose that motion direction is drawn from a Von Mises distribution given by Eq. (B.36). Assume that the measurement distribution is also Von Mises, given by Eq. (B.37).

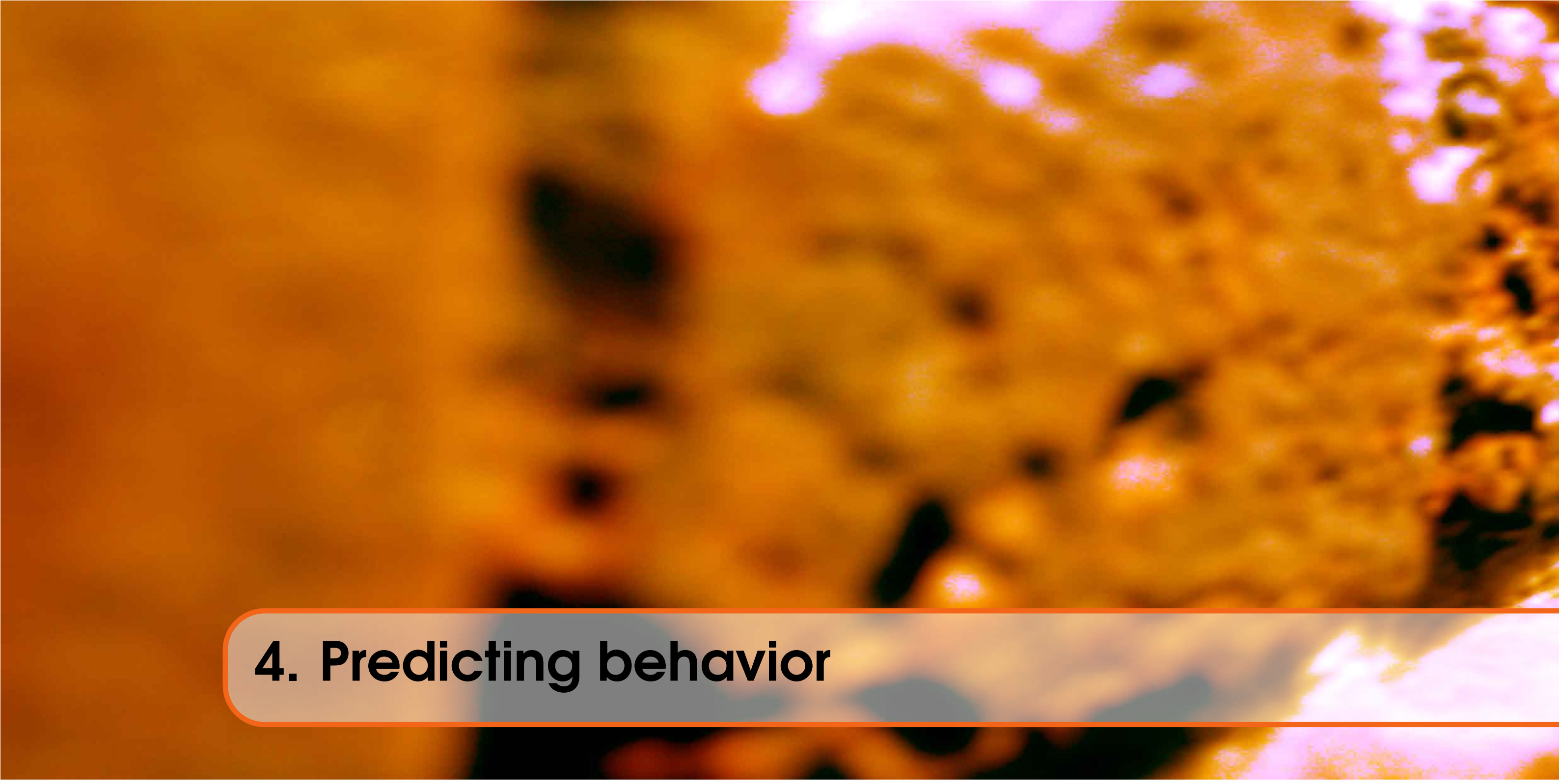
a) Show that the posterior over s, $p(s|x)$, is a Von Mises distribution with (circular) mean posterior $\mu_{\text{posterior}}$ given by

$$\cos \mu_{\text{posterior}} = \kappa_s \cos \mu_s + \kappa \cos x \tag{3.36}$$

$$\sin \mu_{\text{posterior}} = \kappa_s \sin \mu_s + \kappa \sin x \tag{3.37}$$

and concentration parameter $\sqrt{\kappa_s^2 + \kappa^2 + 2\kappa\kappa_s \cos(x - \mu_s)}$.

b) Compare and contrast with the Gaussian case.



4. Predicting behavior

How does a Bayesian model predict a human observer's responses on a perceptual task?

We continue our model of inference under sensory noise from Chapter 3. Ultimately, in experiments we provide stimuli and measure a participant's responses. We thus need to be able to make behavioral predictions. Our model of the observer's inference process (Step 2) predicts the observer's estimate of the world state, given the sensory observations. However, in a psychophysics experiment, the observer's sensory observations are impossible for the investigator to measure, and indeed vary stochastically from trial to trial even when the stimulus is held fixed. Thus, the estimate of the world state (or the observer's behavior) is itself a random variable and has a probability distribution. Therefore, to compare the model's prediction with behavior, we have to compute the probability of each possible estimate in a particular experimental condition.

In Chapter 3, we derived for a simple model with a Gaussian stimulus distribution and a Gaussian measurement distribution that the posterior distribution is also Gaussian, with mean given by Eq. (3.18),

$$\mu_{\text{post}} = \frac{Jx_{\text{obs}} + J_s\mu}{J + J_s}. \tag{4.1}$$

and variance given by Eq. (3.19),

$$\sigma^2_{\text{post}} = \frac{1}{J + J_s}. \tag{4.2}$$

We discussed in Section 3.3.6 that a reasonable way for the observer to estimate the stimulus is the mean of the posterior:

$$\hat{s}_{\text{PM}} = \mu_{\text{post}}. \tag{4.3}$$

The "hat" denotes "estimate" and we refer to this particular estimate as the *posterior mean estimate*. In our simple model, the estimate is also the *response*, i.e. what the observer would respond in an experiment. The topic of the present chapter is Step 3 of Bayesian modeling, in which the probability distribution of the response is computed. This distribution – the estimate distribution or the response distribution – can be compared to experimental data.

Plan of the chapter

We show that, upon repeated presentations of the same stimulus, a Bayesian observer's posterior mean estimate (the observer's response) is a random variable, because it depends on the measurement, which varies stochastically from trial to trial. We derive the probability distribution of the Bayesian observer's responses. This response distribution allows investigators to compare the predictions of the Bayesian perceptual model to the observer's actual behaviour in the context of a psychophysical experiment.

4.1 Inherited variability

In the most basic form of Bayesian modeling, even when the measurement is noisy, everything that comes after the measurement is a deterministic function of the measurement: the likelihood function is a deterministic function of the measurement, the posterior is deterministically computed from the likelihood, and the estimate (e.g. a posterior mean estimate) is deterministically computed from the posterior. Thus, once I give you the value of the observed noisy measurement, x_{obs}, you can compute everything else up to the observer's response[1]. However, the measurement itself is variable from trial to trial even when the true stimulus is kept fixed. This means that from the experimenter's point of view, the above quantities (likelihood, posterior, and posterior read-out) will all vary along with the measurement. **Fig. 4.4** illustrates this variation.

Specifically, because x_{obs} is a random variable for given s, so is the stimulus estimate. Hence, in response to repeated presentations of the same stimulus, the posterior mean estimate will be a random variable with a probability distribution.

> **Box 4.1 — Take-away: Variability of the stimulus estimate.** A stimulus estimate will vary from trial to trial (i.e. will be stochastic) even when the true stimulus is held fixed, because the measurement varies from trial to trial. ■

Nowhere in our model have we added extra noise beyond the Gaussian noise that we started out with. The mapping from measurement to stimulus estimate, is completely deterministic. The stochasticity in the stimulus estimate is *inherited from* the stochasticity in the measurement x_{obs}*.

The experimenter can only control the stimulus. To compare our Bayesian model with an observer's behavior in a psychophysical task, we need to specify what the Bayesian model predicts for the observer's responses when the true stimulus is s. In other words, we need to know the distribution of the stimulus estimate when the true stimulus is s. We denote this distribution $p(\hat{s}|s)$, and call it the *estimate distribution* or the *response distribution*.

4.2 The response distribution

From Step 1, we know that when the true stimulus is s, x follows a Gaussian distribution with mean s and variance σ^2. Moreover, we know that the random variable $\hat{s}$ is linearly related to the random variable x. Using the properties of random variables with Gaussian distributions (see Box 4.2 and Exercise 4.1), one can show that when the true stimulus is s, the stimulus estimate $\hat{s}_{\text{PM}}$ follows a Gaussian distribution with mean

$$\mathbb{E}[\hat{s}_{\text{PM}}|s] = \frac{Js + J_s\mu}{J + J_s} \tag{4.4}$$

and variance

$$\text{Var}[\hat{s}_{\text{PM}}|s] = \frac{J}{(J + J_s)^2}, \tag{4.5}$$

[1] This is no longer true in the presence of response or decision noise

or in other words,

$$p(\hat{s}_{\mathrm{PM}}|s) = \mathcal{N}\left(\hat{s}_{\mathrm{PM}}; \frac{Js + J_s\mu}{J + J_s}, \frac{J}{(J+J_s)^2}\right). \tag{4.6}$$

Box 4.2 — Properties of linear combinations of random variables. The following properties are useful:

1. **General:** If a random variable X has mean μ and variance σ^2, and a and b are constants, then the random variable $aX + b$ has mean $\mu + b$ and variance $a^2\sigma^2$.
2. **General:** If random variables X and Y are independent and have means μ_X and μ_Y, and variances σ_X^2 and σ_Y^2, respectively, then the random variable $X + Y$ has mean $\mu_X + \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$.
3. **Specific to Gaussian distributions:** If X and Y have a Gaussian distribution, then $aX + b$ and $X + Y$ also have Gaussian distributions.

■

Exercise 4.1 In many cases we will have random variables that are made up of weighted sums of other random variables. Let us get a bit of an intuition into how this affects means and variances of the sums:

a) Combine the first two properties in Box 4.2 to show that the random variable $aX + bY$ has mean $a\mu_X + b\mu_Y$ and variance $a^2\sigma_X^2 + b^2\sigma_Y^2$.

b) Using this result, prove Eqs. (4.4) and (4.5).

■

The variance of the posterior mean estimate differs from any variance we have encountered so far: from the variance of the measurement distribution, from the variance of the likelihood function, and from the variance of the posterior distribution. Table 4.1 might help to keep the different distributions apart.

The computation of the distribution of the posterior mean estimate completes our Bayesian model. Recapitulating, the Bayesian model consisted of three steps. We first formulated the generative model, which described how a measurement is randomly generated on a given trial. We then "inverted" the generative model, which meant computing the posterior distribution over the variable of interest, s, given the measurement x. We assumed that the observer reads out the posterior distribution by picking its mean: the posterior mean estimate. Finally, we computed the distribution of the posterior mean estimate across many trials when the true stimulus is held fixed (see **table 4.1**.

4.3 Variance of the posterior mean estimate

Note that the variance of the posterior mean estimate, $\mathrm{Var}(\hat{s}_{\mathrm{PM}}) = \frac{J}{(J+J_s)^2}$, is different from the variance of the posterior (from Eq. (4.2)), $\sigma^2_{\mathrm{post}} = \frac{1}{J+J_s}$. Conceptually, this is possible because they have a completely different meaning: the former is the variability of behavior as measured by the experimenter, the latter is the uncertainty of the observer on a given trial. In general, these two quantities will be different.

The difference becomes clear when we plot both variances as a function of the measurement noise level (**Fig. 4.5**). As the sensory noise level increases, the posterior gets wider and wider. Initially, the same holds true for the distribution of the posterior mean estimate. However, when σ grows large enough, the standard deviation of the posterior mean estimate will decrease again.

	Mean	Variance
Step 1: Generative model	Stimulus distribution, $p(s)$	
	μ	σ_s^2
	Measurement distribution, $p(x\mid s)$	
	s	σ^2
Step 2: Inference	Prior distribution, $p(s)$	
	μ	σ_s^2
	Likelihood function $L(s;x) = p(x\mid s)$	
	x	σ^2
	Posterior distribution $p(s\mid x)$	
	$\frac{xJ+\mu J_s}{J+J_s} = \hat{s}$	$\frac{1}{J+J_s}$
Step 3: Response distribution	Response distribution, $p(\hat{s}\mid s)$	
	$\langle \hat{s} \rangle = \frac{sJ+\mu J_s}{J+J_s}$	$\frac{J}{(J+J_s)^2}$

Table 4.1: Means and variances of all distributions discussed in this chapter. Note: here we adopt the "precision notation" where $J = \dfrac{1}{\sigma^2}$ and $J_s = \dfrac{1}{\sigma_s^2}$.

Exercise 4.2 Why does this make intuitive sense? ■

4.4 Maximum-likelihood estimation

Instead of using the posterior mean as the estimate of the stimulus, the observer could instead simply use the measurement itself. In our example, the measurement is also the *maximum-likelihood estimate*, since the likelihood function $\mathscr{L}(s;x) = p(x_{\text{obs}}|s)$ peaks at x_{obs}:

$$\hat{s}_{\text{ML}} = x_{\text{obs}}. \tag{4.7}$$

The maximum-likelihood estimate is the estimate that ignores the prior altogether. Alternatively, $\hat{s}_{\text{ML}}$ can be regarded as a posterior mean estimate for an observer who erroneously assumes the prior to be flat (i.e. σ_s to be infinite).

Just as we studied the distribution of the posterior mean estimate $\hat{s}_{\text{PM}}$ for given s, we can also study the distribution of the MLE $\hat{s}_{\text{ML}}$ for given s. But we already know this distribution, as it is equal to the measurement distribution $p(x|s)$: Gaussian with mean s and variance σ^2; only, the argument is now $\hat{s}_{\text{ML}}$ instead of x.

Examples of $\hat{s}_{\text{ML}}$ and $\hat{s}_{\text{PM}}$ for stimuli s randomly drawn from a stimulus distribution $p(s)$ are shown in **Fig. 4.1**. This shows that the MLE is on average equal to the stimulus, whereas the posterior mean estimate tends to lie closer to the mean of the prior. The higher the sensory noise, σ, the closer the posterior mean estimate tends to lie to the mean of the prior.

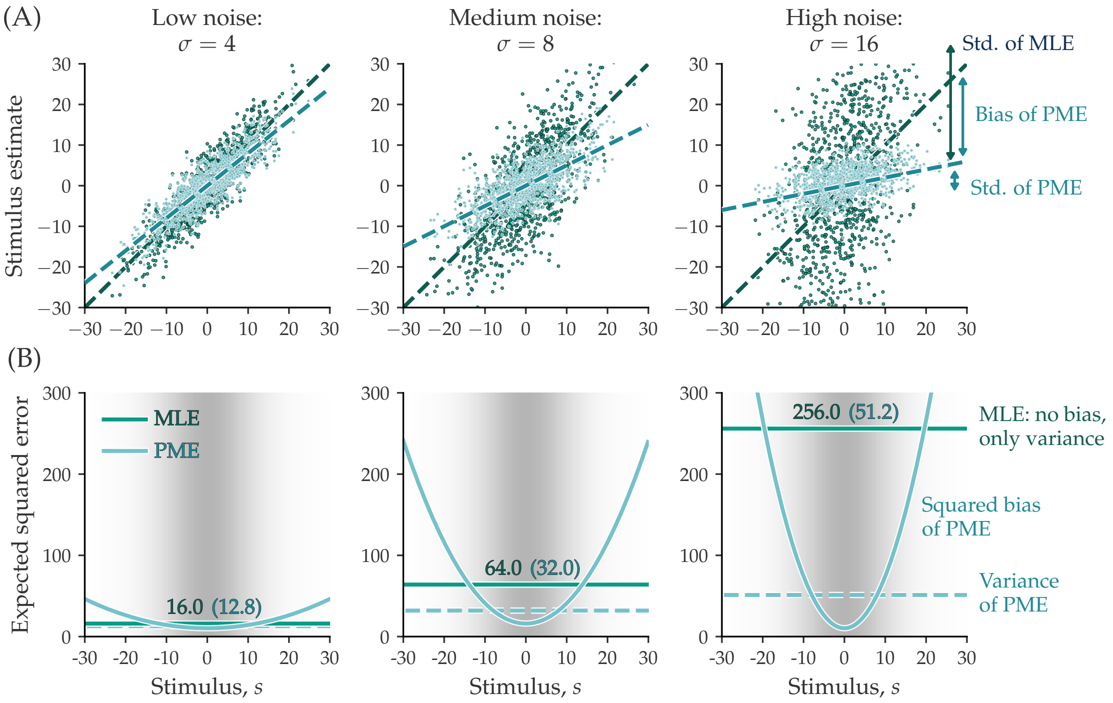


Figure 4.1: Comparison between the posterior mean estimate (PME) and the maximum-likelihood estimate (MLE). In this example, the stimulus distribution has $\mu = 0$ and $\sigma_s = 8$. **(A)** Scatterplots of PMEs and MLEs against the true stimulus. Dashed lines indicate the expected values. The larger the noise, the lower the slope of the expected value of the PME. **(B)** mean squared error as a function of the stimulus for the PMEs and MLEs. mean squared error (solid lines) is the sum of squared bias and variance. Although the PME is biased, its variance (dashed light blue line) is lower than that of the MLE (green line). The stimuli that occur often according to the stimulus distribution (shading indicates probability) are such that the "overall" (stimulus-averaged) mean squared error of the PME (light blue number, in parentheses) is always lower than that of the MLE (green number).

4.5 Bias and mean squared error

We saw that the posterior mean estimate is a weighted average between the measurement and the mean of the Gaussian stimulus distribution. Therefore, the mean posterior mean estimate is a weighted average between the true stimulus and the mean of the stimulus distribution. As a consequence, the mean of the optimal estimate is not equal to the true stimulus: the posterior mean estimate is *biased* from the stimulus towards the mean of the prior.

Definition 4.5.1 — Bias. The bias of an estimate $\hat{s}$ is defined as the difference between the average estimate and the true stimulus, or equivalently, as the average difference between the estimate and the true stimulus:

$$\text{Bias}[\hat{s}|s] = \mathbb{E}[\hat{s}|s] - s. \tag{4.8}$$

Notation 4.1. *The notation $\mathbb{E}[\hat{s}|s]$ indicates the expected value of $\hat{s}$ under the conditional distribution $p(\hat{s}|s)$. The expected value of a random variable with a conditional probability distribution is also called a conditional expectation. Different notations are used for indicating that an expected value is conditional. A common but – in our opinion – highly confusing notation is to use the variable that is being conditioned on, here s, as a subscript, so that $\mathbb{E}[\hat{s}|s]$ would be written as $\mathbb{E}_s[\hat{s}]$. In this*

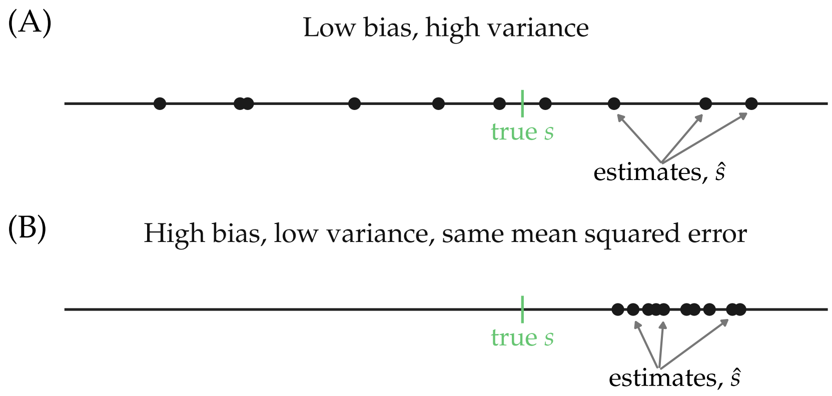


Figure 4.2: The same mean squared error can be produced by an estimator with low bias and high variance (A) and by estimator with high bias and low variance (B).

book, we only use a subscript under a p, $\mathbb{E}$, Var etc. to indicate in case of ambiguity the random variable that is the main argument (see also Box 3.1).

Exercise 4.3 a) Calculate the bias of the maximum likelihood estimate (MLE) $\hat{s}_{\text{ML}}$ for a given s.

b) Calculate the bias of the posterior mean estimate $\hat{s}_{\text{PM}}$ for given s. ■

Box 4.3 — Squared errors. Let s be the stimulus and $\hat{s}$ the stimulus estimate on a given trial. The squared error on a single trial is then $(\hat{s}-s)^2$. The *mean squared error* of $\hat{s}$ for given s is

$$\text{MSE}[\hat{s}|s] \equiv \mathbb{E}_{\hat{s}|s}[(\hat{s}-s)^2|s]. \tag{4.9}$$

This quantity depends on s. The "overall" mean squared error across all s in addition involves an expected value over s:

$$\begin{aligned} \text{MSE}[\hat{s}] &\equiv \mathbb{E}_{\hat{s},s}[(\hat{s}-s)^2] && (4.10)\\ &= \mathbb{E}_s\left[\mathbb{E}_{\hat{s}|s}[(\hat{s}-s)^2|s]\right] && (4.11)\\ &= \mathbb{E}_s\left[\text{MSE}[\hat{s}|s]\right] && (4.12) \end{aligned}$$

This quantity does *not* depend on s. ■

Bayesian modelers often say that Bayesian estimation is optimal. But the posterior mean estimate (and pretty much every other Bayesian estimate) is biased – its mean is different from the true stimulus. This seems contradictory: wouldn't it always be better to have an unbiased estimate?

To resolve this paradox, we have to realize that bias is not all that matters. It turns out that the posterior mean estimator is good in the sense that it minimizes the overall *mean squared error* between the estimate and the true stimulus. Mean squared error has two components: not only bias, but also variance. Both bias and variance are bad, in different ways. **Fig. 4.2** shows how the same expected squared error can arise from low bias and high variance, and from high bias and low variance.

It turns out that the posterior mean estimate is optimal in that it minimizes the overall mean squared error between the estimate and the true stimulus, MSE$[\hat{s}|s]$ (see **Box 4.3**). When the measurements are very noisy, you will be off by least if you pick the mean of the stimulus distribution, μ. If the measurement x is noiseless, you should estimate the stimulus at x. It makes sense that for intermediate noise levels, a strategy of picking an estimate in between μ and x will cause you to be off by the smallest possible amount on average. Thus, an optimal strategy would involve a bias towards the mean of a stimulus distribution.

We can derive an expression for the overall mean squared error of the posterior mean estimate. We will compute the expected value in two steps: first over $\hat{s}$ for a given s, then over s:

$$\mathrm{MSE}[\hat{s}] = \mathbb{E}_s\left[\mathrm{MSE}[\hat{s}|s]\right]. \tag{4.13}$$

We first compute the inside expected value, the mean squared error of $\hat{s}$ given s. This can be rewritten as

$$\mathrm{MSE}[\hat{s}|s] = \mathrm{Bias}[\hat{s}|s]^2 + \mathrm{Var}[\hat{s}|s]. \tag{4.14}$$

This relation is also called the *bias-variance decomposition of the mean squared error*. If $\hat{s}$ is the posterior mean estimate, we can simply substitute the expressions for, respectively:

$$\begin{aligned}\mathrm{MSE}[\hat{s}|s] &= \left(\frac{Jx_{\mathrm{obs}} + J_s\mu}{J+J_s} - s\right)^2 + \frac{J}{(J+J_s)^2} &(4.15)\\ &= \frac{J_s^2(\mu-s)^2}{(J+J_s)^2} + \frac{J}{(J+J_s)^2}. &(4.16)\end{aligned}$$

This can be further simplified, but we keep this form because now the first term is still the square of the bias for given s, and the second is the variance for given s (but as it happens, independent of s). We have plotted the resulting stimulus-dependent mean squared error as a function of s (**Fig. 4.2**). The variance is independent of s, and the bias grows linearly with the distance between s and the mean of the stimulus distribution (this mean is 0 in the plot), so the squared bias grows quadratically with that distance. For comparison, we also plotted the mean squared error of the measurement. This estimate is unbiased, since $\mathbb{E}[\hat{s}_{\mathrm{ML}}|s] = s$, and its variance is constant at $\mathrm{Var}[\hat{s}_{\mathrm{ML}}|s] = \sigma^2$. (It can be shown that the MLE is the estimate that generally has the lowest variance of all *unbiased* estimates.)

The plot allows us to understand why the posterior mean estimate is optimal even though it is biased. For any s, its variance is lower than the variance of the MLE. The squared bias gets added to the variance, but their sum still stays below the variance of the MLE as long as the stimulus is close enough to the mean of the stimulus distribution. Of course, that is exactly where most stimuli will be distributed, since the stimulus distribution is Gaussian (shown as a gray shading). This is why when we evaluate the outside expected value in Eq. (4.13) to calculate the overall mean squared error, we end up with a lower number for posterior mean estimate than for the MLE. We work this out in a problem.

Intuitively, the reason for the optimal estimate to be biased is that if the measurements are very noisy, the best possible guess of the stimulus would be the mean of the stimulus distribution. If the measurement is less noisy, the mean of the stimulus distribution will be less useful. Mathematically, a strategy where the prior mean and the likelihood mean x are combined will actually produce greater rewards in the long run. Thus, a bias towards the mean of a prior is actually a sign of an optimal strategy.

Indeed, as pointed out by E.T. Jaynes, the word "bias," because it carries negative connotations when used in everyday speech, is an unfortunate choice of terminology: "When we call the quantity...‘bias’, that makes it sound like something awfully reprehensible, which we must get rid of at all costs. If it had been called instead the ‘component of error orthogonal to the variance’,...it would have been clear to all that these two contributions to the error are on an equal footing...This is just the price one pays for choosing a technical terminology that carries an emotional load, implying value judgments..." (Jaynes, 2003, p. 514).

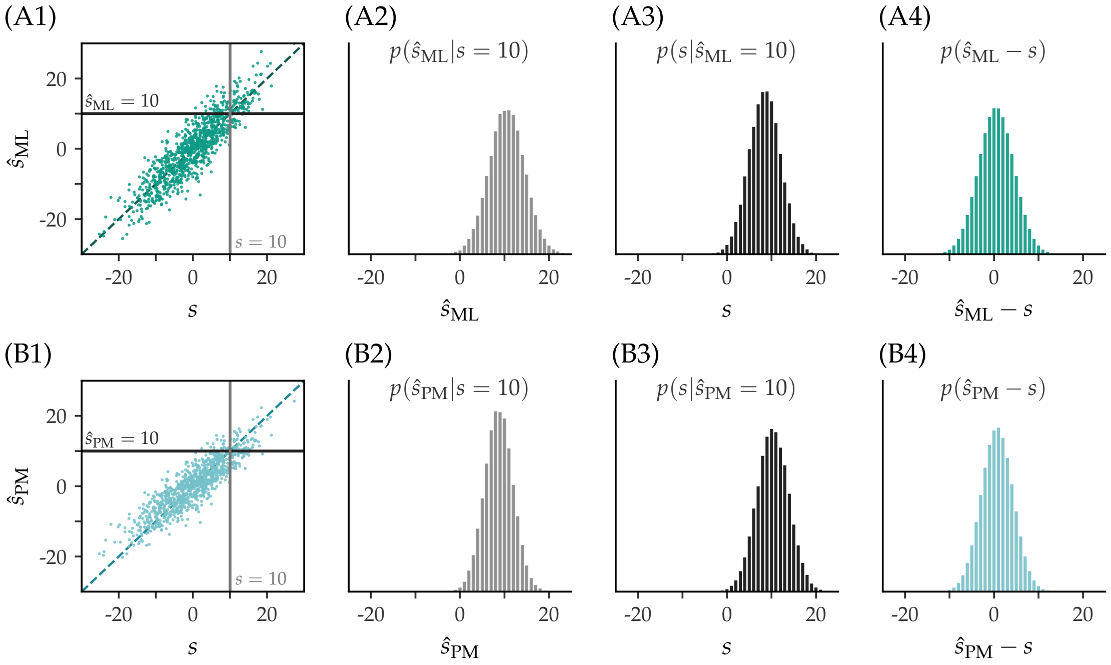


Figure 4.3: The posterior mean estimate (PME, $\hat{s}_{\rm PM}$) is more accurate than the maximum-likelihood estimate (MLE, $\hat{s}_{\rm ML}$). **(A1)** 1,000,000 stimuli were drawn from a normal distribution ($\mu=0$, $\sigma_s=8$), each yielding a noisy measurement ($\sigma=4$). Scatterplot of s vs. $\hat{s}_{\rm ML}$. Diagonal line, $\hat{s}_{\rm ML}=s$; vertical line, $s=10$; horizontal line, $\hat{s}_{\rm ML}=10$. **(A2)**. At each s, $\hat{s}_{\rm ML}$ is distributed with mean s and variance $\sigma^2=16$. The distribution $p(\hat{s}_{\rm ML}|s=10)$ is shown for illustration. **(A3)** For a given $\hat{s}_{\rm ML}$, the distribution of s is not centered on $\hat{s}_{\rm ML}$. The distribution $p(s|\hat{s}_{\rm ML}=10)$ is shown for illustration; this is the posterior distribution over s given $x=10$. **(A4)** For every s, $p(\hat{s}_{\rm ML}-s|s)$ has mean 0 and variance 16. Therefore, the overall error distribution, $p(\hat{s}_{\rm ML}-s)$, also has mean 0 and variance 16. Because this error distribution has mean 0, its variance is MSE$[\hat{s}_{\rm ML}]$. **(B1)** Scatterplot of s vs. $\hat{s}_{\rm PM}$ for the same 1,000,000 data points shown in (A). Diagonal line, $\hat{s}_{\rm PM}=s$; vertical line, $s=10$; horizontal line, $\hat{s}_{\rm PM}=10$. **(B2)** The PME is biased towards the prior mean; $p(\hat{s}_{\rm PM}|s=10)$ is shown for illustration. **(B3)** $p(s|\hat{s}_{\rm PM})$ is the posterior distribution over s, given the measurement that resulted in $\hat{s}_{\rm PM}$. Thus, $p(s|\hat{s}_{\rm PM})$ has mean $\hat{s}_{\rm PM}$ and variance 12.8; $p(s|\hat{s}_{\rm PM}=10)$ is shown for illustration. **(B4)** For every $\hat{s}_{\rm PM}$, $p(\hat{s}_{\rm PM}-s|\hat{s}_{\rm PM})$ has mean 0 and variance 12.8. Therefore, the overall error distribution, $p(\hat{s}_{\rm PM}-s)$, also has mean 0 and variance 12.8. Because this error distribution has mean 0, its variance is MSE$[\hat{s}_{\rm PM}]$. In conclusion, MSE$[\hat{s}_{\rm PM}]<$ MSE$[\hat{s}_{\rm ML}]$, because the variance of the posterior distribution is less than the variance of the noise distribution.

4.5.1 An “inverted bias” perspective

Because it is such an important point to understand, let’s now return to the question: How can a biased estimator be optimal? To answer this question, we will view the scatterplot (**Fig. 4.1A**) from a different perspective (**Fig. 4.3**). We will see that the mean difference between the PME and the stimulus across all trials is in fact zero. The mean difference between the MLE and the stimulus across all trials is also zero, but the MLE is distributed around the stimulus with greater variance. Thus, the PME is a more accurate estimator than the MLE.

Fig. 4.3 plots one million stimuli sampled from a stimulus distribution ($\mu=0$, $\sigma_s=8$), each yielding a noisy measurement ($\sigma=4$). Part **A** shows the statistics for the MLE, and Part **B** for the PME. For a given stimulus value, s, the MLE is distributed with mean s (panel **A2**); thus, the MLE is unbiased. By contrast, for a given s, the PME is not distributed with mean s (panel **B2**); thus, the PME is biased. Crucially, however, on trials when the observer reports a particular PME, the stimulus is indeed centered on that estimate (panel **B3**), and this is not the case for trials with a particular MLE (panel **A3**). In essence, the *stimulus* is biased with respect to the *MLE*. In contrast, the stimulus is unbiased with respect to the PME. Most importantly, while the mean difference across all trials between either estimate and the stimulus is zero, the mean squared error is lower for the PME (panel **B4**) than for the MLE (panel **A4**).

To further understand the above, note that, given a particular MLE (i.e., a particular measurement, x), the distribution of stimuli is the posterior distribution, $p(s|x)$. The posterior distribution is not centered on the measurement; rather, it is shifted away from the measurement and towards the prior mean (Eq. (4.1)). Consequently, the stimulus is biased with respect to the MLE. In contrast, the stimulus is unbiased with respect to the PME. To understand why, note that each MLE (i.e., each measurement) maps onto a particular PME (Eq. (4.1)). Consequently, each horizontal row of dots in panel A1 is simply displaced vertically towards the prior mean (0, in this example) in panel B1, such that its position equals the PME. For instance, the reader may verify from Eq. (4.1) that the row at MLE $= 10$ in panel A1 is shifted to PME $= 8$ in panel B1. Since panel B1 plots the posterior mean on the y-axis, it necessary follows that the mean of the distribution of stimuli in each horizontal row (given sufficiently large samples) coincides with the identity line.

Following this line of reasoning, we can see that the variance of each horizontal row of stimuli in both panel A1 and Panel B1 is simply the variance of the posterior distribution (Eq. (4.2)). Since the variances of the horizontal rows are equal in panels A1 and B1, but the stimulus is biased only with respect to the MLE, it follows immediately that the mean-squared error of the MLE must be greater than that of the PME. In fact, without doing additional math, we can see that the overall mean squared error of the PME equals the variance of the posterior PDF, whereas the overall mean squared error of the MLE equals the variance of the noise distribution. To see this, let's accumulate the errors $(\text{PME} - s)$ from all of the horizontal rows in panel B1. As the distribution of stimuli is centered in each row on the PME, the resulting overall error distribution has mean zero (panel B4). Consequently, the variance of this distribution (i.e., the variance of the posterior distribution) is the overall mean squared error of the PME. Following a similar line of reasoning, we can accumulate the errors $(\text{MLE} - s)$ from all vertical columns in panel A1. As the distribution of MLE is centered in every column on s, the resulting overall error distribution has mean zero (panel A4). Consequently, the variance of this distribution (i.e., the variance of the noise distribution) is the overall mean squared error of the MLE.

4.5.2 A discrete detour

So far, we have considered only continuous variables for comparing posterior-based estimation with maximum-likelihood estimation. How would this comparison play out for discrete (categorical) variables? We consider in particular *nominal* variables, where the categories have labels but no particular ordering. (Ordinal discrete variables are somewhere in between nominal variables and continuous variables.) For nominal variables, the mean of the posterior does not make sense, because there is no underlying space. Instead, as we already discussed in Chapters 1 and 2, the Bayesian observer would pick the category with the highest posterior probability, or, in other words, perform *maximum-a-posteriori* (MAP) estimation. Like the posterior mean estimate in the continuous case, the MAP is biased towards more frequent values, and therefore, it is equally valid to ask if this bias is “worth it".

To examine this question, let's consider the following example: A particular country consists of three states s. The table below shows, for each of the states, the proportions of workers in various occupations, x. These proportions are likelihoods $p(x = \text{occupation}|s)$. Now suppose that, as part of a public-relations campaign, a person is randomly-selected from the country to win an all-expenses-paid vacation. Given this person's occupation, what state do we think they are from?

Maximum-likelihood estimation would lead us to report that, if this person happens to be a teacher, they are from State 1; if they happen to be a farmer, they are from State 3; and if they happen to be a retail worker, they are from State 2 (highlighted, see **table 4.2**).

	$s = 1$	$s = 2$	$s = 3$	% correct ML
Total population (1000s)	19771 (79.0%)	3446 (13.8%)	1805 (7.2%)	estimator
$x =$ teacher	**1.5%**	0.8%	1.1%	86%
$x =$ farmer	0.4%	2.6%	**3.6%**	28%
$x =$ retail	8.4%	**9.3%**	8.3%	15%
$x =$ other	**89.7%**	87.3%	87.0%	79%

Table 4.2: Percentages of workers in a hypothetical three-state country, computed state by state. Boldfaced per row is the maximum-likelihood decision if a randomly selected person has the indicated occupation.

It is easy to see what is wrong with this reasoning. The three states have vastly different populations (see second row in the table; numbers are in thousands), and it is more probable, *a priori*, that a randomly selected citizen will be from a more populous state. Upon consideration, it is clear that, if a randomly selected citizen is a farmer, in order to maximize the probability of being correct about that person's state-of-residence, it is not sufficient to consider only the *proportion* of each state's population that are farmers. Instead, we must calculate the *absolute number* of farmers in each state, i.e. the proportion of the state's population that are farmers, multiplied by the state's total population. These numbers are computed in **table 4.3**.

	$s = 1$	$s = 2$	$s = 3$	% correct MAP estimator
$x =$ teacher	**297**	28	20	86%
$x =$ farmer	79	**90**	65	38%
$x =$ retail	**1661**	320	150	78%
$x =$ other	**17735**	3008	1570	79%

Table 4.3: Same data in absolute numbers. Boldfaced per row is the MAP decision if a randomly selected person has the indicated occupation. The probability that the MAP decision is correct is obtained by dividing the bold number in each row by the sum of numbers in the row. This is also the posterior probability of the MAP estimate. For example, $p(s = 2|x = \text{farmer}) = \frac{90}{79+90+65} = 0.38$.

The most probable state-of-origin for a person of a given occupation is the state that has the greatest number of people in that occupation. This gives a partially different set of decisions (highlighted). Whereas the best guess for the teacher's home state is unchanged, the best guess for the farmer's home state is now State 2, and the best guess for the retail worker's home state is now State 1. This is MAP estimation, and we see in the right column of the table that for the farmer and the retail worker, the MAP estimator is more accurate than the ML estimator.

Although MAP estimation is thus "biased" towards states with higher populations, this bias is entirely rational and indeed optimal. Note also that the MAP estimator does not *always* go with the state with the highest population. In the case of the farmer, the proportion of farmers in State 2 is large enough to overcome the overall population advantage of State 1. This indicates that the observation, acting through the likelihood function, can be strong enough to overcome the prior.

4.6 Other estimates

If we had not heard about Bayesian inference, we might have thought that a reasonable estimate of the stimulus would be the average of the mean of the stimulus distribution, μ, and the measurement

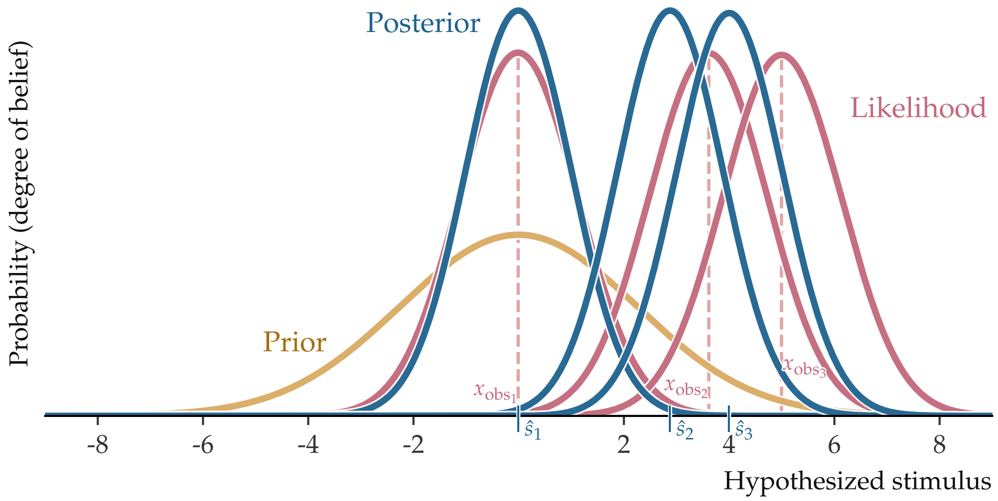


Figure 4.4: Likelihood functions (red) and corresponding posterior distributions (blue) on three example trials. The prior distribution is shown in in grey. The message here is that the likelihood function, the posterior distribution, and the maximum-a-posteriori estimate are not fixed objects: they move around from trial to trial because the measurement x does. Of interest to a Bayesian modeler (in step 3) are the mean and the variance over trials of the maximum-a-posteriori estimate.

x_{obs}.

$$\hat{s}_{\text{average}} = \frac{x_{\text{obs}} + \mu}{2}. \tag{4.17}$$

Another way to view this "average estimate" is that it is the posterior mean estimate under the wrong assumption that σ is equal to σ_s. In general, non-Bayesian estimates can often be interpreted as posterior mean estimates under a wrong assumption about the generative model (Step 1). However, no estimate can ever have a lower mean squared error than the posterior mean estimate (that uses the correct generative model); we explore this in a problem.

4.7 Misconceptions

To build a Bayesian model, we must formulate our generative model and do proper inference on it. Some aspects of this inference are counterintuitive, and scientists doing research on Bayesian models occasionally become confused about the resulting relations between variables. Here we address several misconceptions that may arise during Bayesian modeling.

"The likelihood function is determined by the stimulus." One misconception is that a Bayesian observer's likelihood function is determined by the true stimulus, and therefore it is always the same function as long as the stimulus is held fixed. In reality, we saw in **Fig. 4.4** that the likelihood function varies from repetition to repetition because it is based on the measurement, so a single value of the stimulus gives rise to a different likelihood function whenever the stimulus is given. The misconception arises from confusing the likelihood distribution with the noise distribution. The noise distribution *is* indeed constant as it generally is assumed to depend only on the stimulus. However, the likelihood function always varies from trial to trial.

A related misconception is that the posterior distribution is determined by the true stimulus. In reality, the posterior also varies from trial to trial even when the stimulus is fixed (**Fig. 4.4**).

The response distribution is equal to the likelihood function. Sometimes one reads something like, "We plot the likelihood of the response (or estimate, or percept)". In a Bayesian model, the response is an estimate read out from the posterior distribution, such as the PME estimate. We have already seen that the PME estimate is in general different from the measurement. The distribution of the measurement is the noise distribution, which gives rise to the likelihood function but has a different argument. The likelihood function is a function that indicates the observer's beliefs about the stimulus as derived from the sensory measurement(s) on an individual trial, while the distribution of the estimate is the distribution of the observer's estimate of the stimulus (which is derived from the likelihood and the prior) across many trials. Correct sentences would be "We plot the distribution of the estimate" (if the x-axis represents the PM estimate), "We plot a likelihood function over the stimulus on a particular trial" (if the x-axis represents the hypothesized stimulus), or "We plot the distribution of the measurement" (if the x-axis represents the measurement).

The response distribution is the product of the noise distribution and the prior. Here is another tempting mistake: "Assuming that an observer reports the posterior mean estimate of the stimulus, to obtain the probability density of an observer's response for a given stimulus, I multiply the noise distribution $p(x|s)$ with the prior probability density. This is correct because it will give me a density function that is centered in between the prior mean and the true stimulus." This misconception frequently gets combined with the first one, confusing the noise distribution with the likelihood function. Then the multiplication statement seems even more correct, even though it is not.

One might think that it is possible to take a shortcut by immediately calculating the estimate distribution instead of going through Steps 2 and 3 of the modeling process in sequence. This argument, which is a more sophisticated version of the argument in the previous subsection, could go as follows: "The ML estimate of the stimulus is equal to the measurement x [true]. Thus, the distribution of the ML estimate for a given true stimulus s is equal to the noise distribution, with s [true]. Therefore the PME estimate distribution given a stimulus s can be obtained by multiplying the measurement distribution with the prior [false]." In math, this reasoning would amount to

$$\text{(wrong)} \quad p(\hat{s}|s) \propto p_{x|s}(\hat{s}|s)p_s(\hat{s}), \tag{4.18}$$

where $p_{x|s}(\hat{s}|s)$ is the measurement distribution $p_{x|s}$ evaluated at $\hat{s}$, and $p_s(\hat{s})$ is the prior distribution evaluated at $\hat{s}$.

To illustrate this reasoning and why it is faulty, we could substitute the distributions we used in Chapter 3 and the current chapter, $p_s(s) = \mathcal{N}(s;\mu,\sigma_s^2)$ and $p_{x|s}(x|s) = \mathcal{N}(x;s,\sigma^2)$. After normalization, this would give a Gaussian distribution with a mean of $\frac{Js+J_s\mu}{J+J_s}$ and a variance of $\frac{1}{J+J_s}$. The mean would be correct, per Eq. (4.4). The variance, however, would differ from the correct variance of $\frac{J}{(J+J_s)^2}$ from Eq. (4.5). An intuitive reason why the answer must be wrong can be obtained in the zero-reliability limit ($\sigma \to \infty$). Then, the observer will always estimate the stimulus to be the mean of the prior, and thus the variance of the estimate distribution should be 0. The wrong argument would give a variance of σ_s^2. More formally, the principal mistake made here is that Eq. (4.18) is not a correct application of Bayes' rule. First of all, it does not have the mathematically correct form $p(y|x) \propto p(x|y)p(y)$. Moreover, in Bayesian modeling of behavior, Bayes' rule does not act at the level of estimates over many trials, but at the level of a single trial. Therefore, both components on the right-hand side of Eq. (4.18) are incorrect. The likelihood function should not be a function of the ML estimate or the measurement, but of the hypothesized value of the stimulus s, again on a single trial. The argument of the prior distribution is not the stimulus estimate, but the hypothesized stimulus, again on a single trial.

When the measurement is equal to the true stimulus, the response distribution is equal to the posterior. The next misconception we consider is the following: "For the distribution of the observer's PM estimate for a given stimulus, I can simply use the 'typical' posterior, the one obtained when the measurement happens to be equal to the true stimulus s. It will give me a distribution that is centered in between the prior mean and the true stimulus."

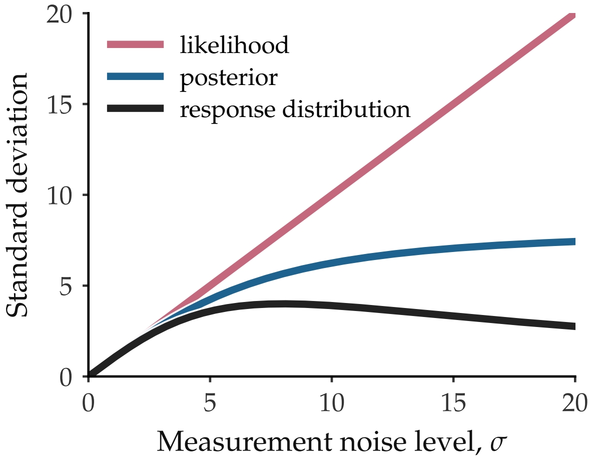


Figure 4.5: The standard deviation of the likelihood function, the posterior, and the response distribution as a function of the measurement noise level, when the stimulus distribution has a standard deviation of $\sigma_s = 8$. As σ grows very large, the standard deviation of the posterior will converge to the standard deviation of the prior, σ_s, while the standard deviation of the response distribution will eventually decrease to 0.

This mistake is mathematically identical to the mistake of the previous subsection but arrived at by following a slightly different line of reasoning. Suppose we correctly calculate the posterior distribution, $p(s|x)$. Now we substitute the true stimulus for x: $p(s|x = s)$. This is a legitimate, though not particularly meaningful, function of s: it reflects the observer's beliefs about the stimulus when the measurement x just happens to coincide with the true stimulus. The final step of the faulty argument would be to regard the distribution $p(s|x = s)$ as the response distribution, $p(\hat{s}|s)$. We further examine this in Problem 4.11.

In our Gaussian example, this would again lead to the conclusion that the variance of the estimate distribution is $\frac{1}{J+J_s}$, while in reality this is the variance of the posterior. Equating the variance of the estimate distribution and the posterior is a frequent mistake. To correctly describe the relationship of prior, noise distribution and distributions of PM estimates, there is no alternative to going through the three steps described in Chapter 3 and this chapter.

The prior probability is equal to the overall probability of responding. This misconception can take different forms, such as:

- "The prior probability of responding rightward is 0.5."
- "I have a simple way to derive the prior distribution used by the observer directly from the data. I can simply tally up the observer's estimates across all trials in the experiment. After all, the higher the prior probability of a stimulus value, the more often the observer will report that stimulus value."

Over the course of an experiment, we could keep track of the distribution of the observer's response. It is tempting but incorrect to regard the overall distribution of the observer's response as the observer's prior distribution. In other words, a phrase like "$p(s)$ is the prior probability of responding s" is incorrect. The correct phrase is "$p(s)$ is the prior probability (observer's belief) that the state of the world is s".

To see the distinction between prior distribution and overall response distribution, consider our Gaussian example. Intuitively, if the observer uses the correct prior distribution, then all their responses will be biased towards the mean of the prior and the response distribution will thus be narrower than the prior distribution.

We can demonstrate the difference between the two distributions more formally. The PME estimate on a single trial when the measurement is x is $\frac{xJ+\mu J_s}{J+J_s}$. This is a random variable because x is a random variable. To calculate the distribution of the PME estimate conditioned on the true stimulus s, we used the fact that x given s was Gaussian with mean s and variance σ^2. Similarly, we now consider the distribution of the PME estimate across all s. The distribution of x across all s is Gaussian with mean μ and variance $\sigma^2+\sigma_s^2$. Then, the PME estimate will have mean μ and variance $\frac{\sigma_s^4}{\sigma^2+\sigma_s^2}$ (see Problem 4.9). This expression, which is plotted as a function of σ in **Fig. 4.5**, shows that the overall response distribution is not identical to the prior distribution (whose variance would, of course, be independent of σ).

4.8 Summary and remarks

We described how Bayesian modeling consists of three steps: defining the generative model, deriving an expression for the observer's posterior mean estimate, and deriving the distribution of the posterior mean estimate over many trials. If the generative model is known to and used by the observer, probability calculus specifies the inference process exactly. Although many of the functions and distributions encountered look similar (in this chapter, they are all Gaussian), they must be distinguished carefully. All relevant quantities are shown in **Fig. 4.4**.

The Bayesian model discussed in this chapter, although simple, is in many ways representative of Bayesian modeling in general. The essence of Bayesian observers is that a Bayesian observer considers all possible values of a state-of-the world variable, and computes their respective probabilities. In other words, the Bayesian observer does not commit to a limited set of hypotheses unless so directed by the evidence.

The power of normative modeling. One of the great powers of Bayesian modeling is that it allows one to build a complete model of a perceptual task before having seen any experimental data: the Bayesian model specifies how an observer *should* do the task in order to be optimal. Bayesian modeling is therefore an example of *normative* modeling: the Bayesian model sets the norm – the highest performance that can be achieved by an observer. This stands in contrast to common practice in much of psychology, in which modeling, if done at all, is often done *after* having observed certain patterns in the data. In Bayesian modeling, one can write down the model and perform model simulation without having even started an experiment.

The Bayesian model we discussed is a backbone model. There are various extensions that can make the model more realistic. These include:

- *Decision noise.* We have described the mapping from the measurement to the stimulus estimate to be deterministic, namely as defined by the posterior mean. This is optimal, but not necessarily realistic. Noise could be added to this mapping. One form of noise is Gaussian noise with constant variance, reflecting generic stochasticity. Another form is that the posterior is read out in a noisy way. Specifically, an estimate could be *sampled* from the posterior distribution instead of always being the posterior mean.
- *Response noise.* We assumed that the observer's response is equal to the stimulus estimate. In practice, every response on a continuum will be subject to some response (e.g., motor) noise, reflecting for instance the accuracy with which the observer is able position the computer cursor. Moreover, the observer's memory of the estimate might decay slightly between the moment the estimate is made and the moment that the response is submitted. Thus, the observer's response distribution is not necessarily the same as the estimate distribution. A complete model of the task would include the response noise. For example, the observer's response could be drawn from a Gaussian distribution with mean equal to the stimulus estimate but with some variance σ^2_{motor}. We will treat this case in a Problem. However, response noise is not central to the Bayesian formalism.

4.9 Suggested readings

- Felix A Wichmann and N Jeremy Hill. “The psychometric function: I. Fitting, sampling, and goodness of fit”. In: *Perception & psychophysics* 63.8 (2001), pages 1293–1313
- Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003
- Shane T Mueller and Christoph T Weidemann. “Decision noise: An explanation for observed violations of signal detection theory”. In: *Psychonomic bulletin & review* 15.3 (2008), pages 465–494

4.10 Problems

Problem 4.1 Match the following functions that play a role in Bayesian modeling with the descriptions:

FUNCTIONS

1. Distribution of the posterior mean estimate
2. Prior distribution
3. Likelihood function
4. Posterior distribution
5. Measurement distribution

DESCRIPTIONS

a) Result of inference on an individual trial
b) Describes how potentially noisy observations are generated
c) Can be directly compared to human responses in a psychophysical experiment
d) Often modeled as a Gaussian shape centered at the measurement
e) May reflect statistics in the natural world

Problem 4.2 In the Gaussian model discussed in this chapter:

a) Prove mathematically that the variance of the posterior mean estimate is smaller than or equal to the variance of the MLE.
b) How do the two compare when the variance of the prior is much larger than the variance of the measurement? Explain why the answer makes intuitive sense.

Problem 4.3 True or false? If false, explain.

a) The likelihood function is always equal to the measurement distribution.
b) The value of the stimulus s that maximizes posterior probability is the value of the measurement x.
c) We can always estimate the distribution of subject’s responses by multiplying the measurement distribution with the prior.
d) There are some cases where the distribution of subject’s responses is Gaussian.
e) If, over the course of an experiment, a Bayesian observer reports one value of the stimulus estimate more often than another value, it means that the prior probability of the former is higher.

Problem 4.4 Show mathematically that the maximum of the standard deviation of the response distribution as a function of noise level (**Fig. 4.1**) is at $\sigma = \sigma_s$. (Recall from calculus that for smooth functions the maximum of a function is where the derivative is zero.)

Problem 4.5 An observer infers a stimulus s from a measurement x_{obs}. As in the chapter, the measurement distribution $p(x|s)$ is Gaussian with mean s and variance σ^2. Unlike in the chapter, we use the prior

$$p(s) = e^{-\lambda s}, \tag{4.19}$$

where λ is a positive constant. This is an *improper prior* (see Box 3.10) but that does not stop us.

a) Derive an equation for the posterior mean estimate.

b) Derive an equation for the probability distribution of the posterior mean estimate for given s.

Problem 4.6 We define *relative bias* as the ratio between bias and the difference between the mean of the stimulus distribution and the true stimulus.

a) For the posterior mean estimate, derive an expression for relative bias as a function of the ratio $R \equiv \frac{\sigma}{\sigma_s}$. (Hint: the expression will only contain R, no other variables.)

b) Plot relative bias as a function of R.

c) Does this plot show what you would expect from the Bayesian observer? Explain intuitively.

Problem 4.7 Let us look at the prior mean estimate. Consider the estimate $\hat{s}_{\text{prior mean}} = \mu$, which completely ignores the measurement and simply returns the mean of the prior. One of the following questions is a trick question.

a) Derive an expression for (overall) mean squared error of this estimate.

b) How large must σ be for this estimate to have a lower mean squared error than the maximum-likelihood estimate?

c) How large must σ be for this estimate to have a lower mean squared error than the posterior mean estimate?

Problem 4.8 Let us calculate the overall mean squared error. Eq. (4.14) represents the mean squared error of the posterior mean estimate for given s. Using that equation, we now calculate the average of this quantity over all s, i.e. not only averaged over $\hat{s}$, but over both $\hat{s}$ and s (see Eq. (4.13)). This *overall mean squared error* is the mean squared error that one would expect in a sufficiently long experiment.

a) Show mathematically that the overall mean squared error of the posterior mean estimate is equal to $\frac{1}{J+J_s}$. Hint: $\mathbb{E}_s[(s-\mu)^2] = \sigma_s^2$.

b) Verify that the expression in (a) returns the purple numbers in **Fig. 4.2B**.

c) Consider any estimate that is a linear transformation of the measurement: $\hat{s}_{\text{linear}} = ax_{\text{obs}} + b$, where a and b are constants. Calculate its overall mean squared error. Hint: $\mathbb{E}_s[s^2] = \sigma_s^2 + mu^2$.

d) Optional: Prove mathematically that among these estimates, the posterior mean estimate is the one with the lowest overall mean squared error. Hint: calculate partial derivatives with respect to a and b.

Problem 4.9 Let us look at the overall (stimulus-averaged) variance of the measurement and posterior mean estimates.

a) Show that across all trials in an experiment, the MLE (on expected value) has mean μ and variance $\sigma^2 + \sigma_s^2$. (Hint: Use Box 4.2 on linear combinations of random variables.) We will call this the "overall variance" of the MLE.

b) Making use of the variance in (a), show that across all trials in an experiment, the posterior mean estimate has mean μ and variance $\frac{\sigma_s^4}{\sigma_s^2+\sigma^2}$. We will call this the "overall variance" of the posterior mean estimate.

c) Show that the overall variance of the posterior mean estimate is always smaller than the overall variance of the MLE.

d) In the absence of sensory noise ($\sigma = 0$), part (c) predicts that the overall variance of the posterior mean estimate is σ_s^2. Explain why this makes sense.

e) Perform a similar sanity check corresponding to $\sigma \to \infty$. (Hint: if sensory noise is extremely large, what can be said about the observer's estimate?)

Problem 4.10 A student claims "To obtain the probability density of an observer's posterior mean estimate for a given stimulus, I multiply the measurement distribution $p(x|s=\mu)$ where μ is the mean of the prior, with the prior probability density $p(s)$."In other words,

$$\text{(wrong)} \quad p(\hat{s}_{\text{PM}}|s) = p_{x|s}(\hat{s}_{\text{PM}}|\mu)p_s(\hat{s}_{\text{PM}}). \tag{4.20}$$

a) Although this would produce a distribution that is centered in between the prior mean and the true stimulus, this claim is conceptually wrong. Why?

b) Show mathematically that the variance of the resulting distribution is incorrect.

c) As a specific example, determine what this student would predict for the variance if measurement noise were extremely large (the limit $\sigma \to \infty$). Explain both answers intuitively.

Problem 4.11 A student claims "To obtain the response distribution $p(\hat{s}_{\text{PM}}|s)$ in the model, I can simply use the posterior $p(s|x_{\text{obs}})$ obtained when the measurement x_{obs} happens to be equal to s." In other words,

$$\text{(wrong)} \quad p(\hat{s}_{\text{PM}}|s) = p_{s|x}(\hat{s}_{\text{PM}}|s). \tag{4.21}$$

a) Although this would produce a distribution that is centered in between the prior mean and the true stimulus, this claim is conceptually wrong. Why?

b) Show mathematically that the variance of the resulting distribution is incorrect.

c) As a specific example, determine what this student would predict for the variance if measurement noise were extremely large (the limit $\sigma \to \infty$), and what the correct calculation would predict. Explain both answers intuitively.

Problem 4.12 A student claims "To obtain the response distribution $p(\hat{s}_{\text{PM}}|s)$ in the model, I can average the posteriors $p(s|x_{\text{obs}})$ over all x_{obs} for a given s." In other words,

$$\text{(wrong)} \quad p(\hat{s}_{\text{PM}}|s) = \int p_{s|x}(\hat{s}_{\text{PM}}|x_{\text{obs}})p(x_{\text{obs}}|s). \tag{4.22}$$

a) Although this would produce a distribution that is centered in between the prior mean and the true stimulus, this claim is conceptually wrong. Why?

b) Show mathematically that the variance of the resulting distribution is incorrect.

c) As a specific example, determine what this student would predict in the infinite-noise limit ($\sigma \to \infty$), and what the correct calculation would predict. Explain both answers intuitively.

Problem 4.13 We mentioned that an observer's response might be corrupted by response or motor noise. Assume motor noise that follows a Gaussian distribution with mean equal to the posterior mean estimate and with standard deviation σ_m.

a) What is the distribution of the observer's response when the true stimulus is s?

b) Think of ways to experimentally distinguish motor noise from noise in the measurement.

Problem 4.14 Repeat parts (a) to (f) of 3.5, but instead of using a single value of the measurement x_{obs}, start with a fixed value of $s = 10$. From this value of s, draw ten values of x_{obs} from the measurement distribution. You should observe that, from trial to trial, the likelihood function and posterior probability density function both "jump around".

Problem 4.15 Assume a Gaussian stimulus distribution $p(s)$ with mean 20 and standard deviation 4; this also serves as the prior distribution. The measurement distribution $p(x|s)$ is Gaussian with standard deviation $\sigma = 5$.

a. Draw 1000 values of s from the stimulus distribution. For each value of s, draw a single x from the measurement distribution $p(x|s)$.

b. Scatter in a scatterplot the measurement x against the true stimulus s. In a separate scatterplot, scatter the posterior mean estimate against the true stimulus. Set your axes suitably. Draw the diagonal as a dashed black line. Draw a horizontal dashed black line at the prior mean.

c. Repeat (a) and (b) using $\sigma = 1$ instead of $\sigma = 5$. The two scatterplots should look very similar. Explain.

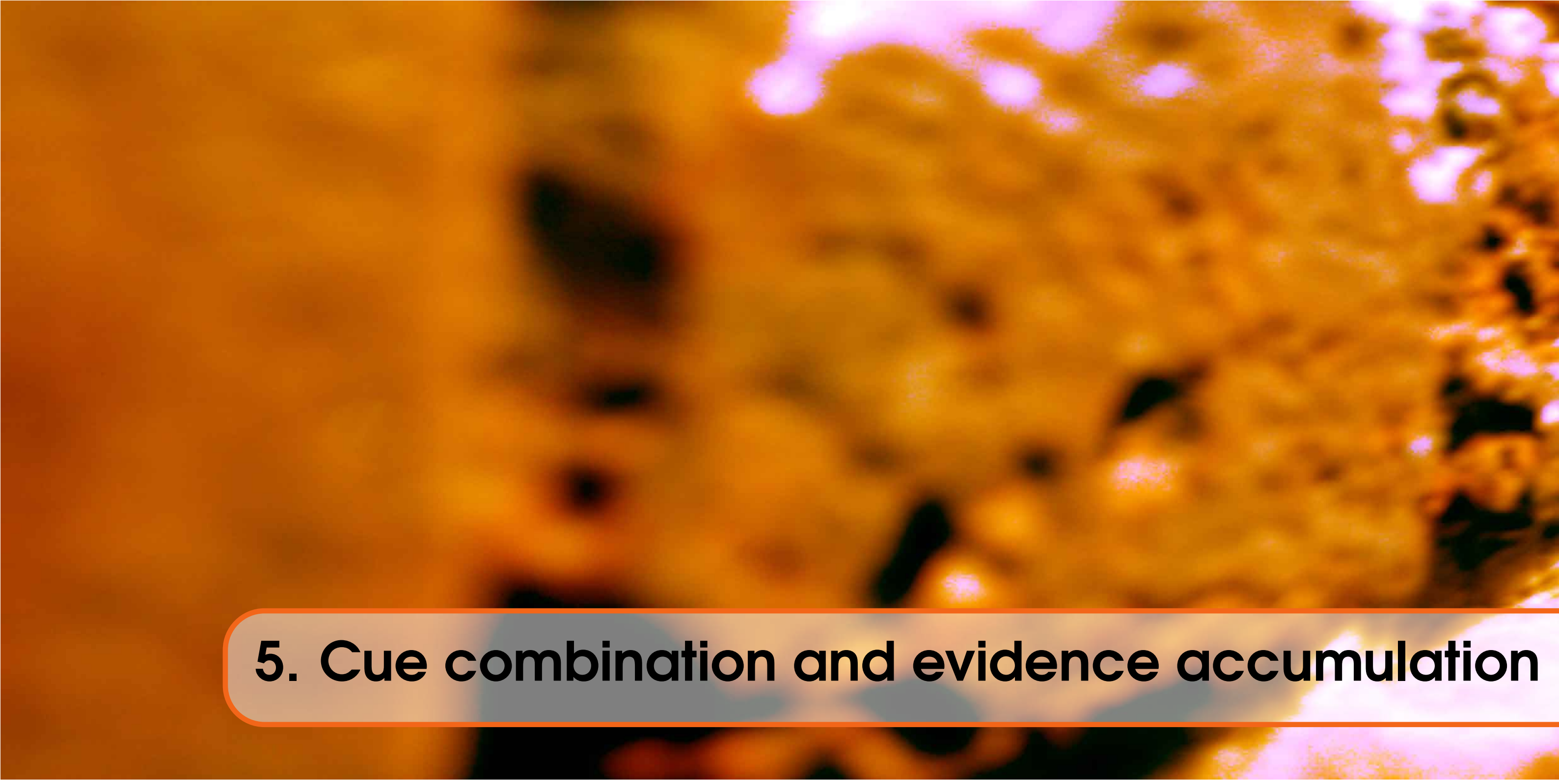
d. Repeat (a) and (b) using $\sigma = 20$. The first scatterplot should look very noisy, whereas the second one should look flat. Explain.

Problem 4.16 In this problem, we will compare the properties of the MAP and the MLE estimators. Refer to the parameter settings in **Fig. 4.1**.

a) Reproduce **Fig. 4.1A**.

b) Reproduce the MLE and PME curves in **Fig. 4.1B** by plotting the corresponding mathematical expressions.

c) For each of the three noise levels, simulate 10,000 values of the stimulus s. For each s, compute the squared bias and variance of the posterior mean estimate using the expressions from (b). Add the two quantities to obtain the mean squared error at that s. Average across all values of s that you drew to obtain the *overall mean squared error* of the posterior mean estimate. The resulting values should be close to the PME-related numbers (12.8, 32, 51.2) in **Fig. 4.1B**.



5. Cue combination and evidence accumulation

How can we integrate multiple sensory cues into a single percept?

In Chapters 3 and 4, we studied the simplest possible generative model, with a single stimulus and a single measurement. Here, we study an extension in which there are two measurements. These are based on sensory inputs that are also called cues. The measurements could correspond to an auditory and a visual measurement of a stimulus, such as the location at which a ball drops on the ground. The observer estimates the stimulus value based on both cues. Mathematically, the model is a straightforward extension of the one in Chapter 3 and 4. Yet, there are three important reasons to study this generative model. First, cue combination occurs very often in daily life. Second, it is a historically early and still prominent domain of application of Bayesian modeling. Third, this generative model is our first example in which the Bayesian observer computes the likelihood function over the world state of interest from two simpler likelihoods. Stronger yet, one important take-away message of this chapter is that Bayesian inference does not need to involve priors to be interesting. The key aspect of Bayesian decision-making is not the prior, it is the reasoning with probability distributions rather than point estimates.

Plan of the chapter

We will begin by discussing the intuitions behind cue combination. We will develop Steps 1 through 3 of Bayesian inference for cue combination. We will show how integration of sensory evidence over time is mathematically equivalent to cue combination. Finally, we will discuss the empirical literature that addresses how well people actually combine cues.

5.1 What is cue combination?

When trying to understand someone's speech, it helps not just to listen carefully but also to simultaneously view the speaker's facial movements and nonverbal gestures. This is an example of cue combination. Combining cues is especially important when an individual cue is noisy, for example when you are trying to understand speech in the presence of background noise. The ability to combine cues enhances performance in perceptual tasks.

In numerous daily perceptual situations, we receive and combine cues from different sensory modalities, yet we do this so effortlessly that we may be unaware it is happening. When tasting food, we may think we are engaging in a purely gustatory activity, but in fact we perceive the flavor of food by combining gustatory, olfactory, thermal, and mechanical (texture) cues. When estimating our acceleration while traveling in a moving vehicle, we may think we are relying only on vision, but we are relying as well on proprioceptive cues conveyed by sensors (muscle spindles and Golgi tendon organs) that signal muscle length and tension, and on vestibular cues conveyed by sensors in our inner ears (semicircular canals and otolith organs) that signal rotational and linear acceleration of the head.

In fact, we combine cues not only across sensory modalities but also within a single modality. Each modality provides an array of distinct cues. In vision, for example, the relative activation of photoreceptors tuned to different wavelengths tells us an object's color; the pattern of reflected light indicates the object's surface texture; and comparison of the images in the two retinae informs us about the object's depth. We effortlessly combine these and other visual cues to infer object identity. Similarly, distinct receptors in the skin provide us with mechanical, thermal and nociceptive information, and even within each of these somatosensory divisions we obtain multiple cues. For example, different mechanoreceptor subtypes provide information about static pressure (Merkel receptors), skin stretch (Ruffini receptors), low frequency vibration (Meissner receptors) and high-frequency vibration (Pacinian receptors). When we run a fingertip across an unknown surface, we obtain information about surface texture, friction, hardness, and other qualities by combining cues from these receptors. This allows us to achieve fine perceptual inferences, distinguishing for instance the feel of silk from that of velvet or wool.

Why should the brain combine cues? To answer this question, let's explore the consequences of an obvious alternate strategy: the brain could simply use the single most informative cue that it has at hand, and ignore the rest. Upon reflection, it is clear that this winner-take-all strategy is suboptimal for two reasons. First, our sensorineural responses are noisy, with the consequence that any parameter estimate based even on the most reliable cue is subject to some uncertainty; a strategy that does not include the other, albeit less reliable cues, discards information that can be used to sharpen the precision of the estimate. Second, even when sensorineural noise does not impose serious limitations, an individual cue is often ambiguous; a strategy that does not include all available cues will often fail to overcome ambiguities. To illustrate these points, we consider two examples.

First, let's suppose we wish to infer the location at which a dropped ball hits the ground. This event provides both visual and auditory cues. Now suppose that we base our inference about location entirely on the visual cue because the ball is dropped in a well-lit environment under direct view, a condition in which vision is more informative than audition. Because our photoreceptors and neural responses are noisy, even the estimate based on this most reliable cue will have some uncertainty, as reflected in the width of the posterior distribution over location. We will show below that inclusion of a less reliable cue (e.g., the auditory cue in this example) nevertheless contributes useful information. Thus, by combining cues, we obtain a more precise estimate than the one obtainable from the best cue alone.

Second, let's suppose we wish to infer the identity of a spherical object, placed on a tabletop, in the dark. As we rest our hand on the object, our proprioceptors inform us about its size, but we cannot unambiguously identify an object from its size alone. In this case, our uncertainty about the object's identity (as opposed to its exact size) does not result primarily from sensorineural noise. Rather, our uncertainty derives from the inherent ambiguity of size as a cue to object identity: even if we knew its size exactly, we would not know the object's identity, because different individual objects (e.g., an apple and an orange) can have the same size. Although size might provide the single most informative cue in this scenario (e.g., greatly narrowing down the set of possible objects to those as big as apples and oranges), with further exploration – lifting and manipulating the object

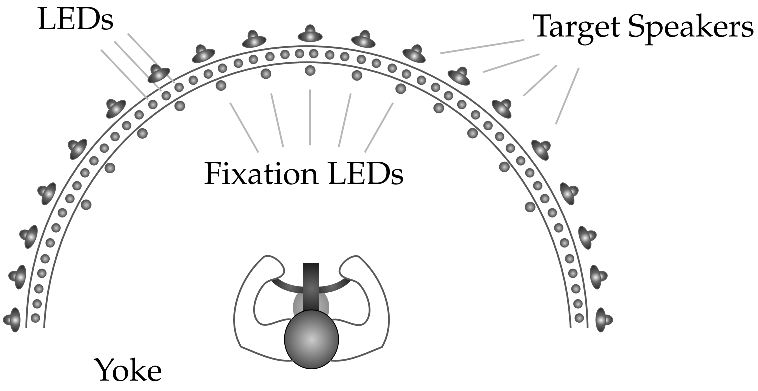


Figure 5.1: Experimental set-up for testing auditory or multisensory localization. A speaker produces a brief tone. At the same time, an LED might produce a brief flash. The subject points a laser pointer at the perceived location of the tone. Figure reproduced from Wallace et al. (2004).

- we could glean from our muscle and mechanoreceptors an understanding of the object's weight and surface texture, further narrowing down the possible set of objects. In short, we overcome ambiguity by combining cues.

How should we expect the nervous system to combine two pieces of information? Obviously, although the winner-take-all strategy is too extreme, we do expect people to rely more on those cues that are most informative to the task at hand. If, in a particular scenario, vision is more informative than audition, for example when we want to locate a person who is talking in an environment with loud ambient noise, then we should mostly rely on vision. When audition is more informative than vision, we expect the reverse – and indeed at night we often rely primarily on our auditory system.

Over the last decades, many scientists have studied cue combination in the laboratory. In a typical experiment, the subject is surrounded by an array of loudspeakers and light-emitting diodes (LEDs, see **Fig. 5.1**). An audiovisual stimulus is produced by the simultaneous occurrence of a brief auditory beep and a light flash. The subject is instructed to indicate the perceived location of the beep. With this apparatus, scientists can probe how visual and auditory cues are combined by the nervous system.

The results of these experiments reveal that when the beep and flash occur at the same – or nearly the same - location, subjects use the visual stimulus to help estimate the location of the auditory stimulus, even when they are instructed to ignore the visual stimulus. The subjects thus naturally and intuitively combine the cues, apparently operating under the assumption that the beep and flash originate from a single source. Indeed, under conditions in which the auditory and visual cues originate from slightly offset locations, subjects are easily led astray as their "auditory localization" estimates are biased by the presence of the visual cue. Importantly, the more precise the visual cue relative to the auditory one, the more strongly subjects rely on the visual cue in formulating their localization estimate.

5.2 Formulation of the Bayesian model

Here we formulate a Bayesian model for optimal cue combination. In developing our formulation, we use the example of auditory-visual location estimation, but the same approach can be applied to other cue combination scenarios. Our approach follows the same three steps of Bayesian modeling outlined in Chapters 3 and 4.

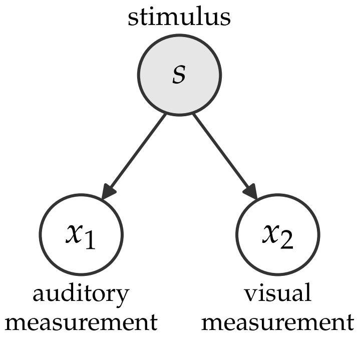


Figure 5.2: Generative model of cue combination. The stimulus affects both auditory and visual measurement.

5.2.1 Step 1: Generative model

Our model consists of three nodes: the stimulus s and two measurements, x_1 and x_2 (see **Fig. 5.2**). Associated with the stimulus is a stimulus distribution $p(s)$. Unlike in Chapter 3, we assume here that the prior distribution is flat. We do this not only because it is the most common assumption in cue combination studies, but also because it illustrates nicely that a non-uniform prior is not necessary for a Bayesian model to be interesting and important.

Box 5.1 **Myth:** Bayesian inference is all about priors.
Truth: In many interesting Bayesian models, the prior is flat. We already saw this in Section 2.5, but cue combination is the best-known example. What makes the Bayesian model for cue combination interesting is that the posterior mean estimate is not just determined by the measurements but also by their inverse variances. ■

Any measurement is noise and we need to make assumptions about their distribution. Associated with the measurements are thus noise distributions $p(x_1|s)$ and $p(x_2|s)$. The separate arrows pointing to x_1 and x_2 reflect a key assumption, namely that these measurements are conditionally independent. Conditional independence of two random variables (see Box) means that they are independent of each other when conditioned on another variable, in this case the actual stimulus s. Specifically, this means that while vision is noisy and audition is noisy, the noise corrupting the two streams is uncorrelated: there is zero noise covariation between the two modalities. This assumption is easier to justify when the two measurements come from two different sensory modalities (e.g. auditory and visual) then from the same one (e.g. two visual measurements, see e.g. Section 5.9). Our calculations are thus based on an assumption of conditional independence.

It is important to distinguish conditional independence from independence. Our visual and auditory percepts are obviously not going to be independent from one another. When the stimulus is to the left, both modalities will tend to indicate a position to the left, and when the stimulus is to the right, both modalities will tend to indicate a position to the right. However, we assume that upon repeated presentations of the same stimulus, the trial-by-trial variability in the auditory and visual measurements will be uncorrelated. Thus, the visual and auditory measurements are independent of one another when conditioned on s.

Mathematically, the conditional independence of x_1 and x_2 given s is expressed as

$$p(x_1,x_2|s) = p(x_1|s)p(x_2|s) \tag{5.1}$$

Independence would have been $p(x_1,x_2) = p(x_1)p(x_2)$, but this is not true here.

Exercise 5.1 What would the generative model look like if the two measurements were not assumed conditionally independent? ■

Box 5.2 — Conditional independence. Conditional independence occurs when two random variables are independent only given the value of a third one. For example, having Alzheimer's and needing reading glasses are two events that are not independent, because both tend to occur in older people. However, among only 80-year-olds (i.e. given the age group), the two are probably more or less independent. Another, famous example is that homicide rates in a city and ice cream sales are not independent random variables: they both tend to be more probable on hot days. However, given the temperature in the city, the two are conditionally independent.

The intuition is that you condition on the value of the cause of dependence of the two variables. If X, Y, and Z denote three random variables, then X and Y are independent given Z if

$$p(x,y|z) = p(x|z)p(y|z)$$

for any values x, y, and z. Be careful not to confuse conditional independence with independence!

Another way to interpret Eq. 5.1 is by starting from a rule that is generally true, the product rule for probabilities:

$$p(x,y|z) = p(x|y,z)p(y|z)$$

Then to get to Eq. 5.1 , we have to make the assumption that $p(x|y,z) = p(x|z)$. In other words, knowledge of z fully specifies our understanding of the probability of x. When we know z, also knowing y does not contribute anything to our assessment of the probability of x. ■

Exercise 5.2 Think of another real-world example of conditional independence. ■

For the distribution of each individual measurement, we choose a Gaussian distribution:

$$p(x_1|s) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{\frac{-(x_1-s)^2}{2\sigma_1^2}} \tag{5.2}$$

$$p(x_2|s) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{\frac{-(x_2-s)^2}{2\sigma_2^2}} \tag{5.3}$$

As we did in the previous two chapters, we will use *precision notation* for convenience. We define two precision variables:

$$J_1 \equiv \frac{1}{\sigma_1^2} \tag{5.4}$$

$$J_2 \equiv \frac{1}{\sigma_2^2} \tag{5.5}$$

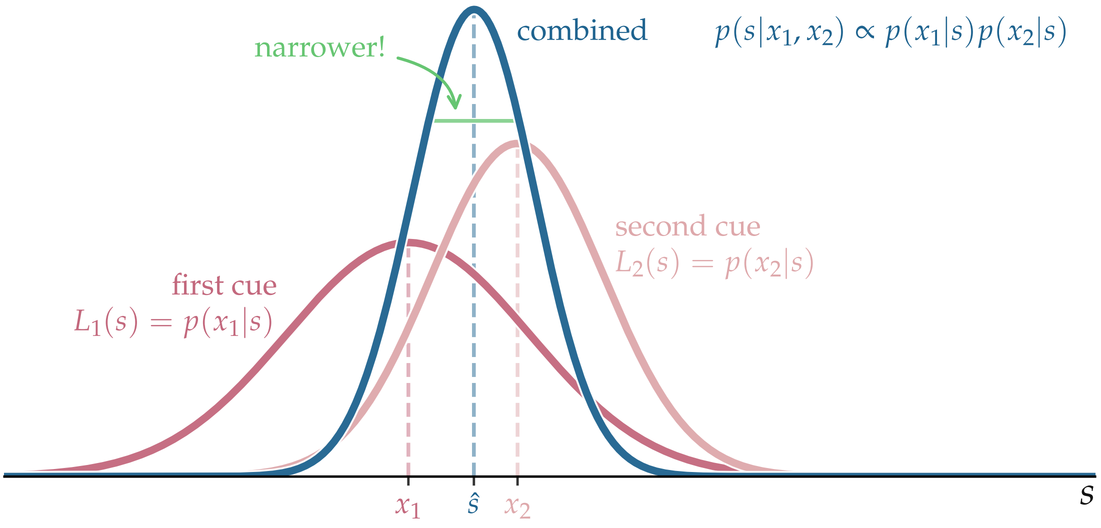


Figure 5.3: The computation of the posterior distribution in cue combination. Note that both red curves are likelihood functions. The prior is flat and not shown.

5.2.2 Step 2: Inference

The observer infers the stimulus s from the measurements $x_{\text{obs},1}$ and $x_{\text{obs},2}$. The likelihoods over the stimulus are the same expressions as the noise distributions but regarded as a function of s:

$$\mathscr{L}_\infty(s;x_{\text{obs},1}) \equiv p(x_{\text{obs},1}|s) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{\frac{-(x_{\text{obs},1}-s)^2}{2\sigma_1^2}} \tag{5.6}$$

$$L_2(s) \equiv p(x_{\text{obs},2}|s) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{\frac{-(x_{\text{obs},2}-s)^2}{2\sigma_2^2}} \tag{5.7}$$

We call each of these an *elementary likelihood function*, defined as a likelihood function over a stimulus feature associated with an individual measurement. The posterior distribution over the stimulus is computed from Bayes' rule:

$$p(s|x_{\text{obs},1},x_{\text{obs},2}) \propto p(s)p(x_{\text{obs},1},x_{\text{obs},2}|s). \tag{5.8}$$

We have left out the factor $\frac{1}{p(x_{\text{obs},1},x_{\text{obs},2})}$ for the reason explained in 3.6: it only acts as a normalization, so if we normalize the distribution in the end, we automatically take this factor into account. Since the stimulus distribution is flat, the prior is flat as well, and the posterior is determined by the likelihood $p(x_1,x_2|s)$ only. To make further progress, we use Eq. 5.1, the assumption of conditional independence of the measurements. Then, the posterior becomes:

$$\begin{aligned} p(s|x_{\text{obs},1},x_{\text{obs},2}) &\propto p(s)p(x_{\text{obs},1}|s)p(x_{\text{obs},2}|s) && (5.9) \\ &\propto p(x_{\text{obs},1}|s)p(x_{\text{obs},2}|s) && (5.10) \end{aligned}$$

This step – making use of the structure of the generative model to express the likelihood in terms of elementary likelihoods – is the only conceptually new element in this chapter compared to Chapter 3. Expressing the likelihood over the state-of-the-world variable in terms of elementary likelihoods is at the core of Bayesian inference models for many tasks. We can substitute the two

Gaussian distributions into this equation and simplify in the same way we did in Chapter 3. The result is that the posterior is another Gaussian distribution (**Fig. 5.3**):

$$p(s|x_{\rm obs,1},x_{\rm obs,2}) = \frac{1}{\sqrt{2\pi\sigma^2_{\rm post}}}e^{\frac{-(s-\mu_{\rm post})^2}{2\sigma^2_{\rm post}}}$$

where the mean is:

$$\mu_{\rm post} = \frac{J_1 x_{\rm obs,1} + J_2 x_{\rm obs,2}}{J_1+J_2}.$$

and the variance is:

$$\sigma^2_{\rm post} = \frac{1}{J_1+J_2}.$$

Exercise 5.3 Show this. Hint: constants added and positive constants multiplied do not change where a function has its maximum. ■

We can also write the posterior mean as

$$\mu_{\rm post} = w_1 x_{\rm obs,1} + w_2 x_{\rm obs,2}, \tag{5.11}$$

where the weights are proportional to the precisions:

$$w_1 = \frac{J_1}{J_1+J_2} \tag{5.12}$$

$$w_2 = \frac{J_2}{J_1+J_2} \tag{5.13}$$

The posterior mean estimate is in this case equal to the maximum-likelihood estimate (as well as to the maximum-a-posteriori estimate). As a reminder, the posterior mean estimate is the estimate that minimizes the expected squared error (see Section 3.3.6).

You might have noticed that the observer's computation of the posterior is exactly analogous to the computation in Chapter 3, where we combined a single cue with a prior. The second cue has now taken the role of the mean of the prior distribution. In the present example, the prior is flat, and therefore the posterior is equal to the normalized likelihood function. The posterior mean estimate is:

$$\hat{s}_{\rm PM} = \mu_{\rm post}.$$

These weights sum to 1: $w_1 + w_2 = 1$, indicating that the posterior mean estimate is a weighted average of the two measurements. Averaging with weights proportional to the inverse of the variance is by far the most frequently used model in cue combination. An analogy is that of a police investigator trying to reconstruct a crime based on the testimonies of two witnesses. One witness was intoxicated at the time of the crime, the other was not. The testimony of the sober witness would be the result of a low-noise process; of the inebriated witness, the result of a high-noise process. A good investigator would take both testimonies into account, but would more heavily weight the testimony of the less noisy witness.

The variance of the posterior is a measure of the observer's uncertainty. Under the assumptions we made, it is smaller than the variances of the elementary likelihood functions.

Exercise 5.4 Prove the fact that the variance of the posterior is never larger than that of either components. ■

Intuitively, this says that combining cues reduces uncertainty; the observer is more confident in the combined estimate than in the estimate that would be obtained from either cue alone.

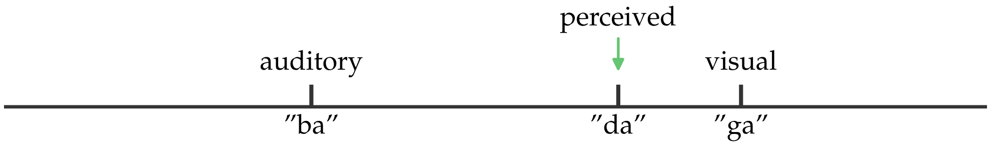


Figure 5.4: Schematic depiction of the McGurk effect.

5.2.3 Step 3: Estimate distribution

As the third step in our Bayesian model, we are interested in the distribution of the posterior mean estimate across many trials. The posterior mean estimate is given as a function of the measurements x_1 and x_2 in Eq. (5.5), but the measurements are themselves random variables – their values vary from trial to trial. As a consequence, the posterior mean estimate varies from trial to trial as well. Since in a behavioral experiment, we never know the measurements on a single trial (they are in the observer's head), we have to compare behavior with the distribution of the posterior mean estimate over many trials. To find the mean and variance of the posterior mean estimate, we apply the rules for linear combinations of normally distributed variables. The means of x_1 and x_2 are both s. Therefore, the model predicts that the mean posterior mean estimate will be $w_1 s + w_2 s = w_1 s + (1 - w_1)s = s$. In other words, for cue combination with a flat prior, the posterior mean estimator is unbiased. (Recall that bias is defined as the difference between the mean estimate and the true stimulus.) The variance of the posterior mean estimate will be:

$$\sigma^2_{\text{post}} = \frac{1}{J_1 + J_2}.$$

Thus, in this cue combination problem, the variance of the posterior mean estimate distribution happens to be identical to the variance of the posterior. This stands in contrast to Chapter 3, where we found that the variance of the estimate distribution was different from the variance of the posterior. This difference arises because the prior was chosen flat in the current chapter. When the prior is Gaussian, then the variance of the posterior mean estimate will differ from the variance of the posterior also for cue combination (see Section 5.4).

5.3 Artificial cue conflict

We saw that the mean posterior mean estimate is equal to the true stimulus. This is not very interesting, since it does not distinguish between the optimal cue combination model and a model in which the observer only uses one of the cues. (The variance of the posterior mean estimate does distinguish, but it is always better to have two measures than one.) In cue combination experiments, therefore, a common trick is to introduce a small conflict between the true stimuli in the two modalities. In other words, unbeknown to the observer, there is not a single s, but rather two slightly displaced stimuli, s_1 and s_2. Everything else remains the same.

A famous example of an artificial cue conflict is the McGurk effect (McGurk and MacDonald, 1976). When we hear the sound of someone saying baba while we play the video of the same person saying gaga, we perceive the person saying dada (**Fig. 5.4**). Plenty of live demos of the McGurk effect can be found online.

The McGurk effect can be understood as an instance of cue combination in which the observer infers a single, common s from the measurements $x_{\text{obs},1}$ and $x_{\text{obs},2}$. Of course, this requires that the observer still believe that there is a single underlying stimulus, in spite of the discrepancy introduced by the experimenter. The experimenter sometimes explicitly instructs the observer to imagine that the two cues are generated by a single stimulus, for example an auditory and a visual measurement generated by a ball hitting the screen.

Distribution	**Argument**	**Mean**	**Variance**	**Precision**
Measurement distributions	Measurement x_i	Stimulus s_i	σ_i^2	J_i
Likelihood functions	Hypothesized stimulus s	Measurements x_i	σ_i^2	J_i
Posterior distribution	Hypothesized stimulus s	$\frac{J_1 x_{\text{obs},1}+J_2 x_{\text{obs},2}}{J_1+J_2}$	$\frac{1}{J_1+J_2}$	J_1+J_2
Response distribution	Stimulus estimate $\hat{s}_{\text{ML}}$	$\frac{J_1 x_{\text{obs},1}+J_2 x_{\text{obs},2}}{J_1+J_2}$	$\frac{1}{J_1+J_2}$	J_1+J_2

Table 5.1: Overview of the distributions in this chapter. *i* can be 1 or 2.

At first consideration, the McGurk effect appears to reveal a form of *suboptimal* inference, because although the investigator has deliberately used two discrepant stimuli, the observer nonetheless incorrectly infers a single common stimulus. Another way to think about this behavior, however, is to consider that the observer is applying a prior based in natural statistics; in the real-world, when the observer simultaneously sees a ball hit the ground and hears a thud, the visual and auditory stimuli nearly always result from the same event, and therefore originate from the same location. In the laboratory experiment, the investigator has contrived a situation that would rarely occur in the world, and therefore is easily misinterpreted by the observer. Keep in mind that even when there is truly a single stimulus, the auditory and visual measurements will differ from each other on each trial because of noise (unless somehow the noise were completely correlated). Thus, the mere fact that the two cues differ does not imply that they resulted from two stimuli at different locations. The observer's inference may still be optimal, then, under a real-world prior.

Of course, the observer will only believe in a single stimulus if the discrepancies introduced are small. Otherwise, the observer will notice a conflict. For example, if the sound of bouncing ball originates at a sufficiently large distance from the visual image of the ball, the observer will realize that two separate stimuli were presented. Similarly, if a movie is poorly dubbed, the discrepancy in time between the speaker's mouth movement and voice will be too large to go unnoticed. When the observer does not necessary believe that there is a single common cause, the observer's inference process changes. This interesting situation will be discussed in a later chapter (causal inference).

If the observer indeed believes that there is a single common cause, then Step 2 above is unchanged. In Step 3, however, the means of x_1 and x_2 are no longer both s, but s_1 and s_2, respectively. As a consequence, the mean posterior mean estimate will be:

$$\mathbb{E}[\hat{s}_{\text{PM}}] = w_1 s_1 + w_2 s_2. \tag{5.14}$$

This estimator is biased. For example, the bias with respect to stimulus s_1, which we denote here by Bias_1, is:

$$\begin{aligned}
\text{Bias}_1(\hat{s}_{\text{PM}}|s_1,s_2) &= \mathbb{E}[\hat{s}_{\text{PM}}] - s_1 && (5.15)\\
&= w_1 s_1 + (1-w_1)s_2 - s_1 && (5.16)\\
&= (1-w_1)(s_2-s_1) && (5.17)
\end{aligned}$$

This predicted bias can be compared against experimental data and is indeed often found to be a good match (see Section 5.7).

5.3.1 Distinguishing the distributions

As in Chapter 4, it is important to distinguish between the posterior (single trial, Step 2), and the estimate distribution (multiple trials, Step 3); see **Table 5.1**. It just happens to be the case that when the prior is flat, as we have assumed so far, the estimate distribution has the same variance as the posterior, but this is not the case in general.

5.4 Generalizations: prior, multiple cues

In Chapters 3 and 4, we studied the combination of a Gaussian prior (mean μ, standard deviation σ_s, precision J_s) with a single measurement. In this chapter, we studied the combination of two conditionally independent measurements. The two combinations can easily be combined. The posterior becomes

$$p(s|x_{\text{obs},1},x_{\text{obs},2}) \propto p(s)p(x_{\text{obs},1}|s)p(x_{\text{obs},2}|s)$$

and the posterior mean estimate becomes

$$\hat{s}_{\text{PM}} = \mu_{\text{post}} = \frac{J_s\mu + J_1 s_1 + J_2 s_2}{J_s + J_1 + J_2} \tag{5.18}$$

We can further generalize to multiple cues. Combining N cues with the same underlying stimulus is as easy as combining two. Assume the measurements are $x_{\text{obs},1}$, $x_{\text{obs},2}$, ..., $x_{\text{obs}}N$ are conditionally independent given s. Then the posterior is:

$$p(s|x_1\cdots x_N) \propto p(s)p(x_1|s)\cdots p(x_N|s) \tag{5.19}$$

$$= p(s)\prod_{i=1}^{N} p(x_i|s) \tag{5.20}$$

where in the last line we used product notation. Thus, the prior gets multiplied with the likelihoods derived from the individual measurements. If the measurements are normally distributed with mean s and standard deviations σ_1, σ_2, $\cdots$,σ_N, respectively, then the posterior will have a mean of

$$\hat{s}_{\text{PM}} = \mu_{\text{post}} = \frac{J_s\mu + \sum_{i=1}^{N} J_i s_i}{J_s + \sum_{i=1}^{N} J_i} \tag{5.21}$$

and a variance of

$$\sigma^2_{\text{post}} = \frac{1}{J_s + \sum_{i=1}^{N} J_i}. \tag{5.22}$$

5.5 Evidence accumulation

A major way in which organisms improve their knowledge of the world is by observing for a longer time. More time allows more evidence to be accumulated. Mathematically, evidence accumulation is cue combination over time and can be described using the same formalism. In fact, evidence accumulation is often regarded as the prototypical example of Bayesian inference, where a posterior gets updated on each time step based on new information.

We now formalize this. Consider an observer who makes a series of conditionally independent, normally distributed measurements x_1, x_2, $\cdots$, x_T, one at each time point. The measurements all have the same mean, s, and standard deviations σ_1, σ_2, $\cdots$, σ_T, respectively. Under these assumptions, Eq. (5.20) applies with N replaced by T:

$$p(s|x_1\cdots x_T) \propto p(s)p(x_1|s)\cdots p(x_T|s) \tag{5.23}$$

$$= p(s)\prod_{i=1}^{T} p(x_i|s) \tag{5.24}$$

Therefore, with the same substitutions, Eqs. (5.21) and (5.22) also apply. In particular, the variance of the posterior will shrink continuously to 0 as more evidence is accumulated.

In the context of evidence accumulation, it makes sense to think of the computation of the posterior as a recursive process, whereby the posterior after obtaining the measurement at a given time point serves the prior used at the next time point. Mathematically, we can rewrite Eq. (5.24) as an *update equation*

$$p(s|x_{\mathrm{obs},1},\cdots,x_{\mathrm{obs},t+1}) \propto p(s|x_{\mathrm{obs},1},\cdots,x_{\mathrm{obs},t})p(x_{\mathrm{obs},t+1}|s), \tag{5.25}$$

where it is understood that $p(s|x_{\mathrm{obs},1},\cdots,x_{\mathrm{obs}}t)$ for $t=0$ is the prior distribution. In words, the posterior at time t gets multiplied with the likelihood at time $t+1$ to produce, after normalization, the posterior at time $t+1$. This process is called the *(Bayesian) updating of the posterior*, and it is the most fundamental concept in the application of Bayesian modeling to temporal data.

Modeling evidence accumulation in this way comes with several important caveats:

- Termination. We have not specified how and when the evidence accumulation terminates. This is a difficult modeling problem that has a long history.
- Conditional independence. We assumed that measurements across different time points are conditionally independent. This assumption is easily violated: in many cases, processes with a long time scale (such as slow fluctuations of attention) will cause measurements to be correlated over time. Then, Eqs. (5.21) and Eqs. (5.22) will no longer apply. In particular, the variance of the posterior might asymptote to a value larger than 0.
- Stationarity. Another important caveat is that we assumed that the true world state, s, does not change over time. Often, the stimulus itself changes as you accumulate evidence. We will discuss an example where it does change in Chapter 12.
- Forgetting. Information does not stay in the brain forever: this loss of information is forgetting. Forgetting can be modeled as noise being added to measurements at each time step. Then, older memories will have accumulated more noise. Formally, this would be part of Step 1 (Generative model). A Bayesian observer will then in Step 2 take into account that older memories are noisier and give them lower weight. The evolving posterior will then be subject to two counteracting forces: sharpening due to the accumulation of new evidence, and widening due to the forgetting of old information.

5.6 Cue combination under ambiguity

So far, we have considered cue combination under sensory noise. However, as we have seen, uncertainty sometimes arises not from sensory noise but because of inherent ambiguities in the inputs. For instance, in attempting to identify an object in the hand based on its size and weight, we may experience uncertainty not because of sensory noise (given sufficient time to obtain reliable measurements of these variables) but rather because multiple objects can have the same size and weight. Nevertheless, the logic of the inference process is the same in this scenario: the different possible objects are hypotheses (typically discrete ones: apple, orange, etc.), and the measurement of each feature (size, weight, etc.) has a certain probability under each hypothesis. As a function of the hypothesis, this is an individual-feature likelihood function. When the features are independent conditioned on object identity, the likelihood of a particular hypothesis is the product of the individual-feature likelihoods. What complicates matters is that the features do not tend to be conditionally independent. For example, even when restricted to oranges, weight and size tend to correlate strongly. Formally, this is a classification task, not an estimation task. We will discuss classification tasks under ambiguity in Chapter 8.

5.7 Applications

There is a long tradition of probing cue combination by analyzing how humans integrate position information from vision and audition. In many cases, vision is very precise (precision ~min arc), while audition is relatively imprecise (precision ~10 degrees). This has the effect that when reliable

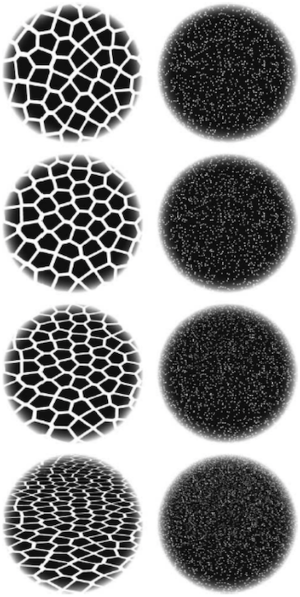


Figure 5.5: Estimating surface slant using texture and stereo cues (Knill and Saunders, 2003). Surface slant was defined using texture (left), stereo (random-dot patterns, right), or both. The four rows shown correspond to 0°, 30°, 50°, and 70° slant. The random-dot patterns have to be presented stereoscopically for the observer to perceive slant; *figure inclusion pending permissions.*

visual information is available, people generally rely primarily on vision, a behavior predicted from Eq. (5.14).

To test the more subtle predictions of the Bayesian model, it was necessary to create situations where vision and audition are similarly precise. In a seminal study, Alais and Burr (2004) accomplished this by blurring visual inputs. The authors used several levels of blur so that visual precision would change unpredictably from trial to trial. They estimated visual precision by presenting visual stimuli alone, making use of the fact that the prior is flat so the variance of the posterior mean estimate for vision alone will equal the variance of the likelihood function (and the noise distribution). Similarly, they presented auditory stimuli alone to estimate the variance of the auditory likelihood function. Using the resulting estimates for σ_A and σ_V, the authors predicted the weights human subjects would place on vision and on audition when combining these cues Eq. (5.13). They found that human behavior was well predicted by Eq. (5.11) with those weights. Moreover, the same model successfully predicted the variance of the posterior mean estimate.

An important technical detail in many cue combination studies is that the experiments usually do not ask for estimates on a continuum (as in **Fig. 5.1**) but instead use a so-called two-alternative forced choice paradigm, in which the subject is presented with stimuli in two intervals and required to make a choice between them. For example, subjects are presented with two sets of auditory-visual stimuli, and asked in which of the two the auditory stimulus was more to the left. This allows the investigator to estimate the variances (precisions) of the cues in a way that is unaffected by the subject's prior. We examine the details of this procedure in a later chapter.

The Bayesian model of cue combination has been tested in other sensory modalities as well. A classic study is that of Ernst and Banks (2002). The authors studied subjects' estimation of the size of an object that could be both seen and felt. Under normal viewing conditions, vision is often more precise than touch; the authors blurred the visual feedback in order to reduce its precision. They found that across different visual feedback precisions, the weight subjects placed on vision was very close to the value predicted by Eq. (5.11).

Many experiments have probed the integration of two cues originating from a single sensory modality. One example is the estimation of slant (orientation of a plane) based on visual texture and visual disparity (**Fig. 5.5**). Texture provides information about slant. Disparity also provides information about slant; the part of the plane that is closer to the viewer will have a smaller binocular disparity.

In a typical study, subjects would be shown a surface with a texture that indicated a given slant. The texture information can be made more or less informative. For example, circles provide a highly informative cue, whereas random white noise provides a very uninformative cue. The disparity cue can also be manipulated, and importantly changed independently of texture. By independently varying the texture and disparity cues, the authors of such studies have generally found that subjects integrate these cues in accordance with the predictions of the Bayesian model. For example, as the texture cue is varied to indicate different slants, it exerts a roughly linear influence on the estimated slant. The slope of that influence, the weight on texture, fits well with the prediction of the Bayesian model.

When we want to estimate the position of our hand in a two-dimensional plane, such as a tabletop, we have to solve a two-dimensional estimation problem. To make this estimate, we can use proprioceptors that signal body posture. We can also use vision. The proprioceptive and visual cues to hand position have different properties. Proprioception is generally noisy, but good at estimating changes in the direction of our smaller joints. Vision is quite good in terms of direction but rather poor at estimating depth. In a seminal study, van Beers and collaborators (1996) probed how the nervous system combines visual and proprioceptive cues in this task. It was found that the cue combination proceeds almost exactly as predicted by the Bayesian model.

Several authors have also studied speech perception from the point of view of the Bayesian model, both in McGurk-like settings with simple syllables, and with monosyllabic words.

There may be some cues that only obtain their meaning through other cues. For example, shadows are not normal cues. If we do not know the direction of the sun then shadows will not actually help us estimate the size of an object. These cues are called pseudo-cues. Recent studies have probed how subjects make use of these non-standard cues.

5.8 Summary and remarks

In this chapter, we have introduced models in which multiple cues have to be combined. We have learned the following:

- Cue combination is a frequent and important perceptual activity that often happens automatically and outside of our conscious control.
- Just as with prior/likelihood combination, all the Bayesian observer needs to do is multiply two probability distributions and normalize.
- Simple Bayesian models can explain how humans combine cues in a wide variety of settings.
- Unlike the winner-take-all strategy, the optimal Bayesian solution, which is followed by humans in many instances, is to weight each cue according to its reliability.
- Cue combination can take place over time, in which case it is sometimes called evidence accumulation, evidence integration, or decision-making. Across subsequent measurements, uncertainty is reduced. The posterior mean estimate is a linear combination of the individual measurements, weighted by their precisions.
- Cue combination can also take place between individuals. In a fascinating study, [24] studied how two individuals combine information about a visual stimulus through verbal communication.

5.9 Suggested readings

- David Alais and David Burr. "The ventriloquist effect results from near-optimal bimodal integration". In: *Current biology* 14.3 (2004), pages 257–262
- Bahador Bahrami et al. "Optimally interacting minds". In: *Science* 329.5995 (2010), pages 1081–1085
- Vikranth Rao Bejjanki et al. "Cue integration in categorical tasks: Insights from audio-visual speech perception". In: *PloS one* 6.5 (2011), e19812
- Anne-Marie Brouwer and David C Knill. "The role of memory in visually guided reaching". In: *Journal of vision* 7.5 (2007), pages 6–6
- Heinrich H Bulthoff. "Bayesian decision theory and psychophysics". In: *Perception as Bayesian inference* 123 (1996), page 1
- Marc O Ernst and Martin S Banks. "Humans integrate visual and haptic information in a statistically optimal fashion". In: *Nature* 415.6870 (2002), pages 429–433
- Robert A Jacobs. "Optimal integration of texture and motion cues to depth". In: *Vision research* 39.21 (1999), pages 3621–3629
- David C Knill and Jeffrey A Saunders. "Do humans optimally integrate stereo and texture information for judgments of surface slant?" In: *Vision research* 43.24 (2003), pages 2539–2558
- Wei Ji Ma et al. "Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space". In: *PloS one* 4.3 (2009), e4638
- Harry McGurk and John MacDonald. "Hearing lips and seeing voices". In: *Nature* 264.5588 (1976), pages 746–748
- Julia Trommershauser, Konrad Kording, and Michael S Landy. *Sensory cue integration*. Oxford University Press, 2011
- Robert J van Beers, Anne C Sittig, and Jan J van der Gon Denier. "How humans combine simultaneous proprioceptive and visual position information". In: *Experimental brain research* 111.2 (1996), pages 253–261

5.10 Problems

Problem 5.1 An observer combines conditionally independent cues A and B with Gaussian measurement noise. When B becomes more reliable, the observer's estimate will

a) shift towards A;
b) shift towards B;
c) stay the same;
d) there is insufficient information to answer

Problem 5.2 True or false? Explain.

a) In the cue combination model of this chapter, the measurements are assumed to be independent of each other.
b) Conflicts between two measurements generated by a single source rarely occur in real-world perception.

Problem 5.3 [1] Within the model for combining two independent cues with Gaussian measurement noise under a flat prior, show, using equations, that the variance of the posterior:

a) Is never greater than 50% of the largest variance of the single-cue likelihoods.
b) Is always between 50% and 100% of the smallest variance of the single-cue likelihoods.
c) Is (a) still true if the prior is Gaussian rather than flat? If so, prove it. If not, give a counterexample.

[1] This problem was suggested in 2016 by Nick Johnson, then a PhD student at New York University.

d) Is (b) still true if the prior is Gaussian rather than flat? If so, prove it. If not, give a counterexample.

Problem 5.4 Suppose $p_1(x), p_2(x), \ldots, p_N(x)$ are Gaussian distributions, where $p_i(x)$ (for every $i = 1, \ldots, N$) has mean μ_i and precision J_i. We multiply these distributions, then normalize:

$$q(x) = kp_1(x)p_2(x)\cdots p_N(x), \tag{5.26}$$

where k is such that $q(x)$ is normalized. We are interested in showing that $q(x)$ is a Gaussian distribution with precision (inverse variance)

$$J_q = \sum_{i=1}^{N} J_i. \tag{5.27}$$

and mean

$$\mu_q = \frac{1}{J_q}\sum_{i=1}^{N} J_i\mu_i. \tag{5.28}$$

a) Method 1: Use direct calculation as in Chapter 3.

b) Method 2: Show by induction; for $N = 1$, the equations are trivially true. Assume it is true for some N, and show that then it would also be true for $N + 1$.

Problem 5.5 Prove Eq. (5.18). The approach you need to take shares a lot with the previous problem.

Problem 5.6 An observer infers a stimulus s from a sequence of $x_{\text{obs},1}, x_{\text{obs},2}, \ldots, x_{\text{obs}}T$ made on a single trial. The distribution of the t^{th} measurement, $p(x_t|s)$, is Gaussian with mean s and variance σ^2 (identical for all measurements). The stimulus distribution is Gaussian with mean μ and variance σ_s^2.

a) What are the mean and variance of the posterior? You may start with the equations in Section 5.4.

b) For a given true stimulus s, we define *relative bias* as the difference between the trial-averaged posterior mean estimate and s itself, divided by the difference between the mean of the stimulus distribution and s. Derive an expression for relative bias in terms of μ, σ, σ_s, and T. Simplify the expression as much as you can.

c) Interpret the expression in (b): explain intuitively how the dependencies on the variables make sense.

d) Compute the variance of the posterior mean estimate for given s.

e) Plot this variance as a function of T for all nine combinations of $\sigma_s \in \{1, 2, 5\}$ and $\sigma \in \{1, 2, 5\}$. Create one plot for each value of σ_s, for a total of three plots, each containing three curves (color-coded).

f) Interpret the plots in (e): explain intuitively how the shapes of the functions make sense.

Problem 5.7 In this problem, we examine *suboptimal* estimation in the context of cue combination. Suppose an observer estimates a stimulus s from two conditionally independent, Gaussian-distributed measurements, $x_{\text{obs},1}$ and $x_{\text{obs},2}$. The prior is flat.

a) Express the posterior mean estimate in terms of the measurements.

b) What is the variance of the posterior mean estimate across trials?

c) Now suppose the observer uses an estimator of the form $\hat{s} = wx_{\text{obs},1} + (1-w)x_{\text{obs},2}$, where w can be *any constant* (not necessarily the one in Eq. (5.13)). Show that this estimate is unbiased (just like the posterior mean estimate); this means that the mean of the estimate for given s is equal to s.

d) What is the variance of this estimate as a function of w? Plot this function. At which value of w is it minimal, and does this value make sense? State your final conclusion in words.

e) Which of the conclusions in (b) and (c) break down when we consider estimates that are general linear combinations of measurements, $\hat{s} = w_1 x_{\text{obs},1} + w_2 x_{\text{obs},2}$, where w_1 and w_2 can be any constants? Explain.

f) What quantity does the posterior mean estimate minimize in this more general setting?

Problem 5.8 In Chapters 3 and 4, we were able to derive analytical expressions for the posterior distribution. For more complex psychophysical tasks (e.g. later in this book), however, analytical solutions often do not exist. In such a case, we can use numerical methods to approximate the distribution of interest. To get some familiarity with such methods, we will reconsider the cue combination model described in this chapter, but we will now compute the distribution of posterior mean estimates numerically. We assume that the experimenter introduces a cue conflict between the auditory and the visual stimulus: $s_1 = 5$ and $s_2 = 10$. The standard deviation of the auditory and of the visual noise is $\sigma_1 = 2$ and $\sigma_2 = 1$, respectively. We assume a flat prior over s.

a) Randomly draw an auditory measurement $x_{\text{obs},1}$ and a visual measurement $x_{\text{obs},2}$ from their respective distributions.

b) Plot the corresponding elementary likelihood functions, $p(x_{\text{obs},1}|s)$ and $p(x_{\text{obs},2}|s)$, in one figure.

c) Calculate the combined likelihood function, $p(x_{\text{obs},1}, x_{\text{obs},2}|s)$, by numerically multiplying the elementary likelihood functions. Plot this function.

d) Calculate the posterior distribution by normalizing the combined likelihood function. Plot this distribution in the same figure as the likelihood functions.

e) Numerically find the posterior mean estimate of s, i.e. the value of s at which the posterior distribution is maximal.

f) Compare with the posterior mean estimate of s computed from Eq. (5.11) using the measurements drawn in (a).

g) In the above, we simulated a single trial and computed the observer's posterior mean estimate of s, given the noisy measurements on that trial. If an analytical solution does not exist for the distribution of the posterior mean estimates, we can repeat the above procedure many times to approximate this distribution. Here, we practice this method even though an analytical solution is available in this case. Draw 100 pairs $(x_{\text{obs},1}, x_{\text{obs},2})$ and numerically compute the observer's posterior mean estimate for each pair as in (e).

h) Compute the mean of the posterior mean estimates obtained in (g) and compare with the mean estimate predicted using Eq. 5.14.

i) Make a histogram of the posterior mean estimate.

j) *Relative (auditory) bias* is defined as the mean posterior mean estimate minus the true auditory stimulus, divided by the true visual stimulus minus the true auditory stimulus. Compute relative auditory bias for your set of estimates.

Problem 5.9 A major assumption in the derivation of our cue combination model was that the measurements x_1 and x_2 are conditionally independent, Eq. (5.1). Here, we consider a generalization in which they are not. We replace Eq. (5.14) by a *bivariate normal distribution* with the same mean s for both measurements, standard deviations σ_1 and σ_2, and correlation ρ:

$$p(x_1, x_2|s) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{1-\rho^2}\left(\frac{(x_1-s)^2}{2\sigma_1^2} + \frac{(x_2-s)^2}{2\sigma_2^2} - \frac{\rho(x_1-s)(x_2-s)}{\sigma_1\sigma_2}\right)}. \tag{5.29}$$

Assume that the prior is flat.

a) Step 2: Derive equations for the mean and variance of the posterior over s. Hint: follow the steps of Problem 3.3 (completing the square).

b) Do a sanity check by setting $\rho = 0$ (conditionally independent measurements). You should get back Eqs. (5.21) and (5.22).

c) Plot the variance of the posterior as a function of ρ for $\sigma_1 = 1$, $\sigma_2 = 2$.

d) Interpret the plot. Specifically, explain how the dependencies of the variables make sense.

Problem 5.10 Combining survey data from multiple people to get at the population mean can also be considered cue combination. Suppose a pollster asks respondents to what extent they agree or disagree with a particular statement. For simplicity, we represent amount of agreement by a real-valued variable x. As individuals in the population differ, we model x as following a normal distribution with mean s and variance σ^2. The variance will generally depend on how and when the data are collected. To the pollster, the variable of interest is s. The pollster has a uniform prior over s. The pollster collects responses $x_{\text{obs},1}, x_{\text{obs},2}, \ldots, x_{\text{obs},N}$ from N people; these responses are assumed to be conditionally independent given s.

a) Show mathematically that the (normalized) likelihood function of s is a normal distribution with mean equal to the sample mean of the responses, denoted by $\bar{x}$, and variance $\frac{\sigma^2}{N}$. Thus, the maximum-likelihood (ML) estimate of s is simply $\bar{x}$.

b) A poll aggregator is trying to do "cue combination" of one poll with N_1 respondents and individual variance σ_1^2, another with N_2 respondents and individual variance σ_2^2 (subscripts now refer to different polls, not to individuals within the poll). We denote the respective ML estimates by $\hat{s}_1$ and $\hat{s}_2$. Show that the combined ML estimate of s is equal to

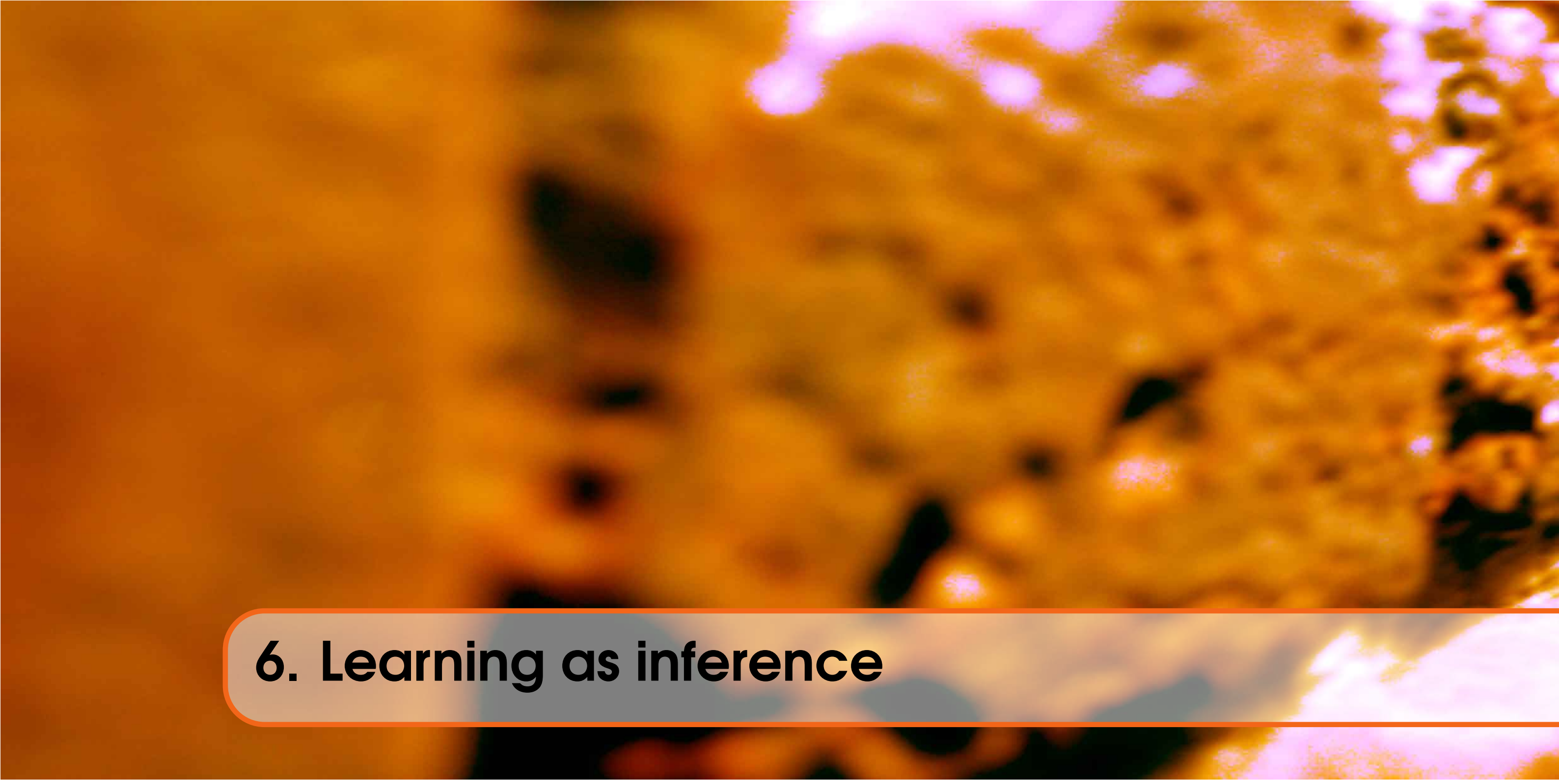
$$\hat{s}_{\text{ML, combined}} = \frac{w_1\hat{s}_1 + w_2\hat{s}_2}{w_1 + w_2}, \tag{5.30}$$

where

$$w_i = \frac{N}{\sigma_i^2} \tag{5.31}$$

for $i = 1, 2$.

c) Interpret this result.



6. Learning as inference

How do we learn from observations?

All animals learn, and learning occurs in all domains of behavior, from learning to categorize objects as an infant to learning to play a new musical instrument all the way down to learning how to control one's muscles while picking up a lightweight object. Learning allows the observer/agent to perform better at a task or to be better adapted to an environment.

In most of this book, we assume that the generative model and its parameters are already known to the decision-maker. This parallels the logic of many empirical Bayesian modeling studies, as they often assume that such learning is already complete. This can be made plausible by some combination of giving clear instructions, exposing the participant to a large number of example stimuli, testing the participant on properties of the relevant distributions, and providing a block of training trials (the data of which do not get analyzed). However, the process of learning generative models is a rich area of study by itself. In the natural world, learning the statistical properties of the environment happens at an evolutionary timescale as well as during the development of an individual. Much of such learning takes place in the absence of explicit feedback; for example, infants learn to parse spoken language in part by keeping track of co-occurrence frequencies of syllables. Learning continues throughout an individual's lifetime.

This chapter focuses on learning as a statistical estimation process[1]. We consider several examples of learning parameters of a generative model from a sequence of imperfect observations, and we model this learning in a Bayesian way.

There is a close connection (see **Table 6.1**) between learning the value of a fixed world state and evidence accumulation over time, as discussed in Chapter 5. In fact, the basic form of the generative model is identical: a single world state, generating conditionally independent observations. Usually, however, evidence accumulation plays out on a shorter time scale (tens of milliseconds to seconds), with all observations being made within a single trial and the world state still changing across trials. In an experimental setting, no judgments are typically required while the evidence accumulation

[1] Some call what we are covering here *statistical learning*, a term we do not use because there is a domain of cognitive science applying this term to language acquisition and a branch of machine learning, statistical learning theory, which focuses on risk minimization

	Sensory evidence accumulation	**Learning**
Variable of interest	World state (stimulus)	Usually a parameter of a world state distribution
Time scale	Trial (up to a few seconds)	Many trials (seconds to years)
Observations	Sensory observations (measurements) within a trial	Stimuli (often noiseless) across trials

Table 6.1: Comparing inference in sensory evidence accumulation with inference during learning. While there is an underlying continuum, the field still mostly distinguishes accumulation and learning.

occurs. In learning, each trial corresponds to one observation, and the world state is fixed across a much longer time scale, typically from minutes to years. In an experimental setting, the progress of learning might be probed by asking the subject for a judgment on each trial. As a result of the different timescale, the reason learning is hard is usually not measurement noise (the effect of which becomes negligible when the observation time is long) but stochasticity in the world.

Plan of the chapter

We start by what is perhaps the simplest form of learning: learning the probability of a binary event from a sequence of observations. Subsequently, we discuss learning the precision of Gaussian distribution from observations. Next, we consider learning the parameter of a relationship between two variables. Finally, we study the learning of categorically distinct causal models of the world. We describe a link between Bayesian learning and reinforcement learning in Section 6.2.

Plan of the chapter

We begin by considering how an observer can learn the probability of a binary event from repeated observations. We will show that the posterior probability can be written in recursive form, updating after each time step. We will see that, for a given number of observations of each outcome, the order of the observations affects the evolution of the posterior, but not the final posterior. We will discuss nonuniform priors and conjugate priors and the link between Bayesian learning and reinforcement learning. We will discuss learning the parameters of a normal distribution, the slope of a linear relationship, and the structure of a causal model.

Note on notation: So far, we used a subscript “obs" to indicate a specific value of the observation in Step 2. However, this causes clutter and is not common in the literature. For these reasons, we will from now on drop the subscript. Please keep in mind that in Step 1, we the observation or measurement is always a *variable*, whereas in Step 2, it is always a *specific value* of that variable.

6.1 Learning the probability of a binary event

Imagine you are shipwrecked and stranded on a desert island. The good news is that it rains frequently. The bad news is that the interval between days with and without rain is unpredictable. To ration the rainwater that you save, you would like to estimate the probability that it will rain on a given day. After three days, you have observed:

dry – rain – dry

Your best guess of the probability at this point may be $\frac{1}{3}$. However, given the small number of days, you are pretty uncertain about that guess. As a devoted Bayesian with a lot of free time, you would

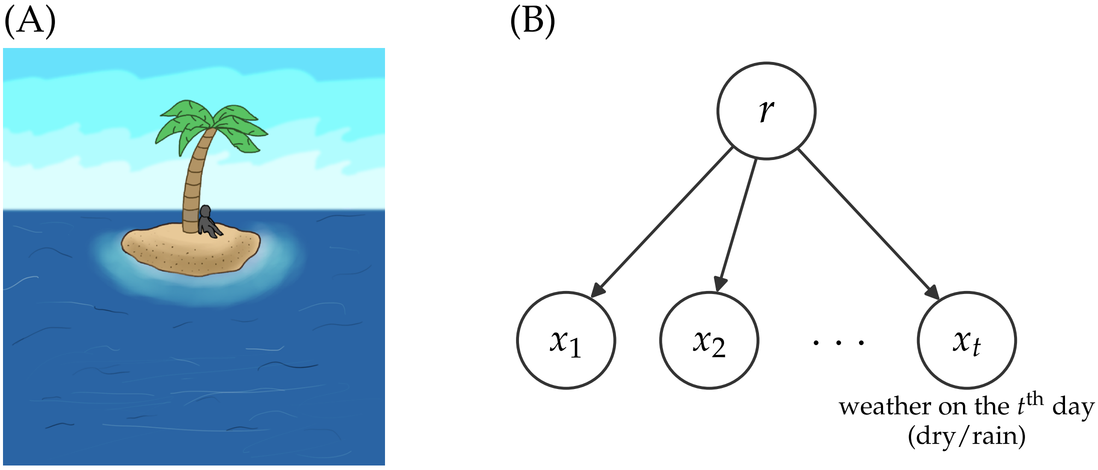


Figure 6.1: Generative model of the task of learning a Bernoulli probability.

like to quantify that uncertainty, or even better, calculate a posterior probability distribution over the probability that it will rain. In other words, what is the probability that that probability of rain is 0.1? 0.3? 0.5? 0.9? etc. So far in this book, we have not yet seen a probability as the world state variable of interest, but it's perfectly legitimate, very common in the real world, and it doesn't really change the math. The only thing to get used to is the slightly awkward expression "probability over probability".

Step 1 is to formulate the generative model. The graphical model is shown in **Fig. 6.1B**.

It has a top-level variable representing the probability of rain on a given day (i.e., the rainfall rate), which we will denote by r. As mentioned in the introductory section, we assume that this world state does not change over a very long time; in our example, that could mean that over the period of a month, the probability of rain is constant. Obviously, this is a simplification, and it is also possible to learn a slowly world state; we will consider this situation later.

We do have to assume a prior distribution over r. You may realize that across the world, there are few places where it rains almost every day (say, $r > 0.9$) or almost never (say, $r < 0.1$). Perhaps, you believe in the absence of any observations, that r is most probable to lie in the interval $[0.4, 0.6]$. These are examples of aspects of the prior distribution over r. For the moment, we assume a completely uniform distribution over r:

$$p(r) = 1. \tag{6.1}$$

Exercise 6.1 Why is this a properly normalized distribution? ■

At the bottom of the generative model, we find the sequence of binary observations corresponding to the observed weather (rain/dry) on days 1 to t, which we denote by $x_1, x_2, \ldots, x_t$. Each x_i can be 0 (dry) or 1 (rain). On a given day,

$$p(x_i = 1) = \pi \tag{6.2}$$

$$p(x_i = 0) = 1 - \pi \tag{6.3}$$

This is an example of a *Bernoulli process*: a random variable with two possible outcomes, with each outcome having a fixed probability. The best-known example of a Bernoulli process is a coin flip and in analogy to this we model rain vs dry as a weighted coin flip. We do not have a single observation but a sequence of them, one on each day. For simplicity, you assume that the weather is independent across days, conditioned on r[2]. Then, we can write for the probability distribution associated with a sequence:

$$p(x_1, x_2, \ldots, x_t | r) = p(x_1|r) p(x_2|r) \cdots p(x_N|r), \tag{6.4}$$

which can be written in product notation as

$$p(x_1, x_2, \ldots, x_t | r) = \prod_{i=1}^{t} p(x_i|r). \tag{6.5}$$

This concludes the specification of the generative model. We are now ready to do inference!

In Step 2, we compute a posterior distribution over r given your observations so far (dry – rain – dry), which we will denote by $\mathbf{x}_{\text{obs}}$. We start with Bayes' rule,

$$p\,(r|\mathbf{x}) \propto p\,(r)\,p\,(\mathbf{x}|r) \tag{6.6}$$

We now use Eq. (6.1) for the prior and Eq. (6.4) for the likelihood:

$$p\,(r|\mathbf{x}) \propto p\,(x_{\text{obs}}1|r)\,p\,(x_2|r) \cdots p\,(x_t|r) \tag{6.7}$$

Now, we realize that every factor in this product is equal to either r or $1-r$, as those are the only possible values. Thus, we can simplify the expression to

$$p\,(r|\mathbf{x}) \propto r^{n_{\text{rain}}}\,(1-\pi_{\text{rain}})^{n_{\text{dry}}}, \tag{6.8}$$

where n_{rain} and n_{dry} are the numbers of rain and dry days observed so far (with $n_{\text{rain}} + n_{\text{dry}} = t$. This type of distribution is called a *Beta distribution* (see Box 6.1). The normalization is provided by the so-called *Beta function*, but this is not essential to our understanding.

Box 6.1 — Beta distribution. The beta distribution is defined over a random variable Y that takes values between 0 and 1. Most often, that random variable is a probability itself, so the beta distribution is a probability distribution over a probability variable. The probability density of the beta distribution is

$$p(y) = \frac{1}{\mathrm{B}(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1}. \tag{6.9}$$

Here, α and β are the parameters of the distribution, both restricted to be positive; $\mathrm{B}(\cdot,\cdot)$ denotes the *beta function*, a special function whose role (and in fact, definition) is to normalize the beta distribution. For the purposes of this book (and for most of Bayesian modeling), we do not need to know anything else about the beta function. In case you need it, all numerical computation packages have the beta function pre-programmed. We show several examples of the beta distribution in **Fig. 6.2**. The mean of a beta-distributed random variable Y is

$$\mathbb{E}[Y] = \frac{\alpha}{\alpha+\beta}. \tag{6.10}$$

Its variance is

$$\mathrm{Var}(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{6.11}$$

[2] Weather across days is, incidentally, very far from conditionally independent, and places can have long spells of dry vs rain days.

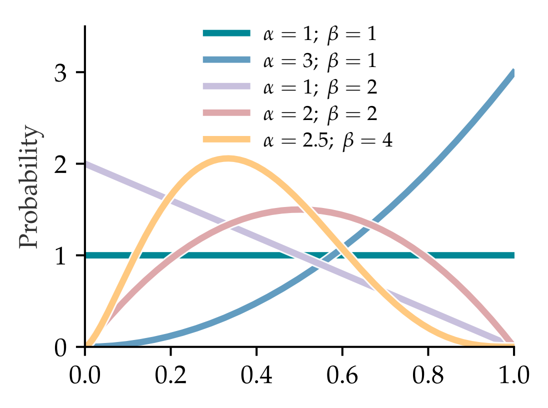


Figure 6.2: Examples of the beta distribution.

What is important about Eq. (6.8) for the posterior is that every time you observe a rain day, it gets multiplied with an increasing function of r, namely $f(r) = r$, and every time you observe a dry day, it gets multiplied with a decreasing function, namely $f(r) = 1 - r$. These individual factors are referred to as *instantaneous* or *momentary likelihood functions*, as they are tied to an additional observation at a single time point. **Fig. 6.3** shows the evolution of the posterior when the observations are dry-rain-dry. We want to note here that the integration of the new data every day is modeled really just an application of Bayes rule with the day's likelihood function.

Exercise 6.2 Plot the posterior you get if you observe only rain days, for different numbers of rain days. How long would it take to convince you that the probability of rain is greater than 90%?

6.1.1 Prediction

Having computed the posterior distribution, $p(r|x_1, x_2, \ldots, x_t) = p(r|\mathbf{x})$, you may wonder, as you fall asleep thirsty after spending several days on the desert island: what is the probability that it will rain the following day? In general, we can compute the probability of any future outcome by integrating its probability under each world state, multiplied by the posterior probability of that world state:

$$p(x_{t+1} = 1|\mathbf{x}_{\text{obs}}) = \int p(x_{t+1} = 1|r)p(r|\mathbf{x}_{\text{obs}})dr \tag{6.12}$$

In essence, we are stating that our belief it will rain tomorrow is our belief that r has a given value AND it will rain if r has that value, OR that r has another value AND it will rain if r has that other value, and so on. Because we are dealing here with a binary outcome whose probability, given r, is simply r itself, our integral reduces to:

$$p(x_{t+1} = 1|\mathbf{x}_{\text{obs}}) = \int rp(r|\mathbf{x}_{\text{obs}})dr \tag{6.13}$$

The reader will recognize that this is the formula for the mean of the posterior distribution. Thus, in this example, the posterior mean estimate is the probability with which we expect rain to occur the next day.

It can be shown that the posterior mean estimate is equal to

$$\hat{r}_{\text{PM}} = \frac{n_{\text{rain}} + 1}{t + 2}. \tag{6.14}$$

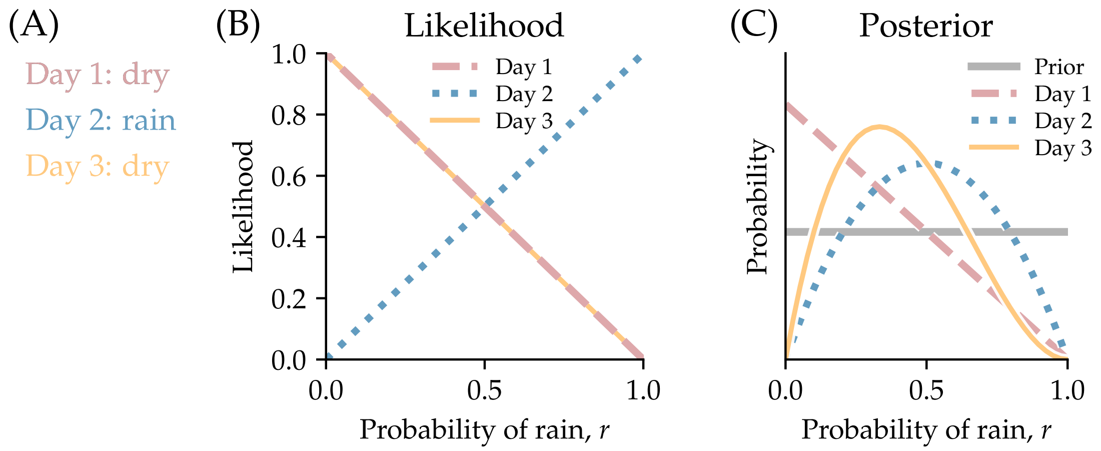


Figure 6.3: Evolution of the posterior over a Bernouilli probability. **(A)** Observations on days 1, 2, and 3. **(B)** Instantaneous likelihoods on days 1, 2, and 3. **(C)** Posterior distributions on days 1, 2, and 3.

This expression was derived in the 18th century by Pierre-Simon Laplace and is known as Laplace's Rule of Succession. This posterior mean estimate is different from the mode of the posterior (the MAP estimate), which is

$$\hat{r}_{\text{MAP}} = \frac{n_{\text{rain}}}{t}. \tag{6.15}$$

Step 3 does not apply, since we are considering neither internal noise nor decision noise. In other words, conditioned on the stimuli, there is no variability in the estimate.

6.1.2 Update equations

Just as when we discussed evidence accumulation in Section 5.5, we can write the posterior in recursive form, updating it after each time step. Specifically, the analogue of Eq. (5.25) is

$$p(r|x_{\text{obs},1},\ldots,x_{\text{obs},t+1}) \propto p(r|x_{\text{obs},1},\ldots,x_{\text{obs},T})p(x_{\text{obs},t+1}|r), \tag{6.16}$$

where it is understood that $p(r|x_{\text{obs},1},\cdots,x_{\text{obs},t})$ for $t=0$ is the prior distribution. This is possible thanks to the conditional independence assumption, Eq. (6.4), which embodies that we believe that days only depend on the overall rain probability, that all days are conditionally independent given this probability. However, writing the posterior using an update equation is not so useful, since the explicit solution, Eq. (6.8) is straightforward.

6.1.3 Uncertainty

As in Section 3.4.1, the standard deviation of the posterior can be used as a measure of uncertainty.

Exercise 6.3 Using Eq. (6.11) for the variance of a beta-distributed random variable, show that in our case, the standard deviation is

$$\text{Std}\,(r|\mathbf{x}) = \frac{1}{t+2}\sqrt{\frac{(n_{\text{rain}}+1)(n_{\text{dry}}+1)}{t+3}}. \tag{6.17}$$

■

6.1.4 Binomial distribution

The order of the observation (rain days and dry days) affects the evolution of the posterior, but not the final posterior. For example, the posterior would be identical for dry-rain-dry as for rain-dry-dry or dry-dry-rain. Stated differently, we could in Step 1 alternatively have summarized the sequence of binary observations as a single count. Then, the generative model would have been

$$p(n_{\text{rain}}|r) = \binom{t}{n_{\text{rain}}} n_{\text{rain}}^{r} (t - n_{\text{rain}})^{(1-r)}. \tag{6.18}$$

This is an example of a *binomial distribution*. A binomial distribution is defined over counts (integers), starting at 0. If we had used this generative model, it would not have changed any of our inference in Step 2.

Exercise 6.4 Why not? ■

6.1.5 Non-uniform prior

So far, we have assumed a uniform distribution over r (see Eq. (6.1)). To generalize this, it is common to choose the prior to be a beta distribution:

$$p(r) \propto \pi^{\alpha_0 - 1}(1-\pi)^{\beta_0 - 1}, \tag{6.19}$$

where α_0 and β_0 are the parameters (both must be positive) and we left out the normalization. When $\alpha_0 = \beta_0 = 1$, the prior is uniform. Using Eq. (6.19), the posterior distribution over r is again beta-distributed, but now with parameters

$$\alpha = \alpha_0 + n_{\text{rain}} \tag{6.20}$$

$$\beta = \beta_0 + n_{\text{dry}}. \tag{6.21}$$

Exercise 6.5 Show mathematically why this is the case. ■

Let's take stock of what we did. We showed earlier that the likelihood corresponding to a Bernoulli or binomial generative model is proportional to a beta distribution. We have now shown that multiplying a beta prior with a beta likelihood produces a beta posterior. This is elegant, because going from prior to posterior, we only need to update the parameters of the distribution, just as we did in Chapter 3 when going from a normal prior to a normal posterior.

Definition 6.1.1 — Conjugate prior. For a given generative model, a prior distribution that is such that the posterior is the same type of probability distribution as the prior is called a *conjugate prior* for that generative model.

In other words, the beta distribution is the conjugate prior for the Bernoulli (or binomial) distribution.

But how can we justify choosing a prior merely based on elegance? Shouldn't a prior reflect the statistics of the world? Yes and no. In practice, it is difficult to know or control the prior distribution that an observer has over a probability. The best we can often do is choose a sufficiently flexible distribution (usually that means, having at least two parameters) and fit the parameters to an individual subject's data (and if applicable, to a particular experimental condition). From this perspective, a beta prior is as good as any alternative, in which case the elegance can be a tiebreaker.

Finally: whereas being stranded on an island may not be particularly relevant to cognitive scientists and neuroscientists, many problems that are relevant have the same mathematical structure.

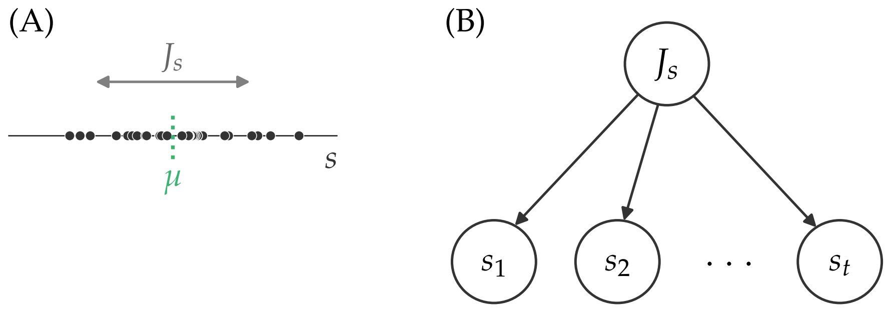


Figure 6.4: Generative model of the precision learning task.

For example, in an iterated trust game, I might be figuring out how trustworthy my partner from binary observations. There, trustworthiness would take the place of the probability of rain in our island example.

6.2 Linking Bayesian learning to reinforcement learning

There is an interesting connection between Bayesian learning and reinforcement learning. In Section 6.1.2, we wrote down an update equation for the posterior when the prior is uniform, Eq. (6.16). We could similarly write down a recursive update equation for the posterior mean estimate at time t (which is the expected value of whether it will rain at time $t+1$):

$$\hat{r}_t = \hat{r}_{t-1} + \frac{1}{t+2}(x_t - \hat{r}_{t-1}). \tag{6.22}$$

(We left out the “PM" subscript to avoid clutter.) This equation resembles the Rescorla-Wagner rule in reinforcement learning (Box 6.2), in which the value of a state gets updated based on the difference between a received reward and the predicted reward – the *prediction error*. In Eq. (6.22), the expectation of rain after day $t-1$ acts as the value of the state at t, and whether it actually rained on day t acts as the received reward. The factor $\frac{1}{t+2}$ acts as the *learning rate*, which represents to what extent the prediction gets updated. It makes intuitive sense that the learning rate decreases as t increases: as the prediction is based on more observations, any one observation has less power to change the prediction. This is true beyond this specific example: a Bayesian learning rule typically has a learning rate that decreases in time, in a way that is fully dictated by the Bayesian formalism. This time dependence contrasts with the standard Rescorla-Wagner rule, in which the learning rate is a constant (although that rule can simply be generalized).

We will prove Eq. (6.22) in Problem **??**, but we can do a numerical example here. Since we start with a uniform prior, the prior expectation of rain is $\hat{r}_0 = \frac{1}{2}$. Let's say that it rains on your first day on the island. Then $x_1 = 1$, and the “prediction error" is $x_1 - \hat{r}_0 = 1 - \frac{1}{2} = \frac{1}{2}$. The posterior expectation of rain after day 1 is $\hat{r}_1 = \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{2} = \frac{2}{3}$, which is indeed the expected value of a posterior $p(r|s_1) \propto r$. Although we have linked Bayesian learning and reinforcement learning here, several caveats are in place. First, the link was only at the level of the *read-out* of the posterior, not of the full posterior distribution. The Rescorla-Wagner rule and many other reinforcement learning models do not treat uncertainty explicitly, although attempts to remedy this are ongoing. Second, even though in our case, the expectation of rain was naturally interpreted as being valuable, not all Bayesian learning can be interpreted as the updating of a value-like quantity; the next section will illustrate this. Third, there is no guarantee in Bayesian learning that the difference between two successive updates is proportional to a “prediction error".

Box 6.2 — Rescorla-Wagner model. One of the simplest learning models is the Rescorla-Wagner model, which describes how the value of state is updated over time based on rewards. We denote by $V_t(s)$ the value of state s at time t, which can be thought of as the expected reward of s in the long run. The agent receives a reward R_t at time t while being in state s (actions are only implicit in this model). Then the value $V_t(s)$ of that state can be updated in the next time step, to

$$V_{t+1}(s) = V_t(s) + \lambda(R_t - V_t(s)). \tag{6.23}$$

Here, $R_t - V_t$ is the difference between the received reward and the expected reward; this difference is also called the *reward prediction error*. The parameter λ is called the *learning rate*[a]. It is a number between 0 and 1 that describes how responsive the agent's value function is to the reward prediction error. If $\lambda = 0$, the value function is wholly unresponsive. If $\lambda = 1$, we have $V_{t+1}(a) = R_t$, which means that the old value is irrelevant and the new value is simply the received reward; this is typically much *too* responsive. In practice, λ is small but depends on the individual and on the experimental condition.

The basic structure of the Rescorla-Wagner model, in which value gets modified by a scaled version of a prediction error, is common to much of *reinforcement learning* and its application to neuroscience. Reinforcement learning is a branch of machine learning that studies how entities interacting with the world can figure out which actions are rewarding in what world states. These entities are typically called *agents* instead of observers, because their actions are part of a causal chain that continues after the action. A reinforcement learning model with the same basic structure but that also takes into account future rewards is *Q-learning*. Machine learning now uses a set of sophisticated algorithms (Deep Reinforement Learning) to solve such problems, although rules similar to the Rescorla-Wagner rule are still widely used to update the connection strengths. A specific case with strong links to neuroscience is the *delta rule*.

■

[a] A more common notation for the learning rate is λ, but we already used that symbol in this chapter for one of the parameters of the beta distribution.

6.3 Learning the precision of a normal distribution

In this section, we consider a very simple form of unsupervised learning, namely how an observer might learn the precision of a distribution from samples. We consider the Gaussian stimulus distribution considered in Chapters 3 and 4:

$$p(s) = \frac{1}{\sqrt{2\pi\sigma_s^2}} e^{-\frac{(s-\mu)^2}{2\sigma_s^2}}. \tag{6.24}$$

For later convenience, we reparametrize this in terms of stimulus precision, $J_s \equiv \frac{1}{\sigma_s^2}$. We also make the dependence of the distribution on stimulus mean μ and stimulus precision J_s explicit:

$$p(s|\mu, J_s) = \sqrt{\frac{J_s}{2\pi}} e^{-\frac{J_s}{2}(s-\mu)^2}. \tag{6.25}$$

In Chapter 3, we assumed that the observer knows this distribution – specifically, that they know μ and J_s. In reality, this information needs to be learned. For simplicity, we consider the case that μ is known but J_s needs to be learned; thus, from now on, we do not mention μ as a variable to be conditioned on.

Step 1: Generative model. The generative model has the world state variable of interest, J_s at the top. The "measurements" – at the bottom – are a set of stimuli $s_1, \ldots, s_T$, which we will collectively denote by a vector $\mathbf{s}$. We assume that the stimuli are independently drawn conditioned on μ and J_s, so that

$$\begin{aligned} p(\mathbf{s}|J_s) &\equiv p(s_1, s_2, \ldots, s_t|J_s) && (6.26)\\ &= p(s_1|J_s)p(s_2|\mu, J_s)\cdots p(s_t|J_s). && (6.27)\end{aligned}$$

In product notation, this can be written as $\prod_{i=1}^{t} p(s_i|\mu, J_s)$. Each individual s_i follows Eq. (6.25), so that

$$p(s_i|J_s) = \sqrt{\frac{J_s}{2\pi}} e^{-\frac{J_s}{2}(s_i-\mu)^2}. \tag{6.28}$$

Step 2: Inference. Inference consists of learning the parameter J_s from specific observed samples s_1, s_2, s_t. Thus, the posterior of interest is $p(J_s|\mathbf{s}_{\text{obs}})$. To calculate this posterior, we apply Bayes' rule and assume a uniform prior:

$$\begin{aligned} p(J_s|\mathbf{s}_{\text{obs}}) &\propto p(\mathbf{s}_{\text{obs}}|J_s)p(J_s) && (6.29)\\ &\propto p(\mathbf{s}_{\text{obs}}|J_s). && (6.30)\end{aligned}$$

We now make use of Eq. (6.27) to write the posterior as proportional to a product of instantaneous likelihood functions, each of which is based on an individual observed stimulus:

$$\begin{aligned} p(J_s|\mathbf{s}_{\text{obs}}) &\propto \prod_{i=1}^{t} p(s_i|\mu, J_s) && (6.31)\\ &= \prod_{i=1}^{t}\left(\sqrt{\frac{J_s}{2\pi}} e^{-\frac{J_s}{2}(s_i-\mu)^2}\right) && (6.32)\\ &\propto J_s^{\frac{t}{2}} e^{-\frac{J_s}{2}\sum_{i=1}^{t}(s_i-\mu)^2} && (6.33)\end{aligned}$$

This expression has several interesting aspects. First, the observed stimuli only come in a specific combination, namely $\sum_{i=1}^{t}(s_i-\mu)^2$. This is the sum of the squares of the observed stimuli to the known mean μ. This combination makes sense, since the larger its value is, the lower precision tends to be. Second, the dependence of J_s is one we have not seen before. This type of distribution is called a *gamma distribution* (see Box 6.3). It is a common distribution for variables that cannot take negative values (such as J_s). Thus, in our example, each individual likelihood as well as the posterior has a gamma distribution form. If we had chosen a non-uniform prior, a convenient choice would have been a gamma distribution as well. Using Definition 6.1.1, we can say that the gamma distribution is the conjugate prior for inferring the precision of a normal distribution.

Just as in Sections 5.5 and 6.1.2, the posterior could be written in the form of a recursive update equation. In **Fig. 6.6**, we show snapshots from the learning process.

Box 6.3 — Gamma distribution. The gamma distribution is defined over a random variable Y that takes values on the positive real axis, such as precision in our case study. The probability density of the gamma distribution is

$$p(y) = \frac{1}{\Gamma(k)\theta^k} y^{k-1} e^{-\frac{y}{\theta}}. \tag{6.34}$$

Here, k and θ are the parameters of the distribution, both restricted to be positive. k is called the shape parameter and θ the scale parameter. $\Gamma(\cdot)$ denotes the *gamma function*, a special function

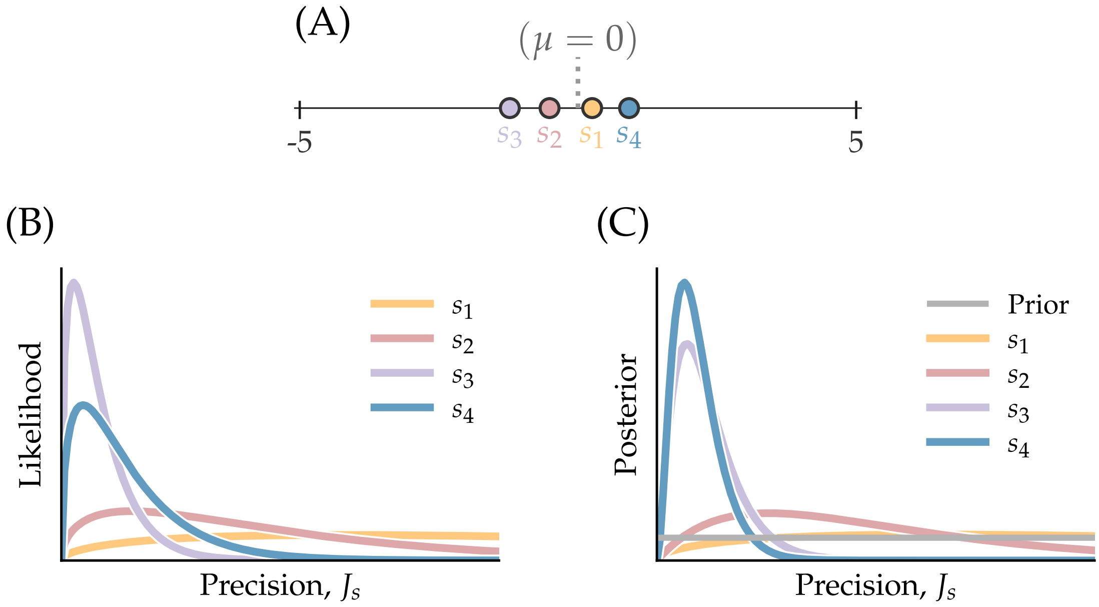


Figure 6.6: Evolution of the posterior over J_s. **(A)** Observations. **(B)** Normalized likelihood functions associated with the most recent observation. **(C)** Posterior distributions given all observations up to the current time. The normalized likelihood functions and the posterior distributions are all gamma distributions. The prior distribution is improper (not normalizable; see Box 3.10), which we indicate with a dashed line.

that is also pre-programmed in all numerical computation packages. We show several examples of the gamma distribution in **Fig. 6.5**. The mean of a gamma-distributed random variable Y is

$$\mathbb{E}[Y] = k\theta. \tag{6.35}$$

Its variance is

$$\mathrm{Var}(Y) = k\theta^2. \tag{6.36}$$

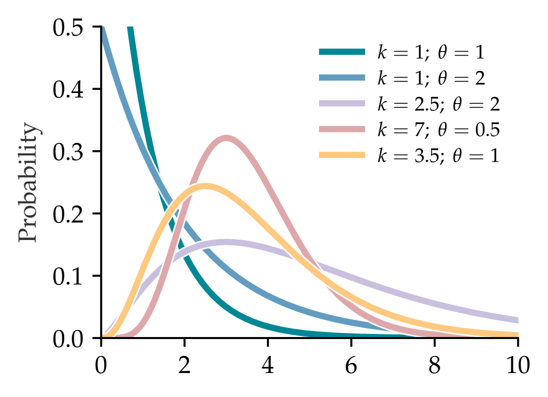


Figure 6.5: Examples of the gamma distribution.

As the optional last part of Step 2, we might want to commit to a read-out. If our objective is to minimize the expected squared error in estimating J_s, we should use the posterior mean estimate.

(This objective can be challenged, as the squared error is mostly meaningful if a variable can take values across the entire real line.) We work this out in Problem 6.6. However, it is now not possible to write the updating of the posterior mean estimate of J_s in Rescorla-Wagner form.

Exercise 6.6 Try this anyway and point out what the problem is. ■

Again, Step 3 does not apply, since we are considering neither internal noise nor decision noise.

6.3.1 Why not infer variance?

We have formulated the problem as one of inferring the precision parameter of a normal distribution. However, we could have alternatively formulated it as an inference of the variance or the standard deviation of a normal distribution. This would not have been equivalent, because a uniform prior over precision is not the same as a uniform prior over variance. (If this is not clear, read Appendix Section B.12.1 on transformations of random variables.) However, if the prior over variance were chosen uniform, then the posterior would have been an *inverse gamma distribution*. It is generally slightly simpler to work with gamma distributions than with inverse gamma distributions.

6.4 Learning the slope of a linear relationship

So far, we have studied how a Bayesian observer would learn a parameter of a probability distribution either over a binary outcome or over a real-valued variable. A slightly more complicated situation arises when the parameter that has to be learned is one that defines the relation between two variables. Yet, the Bayesian approach will not require much modification.

To exemplify this, we consider a toddler who is learning how to control her limbs. To first approximation, the command signal (e.g., firing rate of motor cortex neurons) sent to the spinal cord adjusts muscle force linearly. But what is the slope of this relationship between motor command and force output for a particular muscle? Without this knowledge, the toddler will be clumsy; as she acquires this knowledge, her motor control will improve. We consider the toddler's first attempts to learn this relationship. For simplicity, we assume that the toddler already understands that the relationship between motor command and force output is linear (though, clearly, this too must be learned), but she doesn't know the slope. Her goal is to estimate the slope, k, relating command signal s to force output F:

$$F(s) = ks. \tag{6.37}$$

Fig. 6.7 shows the results of 10 iterations in which the toddler uses different command signal magnitudes to push against the floor with her arms, while she judges the force she produces. She is able to judge the force based on feedback from her proprioceptors. These sensory signals are noisy, so we model her force measurement, f_t, on the t^{th} trial as a sample drawn from a Gaussian distribution around the actual force produced:

$$p(f_t|k,s_t) = \mathcal{N}(f_t;ks_t,\sigma^2). \tag{6.38}$$

The observations are then command-measurement pairs (s_t, f_t), where we assume s_t to be known exactly. Unlike in Section 6.3, we also assume that the toddler knows the measurement noise level σ; if she would not, then the inference described here would have to be combined with the inference in Section 6.3.

After each push, with her knowledge of the noise distribution, the toddler can construct a likelihood function reflecting the probability of the measurement given the slope. We assume a prior $p(k)$. She can then calculate her posterior over slope; it can be written as a variation of Eq.

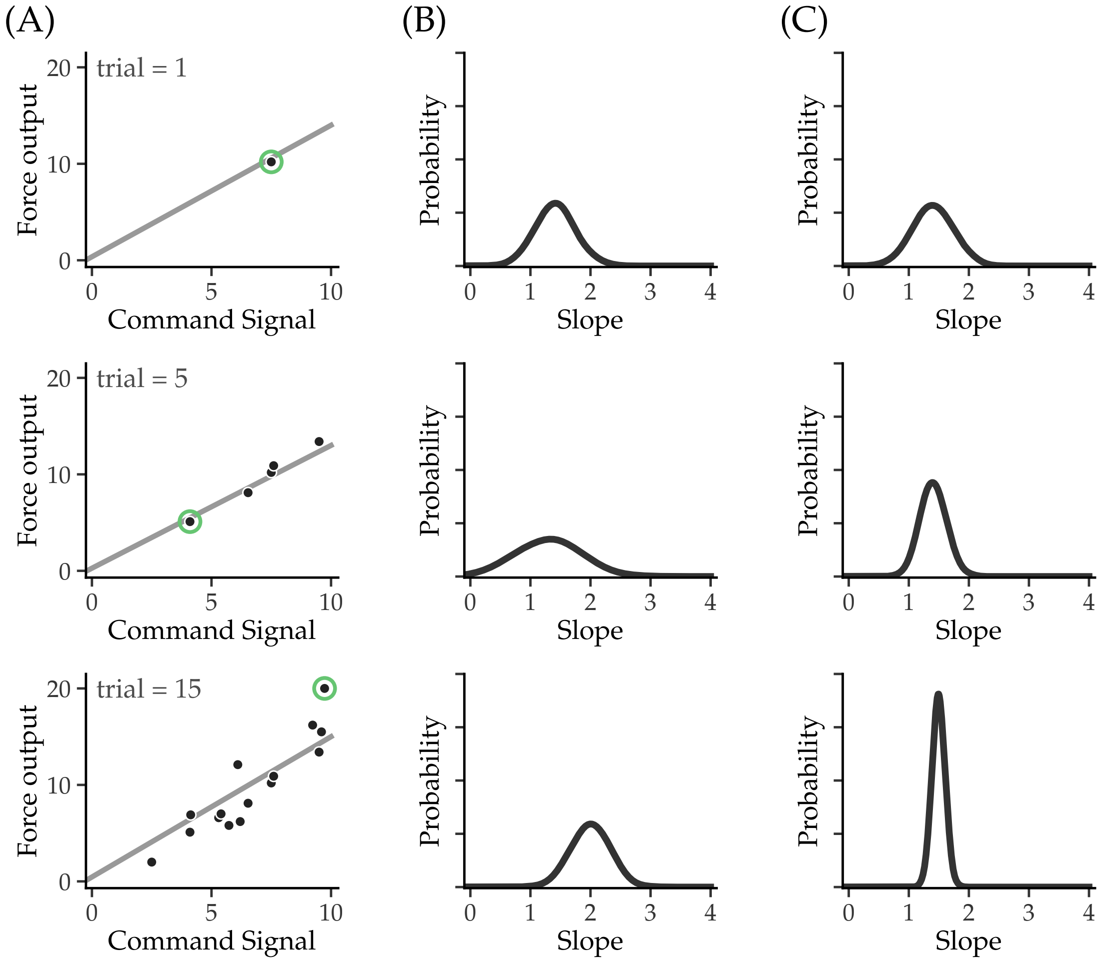


Figure 6.7: An observer learns the slope relating motor command signal magnitude (x-axis) to muscle force production (y-axis). **(A)** Scatterplots showing the data accumulating from trial 1 to 5 to 15. The line in each plot shows the posterior mean slope estimate based on all trials up to and including the one shown (data point circled in purple). **(B)** Single-measurement likelihood functions from the corresponding trials. **(C)** Posterior distribution over slope. The actual slope value used to generate the data was $k = 1.5$, with $\sigma = 2$. We assumed that the prior $p(k)$ is flat.

(5.24) and (6.8),

$$p(k|s_1, f_1, s_2, f_2, \ldots, s_T, f_T) = p(k) \prod_{t=1}^{T} p(f_t|k, s_t), \tag{6.39}$$

or as a variation of the update equations, Eq. (5.25) and (6.16):

$$p(k|s_1, f_1, s_2, f_2, \ldots, s_{T+1}, f_{T+1}) \propto p(k|s_1, f_1, s_2, f_2, \ldots, s_T, f_T) p(f_{T+1}|k, s_{T+1}). \tag{6.40}$$

The difference with the cases in previous sections is that now the commands $s_1, \ldots, s_T$ are needed in the likelihood, but since we assume these commands to be known to the toddler, we can simply condition every probability on them.

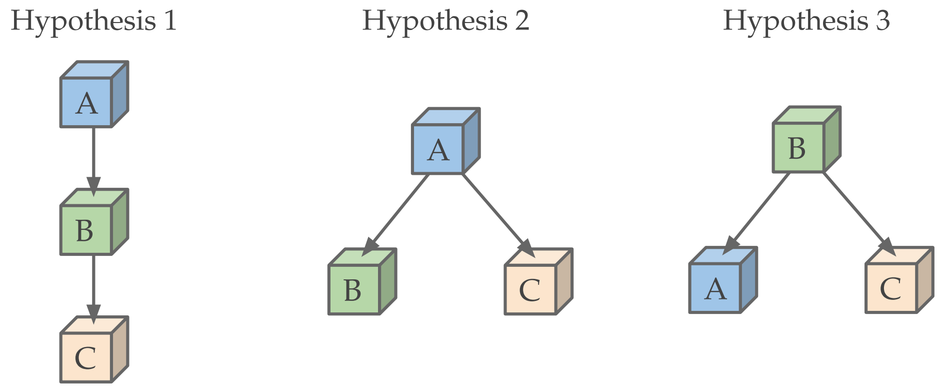


Figure 6.8: Three possible causal structures among three nodes. Although these diagrams look like generative models, they have a different meaning; each of them is a world state in a generative model.

6.5 Learning the structure of a causal model

An important form of learning is learning about the causal structure of the world: clouds may cause rain but not the other way round, this button causes the device to speed up, and smoking causes cancer. Causal learning lies at the basis of classical conditioning and associative learning. Much of science is about understanding the causal structures underlying complex systems – the climate, molecular pathways, neural circuits. In some cases, intervention is possible.

In our case study, we consider an observer who tries to determine the causal relationships among three nodes, A, B, and C (**Fig. 6.8**). For simplicity, we assume that only three structures are possible, and that the observer knows this.

Step 1: Generative model. The three structures, which we will denote by H_1, H_2, and H_3 have a priori the same probability:

$$p(H_1) = p(H_2) = p(H_3) = \frac{1}{3}. \tag{6.41}$$

We next assume that all nodes start out off. If a node in the structure gets turned on, then as a result, only the nodes directly connected by an arrow may also get turned on. However, there is a 0.2 probability of “failure", where despite the presence of a connecting arrow, the node at the end of the arrow fails to turn on. If multiple arrows emanate from the same node, then the causal effects along those arrows are independent. Nodes do not spontaneously turn on. Taking H_1 as an example, the causal rules of that structure include (but are not limited to) the following:

$$\begin{aligned}
&p(\text{A on} \rightarrow \text{B on}|H_1) = 0.8\\
&p(\text{B on} \rightarrow \text{A on}|H_1) = 0\\
&p(\text{C on} \rightarrow \text{B on}|H_1) = 0.8\\
&p(\text{B on} \rightarrow \text{C on}|H_1) = 0\\
&p(\text{A on} \rightarrow \text{C on}|H_1) = 0.8 \cdot 0.8 = 0.64
\end{aligned}$$

Step 2: Inference. Let’s consider a scenario in which an observer sees B being turned on, and as a result, C also gets turned on, but A does not. Thus, the observation is “B on → A off, C on". The likelihood of a hypothesis is the probabilities of this observation under the hypothesis. Using the causal structures, we find

$$\mathcal{L}(H_1) = p(\text{B on} \to \text{A off, C on}|H_1) = 1 \cdot 0.8 = 0.8$$
$$\mathcal{L}(H_2) = p(\text{B on} \to \text{A off, C on}|H_2) = 1 \cdot 0 = 0$$
$$\mathcal{L}(H_3) = p(\text{B on} \to \text{A off, C on}|H_3) = 0.2 \cdot 0.8 = 0.16$$

Combining the priors with the likelihoods, we can now calculate the posterior probabilities of the causal structures to be roughly 0.833, 0, and 0.167, respectively.

Exercise 6.7 Verify numerically that this is true. ■

We have learned something about the causal structure of the world from our observations. More observations will allow us to learn more and perhaps even determine the causal structure with high certainty. Obviously, this was a very simple example and real-world causal inference is more complicated in many ways. First, causal structures might have more than three nodes. Second, we assumed that the causal rules and the associated probabilities were known, which does not have to be the case. Third, we restricted the inference problem to three possible structures; often, there is a combinatorial explosion of the number of possible structures. On the bright side, in the real world, we might not depend on observations that are given to us, but we might be able to intervene with the system ourselves.

Box 6.4 — Non-Bayesian learning in artificial neural networks. A popular model of learning are artificial neural networks (ANNs). In their simplest form, multilayer perceptrons (MLPs) they accept an input x and calculate an output $\hat{y}$ which is meant to be similar to the true output y (say in the mean squared sense). The simplest function would be a linear function: $\hat{y} = \sum_i W_i x_i$. We may call this a single-layer neural network. We could then build a two-layer neural network as $\hat{y} = \sum_i W_i \sum_j W_{ij} x_j$. However, stacking two layers of linear transformations is just another linear transformation, written in a more complicated way. However, dependencies in the real world are usually nonlinear. Consequently, neural networks usually stack linear transformations and nonlinearities. For example, they may use the Rectified Linear Unit function: $\text{ReLU}(z) = \max(0, z)$ and say two layers (this class of functions can approximate most meaningful functions in human behavior). So they may use

$$\hat{y} = \sum_i W_{2i} \text{ReLU} \left(\sum_j W_{1ij} x_j \right)$$

In this nonlinear function artificial neural networks then tend to implement gradient descent with respect to the vector of all weights W on a loss function, e.g. squared error $C(\hat{y}, y) = (\hat{y} - y)^2$.

$$\Delta \mathbf{W_i} = -\eta \frac{\partial C}{\partial W_i}$$

where η is a small constant. This allows performance to progressively get better.

We want to point out that this has exactly the kind of update structure that we saw earlier in the chapter with respect to the Rescorla-Wagner model which can be seen as doing gradient descent on the square difference between the received reward and the expected reward. Learning in such artificial neural networks also shares aspects with the models of learning in this chapter. However, many aspects are quite distinct and go well beyond the scope of this book. However, we will discuss later how ANN learning can be used to obtain a likelihood in domains where that is hard (14.3) and how ANNs can be seen as a competing model to Bayesian ones to describe modeling of behavior[a] (15.1). ■

[a] Within this book we avoid the use of matrix notation, but in the domain of neural networks it is shockingly useful

6.6 Summary and remarks

In this chapter, we described learning from observations as a Bayesian computation. We have learned:

- Phrasing learning as a problem of statistical estimation allows a broad set of Bayesian approaches.
- We formulated a posterior distribution over the rate parameter of a Bernoulli (or binomial) distribution, and over the precision parameter of a Gaussian distribution.
- This computation involves the multiplication of likelihood functions, similar to evidence accumulation in Section 5.5. Conjugate priors can facilitate the integration of priors with new data.
- Sometimes, the causal structure of the world is not known and has to be learned from observations.
- In the examples we considered, the variable to be inferred does not change. This stands in contrast to forms of inference in which the variable of interest changes from trial to trial. We will treat this in Chapter 12.
- Providing observations to allow for learning is called training or teaching. Optimal teaching can also be treated computationally but we do not do so in this book.
- We also did not discuss *active learning*, in which the learner has some control over the acquisition of observations.

6.7 Suggested readings

- Daniel E Acuña and Paul Schrater. “Structure learning in human sequential decision-making”. In: *PLoS computational biology* 6.12 (2010), e1001003
- Patricia W Cheng. “From covariation to causation: A causal power theory.” In: *Psychological review* 104.2 (1997), page 367
- Anna Coenen, Bob Rehder, and Todd M Gureckis. “Strategies to intervene on causal systems are adaptively selected”. In: *Cognitive psychology* 79 (2015), pages 102–133
- Alison Gopnik et al. “A theory of causal learning in children: causal maps and Bayes nets.” In: *Psychological review* 111.1 (2004), page 3
- Charles Kemp, Andrew Perfors, and Joshua B Tenenbaum. “Learning overhypotheses with hierarchical Bayesian models”. In: *Developmental science* 10.3 (2007), pages 307–321
- Pierre Simon Laplace. “Memoir on the probability of the causes of events”. In: *Statistical science* 1.3 (1986), pages 364–378
- Tamas J Madarasz et al. “Evaluation of ambiguous associations in the amygdala by learning the structure of the environment”. In: *Nature neuroscience* 19.7 (2016), pages 965–972
- Jenny R Saffran, Richard N Aslin, and Elissa L Newport. “Statistical learning by 8-month-old infants”. In: *Science* 274.5294 (1996), pages 1926–1928
- Joshua B Tenenbaum. “Bayesian modeling of human concept learning”. In: *Advances in neural information processing systems* (1999), pages 59–68
- Ting Xiang, Terry Lohrenz, and P Read Montague. “Computational substrates of norms and their violations during social exchange”. In: *Journal of Neuroscience* 33.3 (2013), pages 1099–1108

- Fei Xu and Joshua B Tenenbaum. “Word learning as Bayesian inference.” In: *Psychological review* 114.2 (2007), page 245

6.8 Problems

Problem 6.1 In movement science, people build models of motor adaptation (which are quite similar to a model we used above). They talk about how muscles get stronger and weaker. And how the brain learns to deal with weaker muscles by sending stronger neural signals. Let us say we have muscles that (very slowly) fluctuate in strength (s). Whenever the person uses the muscle there will be a noisy observation (say with Gaussian noise) of the strength. Over time, the person can estimate how strong their muscles are. Formulate this problem as a learning problem in line with the models in this chapter.

Problem 6.2 Prospect theory, a popular (and Nobel-prize winning) theory in behavioral economics, posits that people overestimate small probabilities (e.g. p(I will die of Ebola) and underestimate large probabilities (e.g. p(I will die of cancer). Early in the chapter we wrote about thinking of probabilities of probabilities. Use that concept to describe why people should, indeed, overestimate small and underestimate large probabilities.

Problem 6.3 In Section 6.1, we noted that the probability of rain the next day was given by the posterior mean estimate, and was different from the most probable value of the rainfall rate, r, on the island, which is the MAP estimate.

a) Suppose you’ve been on the island for 4 days, and it has rained only once. Show that the MAP estimate for r is 0.25 and, according to Laplace’s rule of succession (Eq. (6.14), the posterior mean estimate is 0.33. Explain intuitively why it makes sense that your estimate for the probability of rain the next day is different from (and, in this case, greater than) your estimate of the most probable rainfall rate.

b) Verify Eq. (6.14) numerically, by discretizing r into several hundred or thousand values equally spaced between 0 and 1. Given 1 rainy day out of 4, calculate the likelihood for each r value, and enter these into Bayes’ formula with a uniform prior. Calculate the posterior mean as $\sum rp(r|\mathbf{x}_{\text{obs}})$.

c) If you are up to the challenge, use calculus to derive Eq. (6.14).

Problem 6.4 Generalize Eq. (6.22) to the case of a beta prior with parameters α_0 and β_0.

Problem 6.5 Evidence accumulation can be seen as a learning rule. This problem builds on Sections 5.5 and 6.2.

a) Show through a mathematical derivation that the posterior mean estimate can be written recursively as $\hat{s}_{t+1} = \hat{s} + \lambda_{t+1}(x_{t+1} - \hat{s}_t)$, and find an expression for the “learning rate” λ_t in terms of σ and t.

b) Does the learning rate increase or decrease as time goes by? Explain intuitively why this makes sense.

c) Generalize (a) and (b) to the situation where each measurement x_t has its own variance σ_t^2.

Problem 6.6 How can we learn a precision parameter?

a) Using the posterior in Eq. (6.33), find an expression for the posterior mean estimate.

b) Interpret this expression.

c) Find an expression for the posterior standard deviation.

d) Modify Eq. (6.33) if the learner has a prior $p(J_s)$ that is a gamma distribution with scale parameter k_0 and shape parameter θ_0.

e) Modify the answers to parts (a) and (c) accordingly.

Problem 6.7 Create a movie in which each frame corresponds to a trial in learning the slope parameter k in the relation $F = ks$ (see Section 6.4). The toddler performs 20 trials. On each trial, she sends a command s that is drawn from a uniform distribution on $[2, 10]$, and her force measurement is drawn from a Gaussian distribution with mean $1.5s$ and standard deviation 2.

The prior over k is flat. Each frame should look like a row in **Fig. 6.7**: (center) the likelihood function over k computed from the measurement on the t^{th} trial; (right) the posterior based on the measurements made up to and including the t^{th} trial; (left) a graphical representation of the data, with the line corresponding to the posterior mean estimate of k based on the measurements made up to and including the t^{th} trial. Make sure that the axes do not change from frame to frame. Choose the ranges on both axes large enough. Make sure that the numbers on the axes are easily legible and that the lines in your plots are sufficiently thick. Save your movie.

Problem 6.8 In the island example of Section 6.1, what is the distribution of the number of days elapsed between rain days? Explain.

Problem 6.9 In Section 2, we assumed that the binary outcome was independent across days. Assume instead that the probability of rain on a day after a rain day is r_r and after a dry day r_d. Derive the posterior probability that it will rain on day $T+1$ after a series of observations $x_1,\ldots,x_t$. Give an intuition of why.

Problem 6.10 This problem combines Problem 5.10 with the present chapter. Even though the present chapter focuses on learning, we also encountered a new distribution, the beta distribution, which is generally useful when inferring from counts of occurrence of two categories the underlying probabilities of those categories. We consider a pollster who is interested in the *proportion* of a population supporting a candidate. Thus, the variable s is now a probability (between 0 and 1). The individual's poll response x_i is binary: whether or not they support the candidate. The pollster's prior is still flat, and individual responses are still conditionally independent given s. Suppose that the pollster polls N people and receives n positive responses.

a) Show mathematically that the posterior distribution over s is a beta distribution with parameters $n+1$ and $N-n+1$. The maximum-likelihood (ML) estimate is $\hat{s}_{\text{ML}} = \frac{n}{N}$. posterior mean estimate is $\hat{s}_{\text{PM}} = \frac{n+1}{N+2}$.

b) A poll aggregator is trying to do "cue combination" of one poll with N_1 respondents and another with N_2 respondents. The respective numbers of positive responses in both polls are n_1 and n_2, and the respective ML estimates are $\hat{s}_{\text{ML},1}$ and $\hat{s}_{\text{ML},2}$. Show that the combined ML estimate of s is

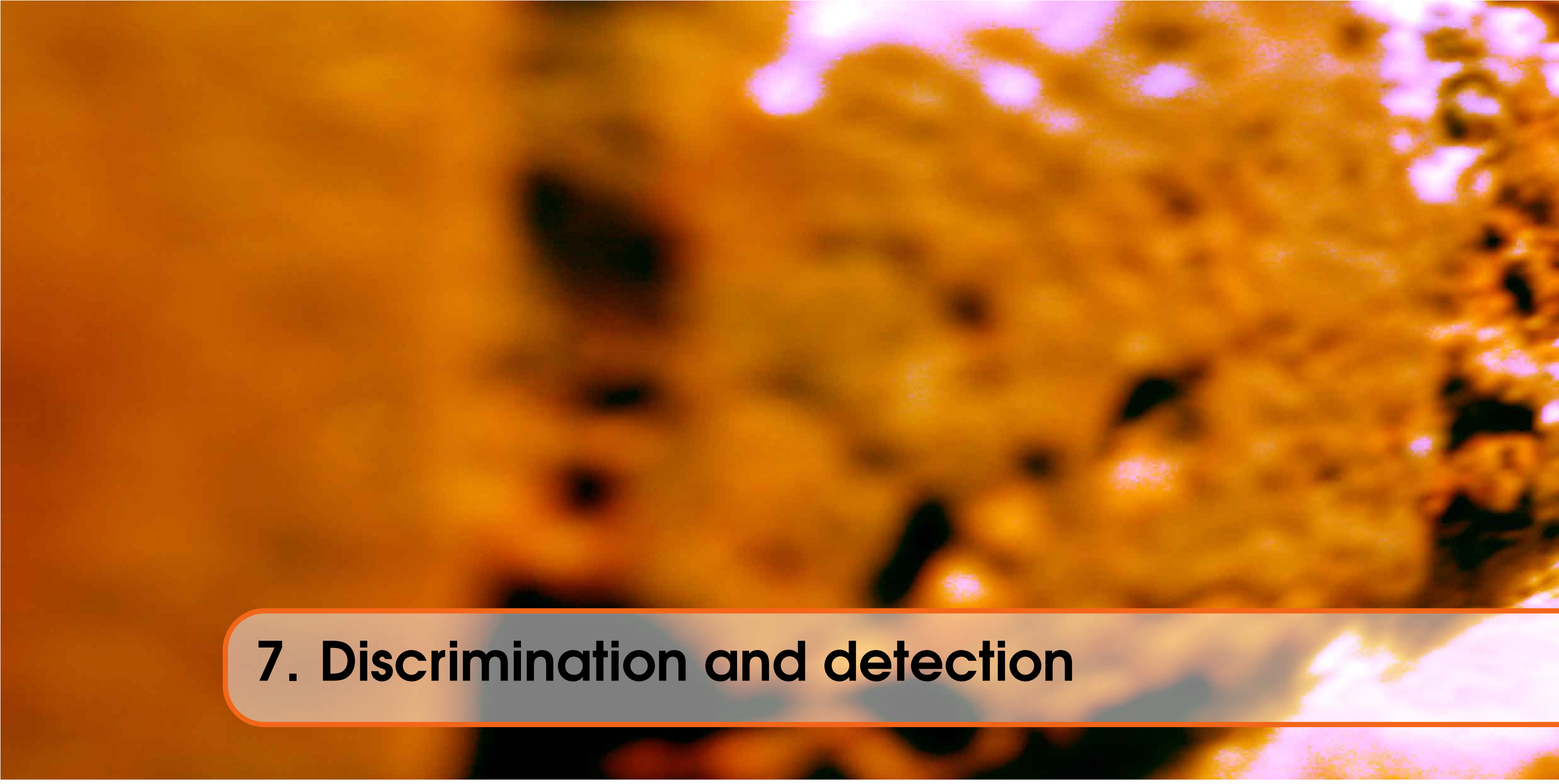
$$\hat{s}_{\text{ML, combined}} = \frac{N_1\hat{s}_{\text{ML},1}+N_2\hat{s}_{\text{ML},2}}{N_1+N_2}. \tag{6.42}$$

In other words, the combined ML estimate is a weighted average of the individual ML estimates. This parallels Eq. (5.11) for combining two normally distributed measurements.

c) Things are less intuitive with the posterior mean (PM) estimate. We denote the individual PM estimates by $\hat{s}_{\text{PM},1}$ and $\hat{s}_{\text{PM},2}$. Show that the combined PM estimate of s is

$$\hat{s}_{\text{PM, combined}} = \frac{(N_1+2)\hat{s}_{\text{PM},1}+(N_2+2)\hat{s}_{\text{PM},2}-1}{N_1+N_2+2}. \tag{6.43}$$

d) At first glance, this result seems bizarre: if we set $N_2=0$, the result will still depend on $\hat{s}_2$. Find out whether this is indeed the case.



7. Discrimination and detection

How do we determine which of two stimuli occurred?

In previous chapters, we discussed the basics of Bayesian modeling, often using the example of a spatial localization task. In such a task, the variable of interest – location on a line – is continuous, i.e. it takes on a continuum of values. The task was to estimate the location on this continuum – so that in principle, the subject has an infinite number of possible responses.

Many if not most tasks in the lab ask for a choice between only two alternatives. This is called a binary choice or binary decision. Binary choices are common in the real world, for example: will it rain today, can I trust this person, can I make it to the bus stop in time when I run, is this email spam or not? Each of these questions has a yes/no answer, and the corresponding random variable (whether it will rain today, etc.) is therefore binary. Many examples we encountered in Chapter 2, such as determining whether a bag on the baggage carousel is yours, also featured binary decisions.

Two types of binary decisions are particularly important, if only because they correspond to popular psychophysical paradigms. Imagine you are a radiologist trying to determine whether a tumor is present on a noisy X-ray. Such tasks, in which the observer decides whether a stimulus is present or absent, are *detection* tasks. Now imagine you are standing by the side of the road and see the silhouette of a moving car in the distance. You are trying to determine whether the car is coming towards or away from you. Tasks in which the observer decides between two nonzero values or categories of the stimulus variable (motion direction) are called *discrimination* tasks. Detection and discrimination tasks appear in many laboratory experiments: was the motion to the left or to the right (discrimination), was a vertical line present in the display (detection), did you feel a stimulus on your finger (detection), etc. Even when the underlying stimulus variable is continuous (e.g., duration), it is common to phrase the task in terms of a choice between two options (e.g., which stimulus lasted longer), while continuously manipulating the stimuli from trial to trial.

For binary decisions, the prior, the posterior, and the estimate distribution are each characterized by a single number, since the probabilities of both possible stimuli have to sum up to 1. This allows us to characterize behavior using a set of specialized tools, such as receiver operating characteristics, which allow a particularly meaningful characterization of the decision process.

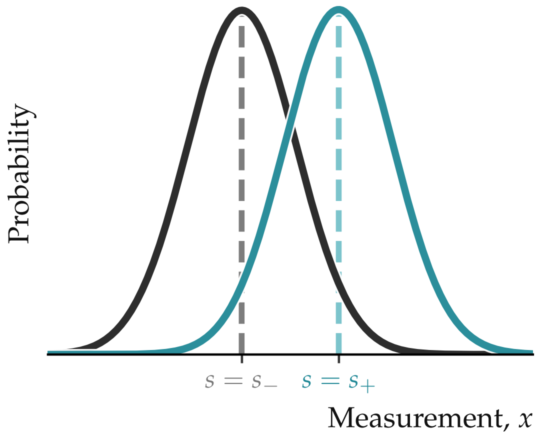


Figure 7.1: Noise distributions for the yes/no task of discriminating between s_+ and s_-.

Plan of the chapter

This chapter is structured around the same three-step process as previous chapters: generative model, inference process, and estimate distribution. We will use the same basic task as in Chapters 3-4, combining a measurement with a prior. This chapter will link Bayesian models to signal detection theory.

7.1 Discrimination

In this chapter, the stimulus s can take just two values, which we call s_+ and s_-, and the observer chooses between them. For example, the observer reports whether an oriented pattern is tilted 1° to the right or 1° to the left of vertical. This is called a "yes/no" discrimination task. (If the observer is presented with two stimuli, one of which is tilted 1° to the right and the other 1° to the left, the task would be a *two-alternative forced choice* discrimination task, also called a *two-interval forced choice* discrimination task. We will study such tasks in Chapter 11. In practice, people often use the term "two-alternative forced choice" also for the yes/no task.)

7.1.1 Step 1: Generative model

The generative model is $s \to x$, as in Chapter 3. The difference with that chapter is that the stimulus s takes only two possible values, also called *alternatives* and denoted by s_+ and s_-. These two are simply numbers, such as -1 and 1, or 0 and 3. They are *not* random variables. We will consider in Chapter 8 what happens when choosing between two *classes* of stimuli. The stimulus distribution is a discrete probability distribution, with values $P(s = s_+) = P(s = s_+)$ and $P(s = s_-) = P(s = s_-)$, which sum to 1 and reflect the frequencies with which the stimulus values are presented. The measurement x follows the usual Gaussian noise distribution $p(x|s)$. The noise distribution is shown in **Fig. 7.1** for both possible values of s.

7.1.2 Step 2: Inference

Suppose $s_+ = 1^\circ$ and $s_- = -1^\circ$, and that on a given trial, your measurement is 0.1°. Would you report that the stimulus was s_+ or s_-? You probably would say s_+ simply because the measurement is closer to s_+ than to s_-. But that is not what a Bayesian observer would necessarily do. To see that, imagine that you knew that s_- was far more common in the world (or in the experiment) than s_+ . In that case, a measurement that is only slightly closer to s_+ than to s_- would probably

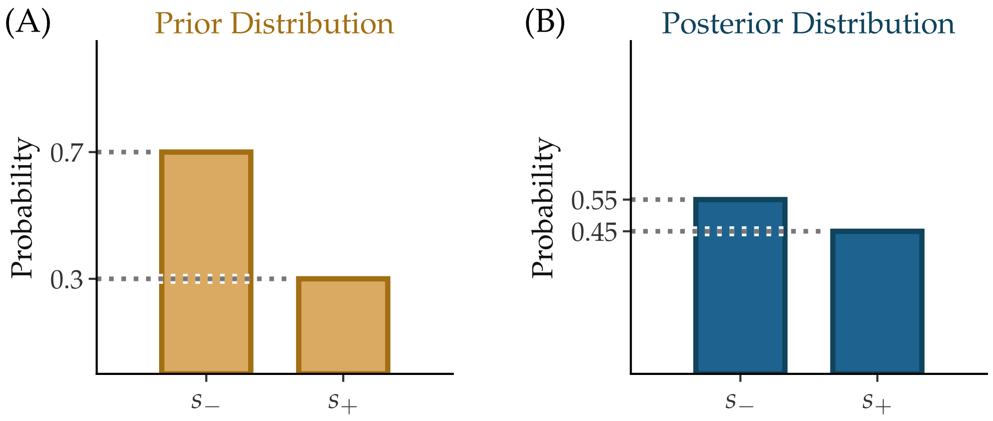


Figure 7.2: Example prior and posterior distribution over a binary variable.

have been produced by s_- . In this subsection, we will work out how the Bayesian observer would decide.

Exercise 7.1 Describe a real-world discrimination task where one stimulus, say s_+ is far more probable than another stimulus s_-. ■

Describing the Bayesian observer requires calculating the posterior over s, $p_{s|x}(s|x)$. Since s takes on two values, the posterior distribution is a discrete probability distribution, with values $P(s = s_+|x)$ and $P(s = s_-|x)$, which have to sum to 1 (**Fig. 7.2**)[1]. Bayes' rule tells us that

$$p(s|x) = \frac{p_{x|s}(x|s)P(s)}{p_x(x)}. \tag{7.1}$$

For a binary variable, the posterior distribution is uniquely determined by the ratio of the posterior probabilities of the two alternatives. We can calculate this ratio using Bayes' rule:

$$\frac{P(s = s_+|x)}{P(s = s_-|x)} = \frac{\frac{p(x|s=s_+)P(s=s_+)}{p_x(x)}}{\frac{p(x|s=s_-)P(s=s_-)}{p_x(x)}} \tag{7.2}$$

$$= \frac{p(x|s = s_+)P(s = s_+)}{p(x|s = s_-)P(s = s_-)} \tag{7.3}$$

This ratio is called the *posterior ratio* or *posterior odds*. Its interpretation is that of the probability of one alternative *relative to* that of the other. We see that the normalization $p_x(x)$ drops out and is thus irrelevant to this ratio. For example, if your posterior probability that the stimulus was s_+ is 80%, then your posterior ratio is $\frac{0.80}{0.20} = 4$. The posterior ratio is always positive but can grow arbitrarily large. For example, if the posterior probability of s_+ is 99%, then the posterior ratio is $\frac{0.99}{0.01} = 99$. Knowing the posterior ratio, one can calculate the probability of each of the alternatives, and vice versa.

In Bayesian calculations, you will often see Eq. (7.3) with the natural logarithm taken of both sides. We denote this log ratio by d:

$$d \equiv \log \frac{P(s = s_+|x)}{P(s = s_-|x)} = \log \frac{p(x|s = s_+)P(s = s_+)}{p(x|s = s_-)P(s = s_-)} \tag{7.4}$$

[1] In expressions such as these, many authors would leave out the subscripts; however, we leave them in to distinguish between a random variable and specific values of the variable. s_+ and s_- are simply numbers, which are plugged into the posterior distribution, likelihood function, and prior distribution, respectively. We only leave out the subscript when there is no possibility of misunderstanding, which is when the arguments are generic rather than specific values.

Taking the logarithm simplifies many mathematical derivations, as we will see below[2]. The quantity d is called the *log posterior ratio* (also: *log posterior odds*). If the posterior probability of s_+ is 0.80, then the log posterior ratio is $\log\frac{0.80}{0.20} = \log 4 = 1.39$. The log posterior ratio contains the same information as the posterior distribution itself; after all, we can exponentiate it and calculate the probability of each alternative from it as we did above.

The posterior probabilities of the two alternatives can be recovered from d as follows:

$$P(s=s_+|x) = \frac{1}{1+e^{-d}} \tag{7.5}$$

$$P(s=s_-|x) = \frac{1}{1+e^{d}} \tag{7.6}$$

The former is called the *logistic function* of d, the latter is 1 minus that.

Exercise 7.2 Show analytically that this relation of their difference being 1 is correct. ■

The log posterior ratio takes values between $-\infty$ and ∞. The log posterior ratio also has a symmetry property: flipping the posterior probabilities of the two alternatives is equivalent to flipping the sign of the log posterior ratio. For example, if the posterior probability of s_+ is 20% and that of s_- 80%, then the log posterior ratio is $\log\frac{0.20}{0.80} = \log 0.25 = -1.39$. When two alternatives have the same posterior probability, the log posterior ratio is equal to 0. If the log posterior ratio is positive, then $P(s=s_+|x)$ is greater than $P(s=s_-|x)$. Therefore, the MAP estimate is

$$\hat{s}_{\text{MAP}} = \begin{cases} s_+ & \text{if } d > 0 \\ s_- & \text{if } d < 0 \end{cases} \tag{7.7}$$

Thus, binary decision-making is concisely described in terms of log posterior ratios.

We now point out some common terminology. The inequality used to determine the MAP estimate, $d > 0$, is also called the *decision rule* of the Bayesian MAP observer, and d is called the *decision variable* (hence our notation d). The decision rule can be thought of as the binary equivalent of the mapping from x to s in estimation tasks (Chapters 3 and 4). The scalar value to which the decision variable is compared in order to make a decision, here 0, is also called the *decision criterion*, or simply the *criterion*. The terminology of decision rule, decision variable, and decision criterion is not specific to Bayesian models. Any inequality of the form $f(x) > k$, with f any function and k any scalar, can serve as a model for how the observer turns a measurement into a decision. In general, a *criterion* is any fixed value to which a variable is compared to produce a decision.

Log posterior ratios also simplify other aspects of the derivations. Since the logarithm of a product is the sum of the logarithms, the right-hand side of Eq. (7.4) can be rewritten as a sum:

$$d \equiv = \log\frac{P(s=s_+)}{P(s=s_-)} + \log\frac{p(x|s=s_+)}{p(x|s=s_-)} \tag{7.8}$$

Each of these terms has an intuitive meaning. The first term on the right-hand side is called the *log likelihood ratio* and reflects the amount of evidence provided by the measurement x. The second term on the right-hand side is the log prior ratio and reflects the observer's relative prior beliefs in the two alternatives. The sum of these terms is the log posterior ratio (**Fig. 7.3**).

Whenever d is greater than zero, s_+ is most probable. The MAP decision rule is thus to report s_+ when the sum of the log prior ratio and the log likelihood ratio is positive:

$$\text{MAP rule: Report } s_+ \text{ when } \log\frac{P(s=s_+)}{P(s=s_-)} + \log\frac{p(x|s=s_+)}{p(x|s=s_-)} > 0 \tag{7.9}$$

[2] When we use the notation "log" in this book, we always refer to the natural logarithm (base e).

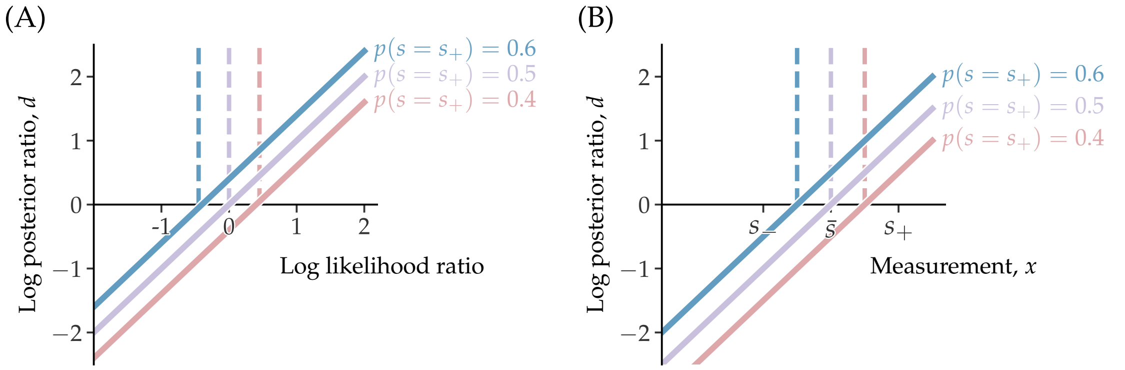


Figure 7.3: (A) Relation between log posterior ratio and log likelihood ratio (general, not just in the Gaussian model). Going from red to purple to blue the prior more strongly favors s_+, the criterion on the log likelihood ratio shifts towards smaller values (dashed lines), indicating that the subject has a stronger tendency to report s_+. **(B)** Relation between log posterior ratio and measurement *in the Gaussian model*. The relation is linear with slope $\frac{\Delta s}{\sigma^2}$. The purple line also represents the log *likelihood* ratio, and the corresponding "neutral" criterion is the midpoint of s_+ and s_-.

When s_- occurs with higher probability than s_+ (in other words, when the the log prior ratio is negative), the optimal decision rule is to report s_+ only when the measurement x provides sufficiently strong evidence in favor of s_+ to overcome the prior bias in favor of s_-.

To understand the intuition behind Eq. (7.9), we consider a world in which a person has either brown eyes or blue eyes. Imagine you are standing in front of a classroom of people, and you are asked about one person's eye color. The quality of information is affected by your distance to the person. Suppose that in your region of the world, brown eyes are more common than blue eyes. If you are asked about someone nearby, your sensory information will be of high quality and you will be able to base your decision predominantly on this sensory information. If you are asked about someone farther away, the quality of the sensory information is worse or even uninformative. The lower the quality of the visual information, the more you will rely upon your knowledge of the prevalence of brown eyes in the general population. When no information is available at all, your best bet is to always respond that the person has brown eyes. This increasing effect of the prior as the quality of sensory information decreases is exactly expressed by Eq. (7.9). The term on the left-hand side – the log-likelihood ratio – will tend to be smaller in magnitude (either positive or negative) when the sensory information is of lower quality ("tend to" because this term is a random variable that inherits its distribution from the distribution of x). The relative quality of prior and likelihood information determine which information dominates. In summary, in a yes/no discrimination task, the prior has the effect of *shifting the decision criterion.*

7.1.3 Gaussian model

If the measurement x follows a Gaussian distribution when conditioned on s, we can further evaluate the log likelihood ratio by substituting the expression for $p_{x|s}(x|s)$ This produces a particularly simple expression for the log posterior ratio,

$$d = \log \frac{P(s=s_+)}{P(s=s_-)} + \frac{\Delta s}{\sigma^2}(x-\bar{s}) \tag{7.10}$$

where we introduced the notation

$$\bar{s} \equiv \frac{s_+ + s_-}{2} \tag{7.11}$$

for the midpoint between s_+ and s_-, and

$$\Delta s = s_+ - s_- \tag{7.12}$$

for their difference. We will derive Eq. (7.10) in Problem 7.5. We plotted the log posterior ratio d as a function of the measurement x in **Fig. 7.3**. It is a linear function of the measurement.

To sum up, in a equal-variance Gaussian measurement model, the log posterior ratio is linear in the measurement. Although very convenient, this property is specific to the equal-variance Gaussian model, and rarely true otherwise (see Problems).

The log likelihood ratio is positive whenever x is greater than the midpoint of s_+ and s_- , and negative otherwise. This is intuitive: the measurement provides evidence for s_+ if it lies closer to s_+ than to s_- . The factor $\frac{\Delta s}{\sigma^2}$ helps determine the magnitude of the log likelihood ratio. This factor tells us that for the same x, the strength of the evidence is larger when the two stimuli to be discriminated are farther apart (i.e., $s_+ - s_-$ is larger) or when the noise in the measurement (σ) is smaller.

7.1.4 Decision rule in terms of the measurement

Substituting Eq. (7.10) for the log likelihood ratio into Eq. (7.9) for the decision rule and solving for x, we arrive at the optimal decision rule for our yes/no discrimination task with Gaussian measurement noise: to report that the stimulus is s_+ when

$$x > k_{\text{MAP}}, \tag{7.13}$$

where the *MAP criterion on the measurement* is

$$k_{\text{MAP}} = \bar{s} - \frac{\sigma^2}{\Delta s} \log \frac{P(s = s_+)}{P(s = s_-)}. \tag{7.14}$$

While the MAP criterion on the log posterior ratio is always 0, the MAP criterion in the measurement depends on the stimuli to be discriminated, the measurement noise, and the prior probabilities. The observer must in general have knowledge of all these variables in order to be optimal.

In the special case that the prior is flat, $P(s = s_+) = P(s = s_-) = 0.5$ (purple lines in **Fig. 7.3**), the log prior ratio is zero, and the observer would report s_+ simply when

$$x > \bar{s}. \tag{7.15}$$

This makes sense: x gets compared to the midpoint of s_- and s_+, which we could call the “neutral criterion". By comparing Eqs. (7.14) and (7.15), we can identify the term $-\frac{\sigma^2}{\Delta s} \log \frac{P(s=s_+)}{P(s=s_-)}$ as the *criterion shift* due to having a non-flat prior. If the prior favors s_+, for example $p_s(s+) = 0.6$ (blue line in **Fig. 7.3B**), then the negative log prior ratio would be a negative number (-0.405). As a consequence, the measurement x could be closer to s_- than to s_+ and yet the observer would respond $\hat{s} = s_+$. The opposite happens when the prior favors s_- (red line). In other words, the prior *biases* the observer towards reporting the alternative with the highest prior probability. This phenomenon is similar to how a Gaussian prior biases the observer towards its mean in the continuous estimation task discussed in Chapters 3 and 4.

7.1.5 Multiple tasks can have the same Bayesian decision rule

Each binary decision has only one Bayesian decision rule, aside from the fact that the same rule can be written in mathematically equivalent forms, for example $x > 0$ would be equivalent to $e^x > 1$. However, different tasks can have the same decision rule. As a simple example, consider the decision rule in Eq. (7.15). There are many ways of choosing pairs of stimuli (s_+, s_-) that have the same mean and therefore the same decision rule. Therefore, it is not possible to reconstruct the task from the decision rule.

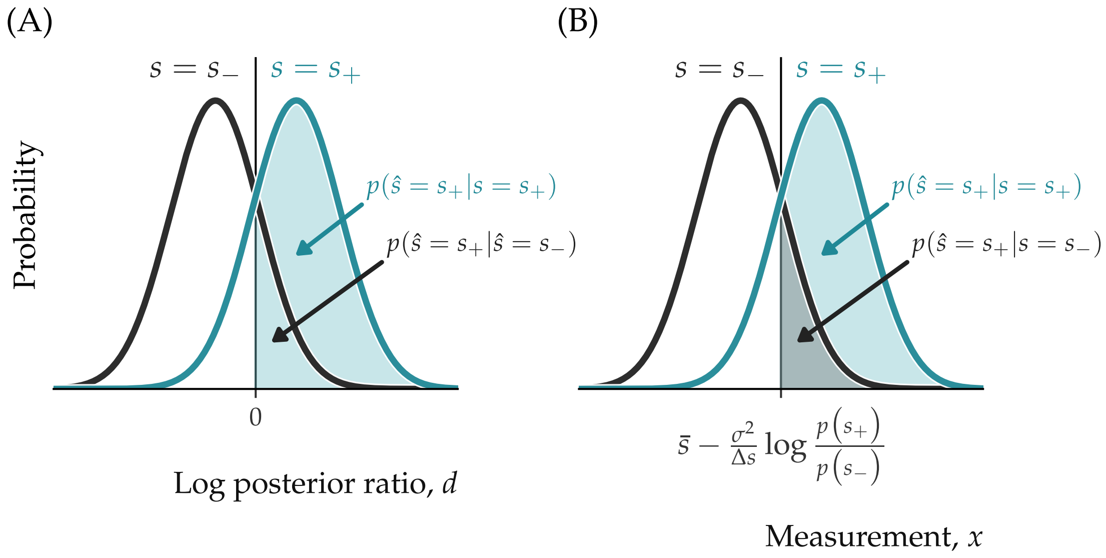


Figure 7.4: Response probabilities in discrimination. The following are two equivalent ways of visualizing Step 3. **(A).** Log posterior ratio space (d-space). Distributions of the decision variable (the log posterior ratio) conditioned on the true world state, s_+ or s_-. The probability that the MAP estimate is s_+ is equal to the shaded area when the true stimulus is s_- (grey) or s_+ (teal). **(B)** Measurement space (x-space). The criterion on the measurement is different but the shaded areas have the same meaning as in **(A)**.

7.1.6 Step 3: Response distribution

In this section, we generalize a bit beyond the Bayesian model and consider a general criterion k, which may or may not be equal to the Bayesian criterion k_{MAP}. Regardless, we are interested in the distribution of the response over many trials in which the experimental condition is held fixed. In our task, the experimental condition is completely specified by s, which can take two values. Therefore, the distribution of the stimulus estimate is given by the probability of reporting either $\hat{s} = s_+$ or $\hat{s} = s_-$, when x is drawn from either $p(x|s = s_+)$ or $p(x|s = s_-)$, for a total of four possibilities. These four numbers can be reduced to two, since the probability of estimating the stimulus as s_+ is 1 minus that of estimating it as s_-. Thus, the distribution of the estimate is determined by the following two probabilities, corresponding to the observer's correct and incorrect reports of s_+:

$$P(\hat{s} = s_+|s = s_+) = P(d > 0|s = s_+) \tag{7.16}$$
$$P(\hat{s} = s_+|s = s_-) = P(d > 0|s = s_-) \tag{7.17}$$

For convenience, we assume in this subsection that the prior is flat. Evaluating Eq. (7.17) allows us to calculate predictions for how we expect an observer to behave across multiple trials.

As the third step of Bayesian modeling, we need to compute the probability that Eq. (7.15) is satisfied when x is drawn from either $p(x|s = s_+)$ or $p(x|s = s_-)$. We can think of this in two equivalent spaces: d-space (**Fig. 7.4A**) and x-space (**Fig. 7.4B**). Here, we focus on the latter (in a problem, on the former). In **Fig. 7.4B**, we copied the plot of both distributions from **Fig. 7.1**. Eq. (7.15) is satisfied whenever the measurement falls to the right of the vertical line at k. (This may or may not be the MAP criterion, k_{MAP}). Thus, graphically, the probability that Eq. (7.15) is satisfied is the area under the probability density function to the right of the line. Mathematically, calculating this area corresponds to integrating the density function from k to infinity:

$$\Pr(\hat{s}=s_+|s=s_+) = \int_k^\infty \mathcal{N}(x;s_+,\sigma^2)dx = \frac{1}{\sqrt{2\pi\sigma^2}}\int_k^\infty e^{-\frac{(x-s_+)^2}{2\sigma^2}}dx \tag{7.18}$$

$$\Pr(\hat{s}=s_+|s=s_-) = \int_k^\infty \mathcal{N}(x;s_-,\sigma^2)dx = \frac{1}{\sqrt{2\pi\sigma^2}}\int_k^\infty e^{-\frac{(x-s_+)^2}{2\sigma^2}}dx. \tag{7.19}$$

These integrals are cumulative distribution functions of the normal distribution. They cannot be evaluated analytically, but they can be rewritten in terms of either the *cumulative distribution* of a normal distribution, Φ, or the *cumulative distribution* of a *standard* normal distribution, Φ_{standard}[3].

$$\Pr(\hat{s}=s_+|s=s_+) = \Phi\left(s_+;k,\sigma^2\right) = \Phi_{\text{standard}}\left(\frac{s_+-k}{\sigma}\right) \tag{7.20}$$

$$\Pr(\hat{s}=s_+|s=s_-) = \Phi\left(s_-;k,\sigma^2\right) = \Phi_{\text{standard}}\left(\frac{s_--k}{\sigma}\right). \tag{7.21}$$

How we obtain these equations is explained in more detail in Box 7.1. Eqs. (7.20)-(7.21) represent predictions for how often subjects would estimate the stimulus s_+ when it is in reality s_+, or when it is in reality s_-. These predictions can be compared against experimental results, in parallel to what we found in Section 4.2 for the case of continuous estimation.

In the special case that $k=k_{\text{MAP}}$, we can substitute Eq. (7.14) into Eqs. (7.20)-(7.21) to find

$$\Pr(\hat{s}=s_+|s=s_+) = \Phi_{\text{standard}}\left(\frac{\Delta s}{2\sigma}+\frac{\sigma}{\Delta s}\log\frac{P(s=s_+)}{P(s=s_-)}\right) \tag{7.22}$$

$$\Pr(\hat{s}=s_+|s=s_-) = \Phi_{\text{standard}}\left(-\frac{\Delta s}{2\sigma}+\frac{\sigma}{\Delta s}\log\frac{P(s=s_+)}{P(s=s_-)}\right). \tag{7.23}$$

Box 7.1 — Cumulative normal distribution. A cumulative distribution is obtained from a regular probability distribution by summing the values up from left to right, keeping a running tally. For example, the cumulative normal distribution belonging to a normal distribution with mean μ and variance σ^2 is defined as

$$\Phi\left(y;\mu,\sigma^2\right) \equiv \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{y} e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx. \tag{7.24}$$

Here, y is the argument of the cumulative distribution, whereas x is just an integration variable. The *standard normal distribution* is a normal distribution with mean 0 and variance 1. We give its cumulative distribution a special notation,

$$\Phi_{\text{standard}}(y) \equiv \Phi(y;0,1). \tag{7.25}$$

The following properties hold:

$$\Phi(y;\mu,\sigma^2) = \Phi(y-\mu;0,\sigma^2) \tag{7.26}$$

$$\Phi(y;\mu,\sigma^2) = \Phi_{\text{standard}}\left(\frac{y-\mu}{\sigma}\right) \tag{7.27}$$

$$\Phi(y;\mu,\sigma^2)+\Phi(-y;-\mu,\sigma^2) = 1 \tag{7.28}$$

$$\Phi_{\text{standard}}(y)+\Phi_{\text{standard}}(-y) = 1 \tag{7.29}$$

[3] Some texts use Φ for the cumulative standard normal distribution.

Exercise 7.3 a) Derive these properties.

b) Show how we can use some of these properties to derive Eq. (7.20) from Eq. (7.18) (or Eq. (7.21) from Eq. (7.19)).

7.2 Detection

In the introduction of this chapter, we provided the example of a radiologist determining whether a patient has a tumor based on an X-ray. This is an example of a *detection task*: is the tumor present or not? There are many other daily-life examples. As we are showering, we have to determine whether or not our phone rang. On the road, we have to determine whether there is a bump ahead or not. If we have a gas stove, detecting the smell of gas can keep us out of danger. In general, the task is to determine whether a signal is present in noise.

Mathematically, detection is closely related to the discrimination task discussed so far. In the discrimination task, the observer had to discriminate between two stimulus values, s_+ and s_-. In its simplest form, detection is the special case in which s_+ is positive and $s_- = 0$, i.e., the observer is discriminating between a certain nonzero value and zero. In many cases, the variable s is somewhat abstract; for instance, s could be a composite of different image features that a radiologist uses to judge tumor presence. For simplicity, however, we still conceptualize s as a one-dimensional variable.

The Bayesian model we described for discrimination can therefore also be used for detection. For example, the decision variable is obtained by substituting $s_- = 0$ into Eq. (7.10) to obtain

$$d = \log \frac{P(s = s_+)}{P(s = 0)} + \frac{s_+}{\sigma^2}\left(x - \frac{s_+}{2}\right). \tag{7.30}$$

Similarly, from Eq. (7.20), the probability of correctly detecting the stimulus is

$$P(\hat{s} = s_+ | s = s_+) = \Phi_{\text{standard}}\left(\frac{s_+}{2\sigma} + \frac{\sigma}{s_+}\log\frac{P(s = s_+)}{P(s = 0)}.\right) \tag{7.31}$$

In a detection task, the probability of reporting "present" when the signal is present is called the *hit rate*, *detection rate*, *sensitivity*, or *true positive rate*, whereas the probability of reporting "present" when the signal is absent is called the *false-alarm rate* or *false-positive rate*. These probabilities can all be recognized as areas under the curves in **Fig. 7.4**. This terminology stems from *signal detection theory*. 1 minus the hit rate is the *miss rate* or *false-negative rate*, and 1 minus the false-alarm rate is called the *correct-rejection rate*, *specificity*, or *true-negative rate* (see **Fig. 7.5**). These four terms can also be applied to a discrimination task, such as discriminating a $-1°$ from a $1°$ orientation, but in such tasks, it is arbitrary which stimulus is regarded as the "signal".

In our example, hit and correct-rejection rates are equal, as are the false-alarm and miss rates. Consequently, hit and false-alarm rates sum to 1. In general, however, hit and false-alarm rate do not need to sum to 1.

7.3 Confidence

In a binary decision, the sign of the log posterior ratio determines the MAP decision. However, the log posterior ratio also has a magnitude or absolute value. A decision made with a log posterior ratio of 0.1 is made less confidently than one with a log posterior ratio of 1: after all, a lower absolute value means that the posterior probabilities of the two alternatives are closer to each other.

		Reported world state		
		present	absent	Total
True world state	present	hits (true positives)	misses (false negatives)	1
	absent	false alarms (false positives)	correct rejections (true negatives)	1

Figure 7.5: Terminology for the four types of response frequencies in a detection task.

Therefore, a natural measure of confidence in a binary decision is the magnitude of the log posterior ratio:

$$\text{confidence2} = \left| \log \frac{P(s = s_+|x)}{P(s = s_-|x)} \right|. \tag{7.32}$$

This is plotted in **Fig. 7.6**. Confidence can decrease due to a non-flat prior. For example, when the log likelihood ratio is 0.3, and the log prior ratio is -0.4, then confidence decreases from 0.3 to 0.1 due to the introduction of the non-flat prior.

In Chapter 3, we introduced a different measure of confidence, namely the posterior probability of the response (estimate). We will here call that measure confidence1:

$$\text{confidence1} = P(s = \hat{s}|x). \tag{7.33}$$

In the current task, the estimate is the MAP estimate, so confidence1 is equal to $P(s = s_+|x)$ if $P(s = s_+|x) > 0.5$, and equal to $1 - P(s = s_+|x)$ if $P(s = s_+|x) < 0.5$. The two measures, confidence1 and confidence2, are related through:

$$\text{confidence1} = \frac{1}{1 + e^{-\text{confidence2}}}. \tag{7.34}$$

We will derive this in a problem.

In Eq. (7.34), the logistic function, which we first encountered in Eq. (7.5), makes an appearance again – and for much the same reason. Since the logistic function – is monotonically increasing, the two measures of confidence are in one-to-one correspondence, and both are legitimate measures of confidence. In **Fig. 7.6**, both measures are plotted as a function of the log posterior ratio.

Exercise 7.4 Does either of the two measures make more sense to you? Why? ■

Having established that confidence (either measure) corresponds to distance from the origin on the decision variable axis, we can use the Bayesian model to predict not only the observer's responses on the discrimination (or detection) task, but also how often this decision is made with high or with low confidence. Of course, the dividing line between low and high confidence is unknown, but this is a number that could be fitted to a human subject's data. This is the idea behind a *confidence rating* experiment: the subject is asked not only for a binary judgment on the discrimination task, but afterwards also to rate confidence, let's say as low, medium, or high. Thus,

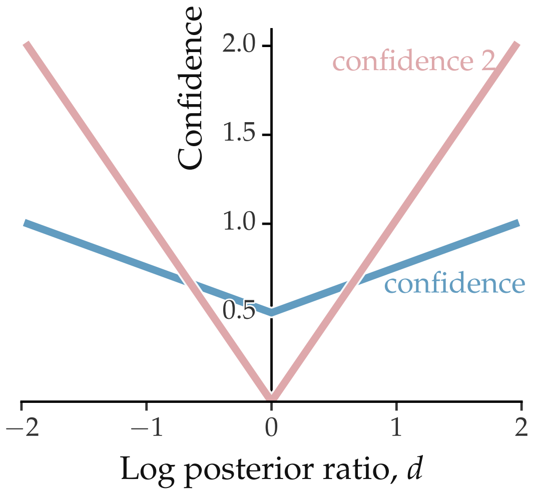


Figure 7.6: Two measures of confidence as a function of the log posterior ratio. Red: the absolute value of the log posterior ratio. Blue: the posterior probability of the MAP choice.

there are now 6 possible responses: 2 estimates times 3 confidence ratings. We saw before that the Bayesian observer makes a binary judgment by determining in which of two regions in the decision space (positive and negative) the log posterior ratio falls. Similarly, the Bayesian observer now chooses one of the six possible responses by determining in which of six decision regions the log posterior ratio falls (**Fig. 7.7**). Three of these regions together form the negative axis, and three the positive axis. From left to right, these regions would correspond to estimating the stimulus as s_- with high, medium, and low confidence, and estimating the stimulus as s_+ with low, medium, and high confidence. A total of five decision criteria separate these regions. When there are M confidence ratings, the number of points in the plot is $2M-1$.

7.4 Further characterizing the response distribution

In the binary tasks that are the topic of this chapter, we can further characterize the response distribution. The following applies both to discrimination and to detection, although the terminology (hit rate, false-alarm rate, etc.) is mostly associated with detection.

7.4.1 Receiver operating characteristic

In Section 7.2, we defined hit and false-alarm rates with respect to one particular decision criterion. In a task with confidence ratings, we can associate a hit and a false-alarm rate with any criterion dividing two adjoining decision regions. In the example with three criteria, the highest criterion would separate s_+ estimates made with medium confidence from those made with high confidence. The generalized hit rate of the Bayesian model is equal to the area under the distribution of the decision variable when $s = s_+$, $p_{d|s}(d|s_+)$, to the right of a particular criterion k (which was previously always zero). Similarly, the generalized false-alarm rate is equal to the area under the $s = s_-$ distribution of the decision variable, $p_{d|s}(d|s_-)$, to the right of the same criterion k. In equations:

$$H(k) = P(d > k|s = s_+) = P(d > k|s = s_+) \tag{7.35}$$

$$F(k) = P(d > k|s = s_-) = P(d > k|s = s_-) \tag{7.36}$$

If there are three confidence ratings, this leads to five pairs of hit and false-alarm rates, one for each criterion. Plotting hit rate $H(k)$ against false-alarm rate $F(k)$ gives us five points in a plot

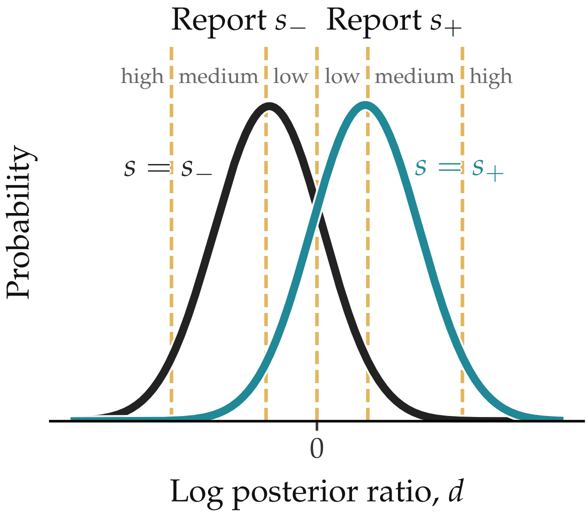


Figure 7.7: Confidence ratings (here low, medium, high) subdivide the regions $d < 0$ (report s_-) and $d > 0$ (report s_+). In this hypothetical experiment, the observer has a total of 6 response categories, ordered as indicated. When d is high in absolute value, the confidence rating will be higher.

with horizontal and vertical axes both equal to [0,1]. In the limit of having a very large number of confidence ratings, the plot would contain a smooth curve passing through the origin and through (1,1). This would correspond to the decision criterion k moving continuously along the decision axis from right to left, at each value producing a hit and a false-alarm rate (**Fig. 7.8**). This curve is called the *receiver operating characteristic* (ROC). It characterizes the distributions of the decision variable given either stimulus in a more complete manner than the original hit and false-alarm rates can; the latter are essentially only one point on the ROC. The ROC is *parameterized by* the criterion. The ROC is one of the most important concepts in signal detection theory.

In the main case under study in this chapter, the hit rate is equal to the correct rejection rate and the false-alarm rate is equal to the miss rate. As a consequence, the ROC is symmetrical around the negative diagonal.

Exercise 7.5 Why is this the case? ■

However, this is not the case in general, and in a problem, we will see an example of an ROC that is asymmetric around the negative diagonal.

In an actual experiment, an empirical ROC is obtained from the response frequencies in each of the $2M$ response categories, for each of the two stimuli. The way to do this is by creating a table of 2 rows and $2M$ columns (**Fig. 7.9**, rows I and II). The top row corresponds to the true stimulus being $s = s_+$, the bottom row to $s = s_-$. Each column corresponds to a response category. The left M columns correspond to $s = s_-$ responses, in order of decreasing confidence. The right M columns correspond to $s = s_+$ responses, in order of increasing confidence. Each cell in the table contains the frequency of responses in each category, divided by the total number of responses across all categories for that stimulus. Thus, the sum of the numbers in each row equals 1. Next, create a new table in which each cell contains the sum of the number in the corresponding cell and all cells to the right of it in the same row in the original table. In other words, the new table (rows II and IV in **Fig. 6.8a**) is built by cumulatively summing the numbers in the original table from right to left, for each row separately. In the new table, each column corresponds to a (hit, false-alarm) rate pair. The leftmost pair should, by construction, always be equal to (1,1). Finally, the hit rate (row II)

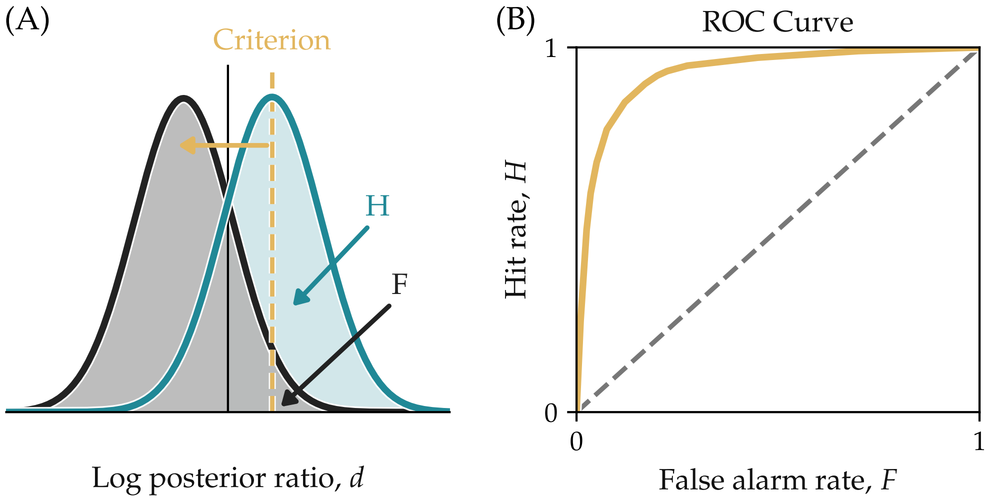


Figure 7.8: Theoretical receiver-operating characteristic. (A) Distribution of the log posterior ratio d when the stimulus is s_+ (teal curve) or s_- (grey curve). The criterion (dashed gold line) defines a hit rate H (teal area), and a false-alarm rate F (grey area). **(B)** By sweeping the criterion from right to left and plotting H against F, we obtain the theoretical receiver-operating characteristic.

is plotted against the false-alarm rate (row IV) (**Fig. 7.9**). When a model accurately describes an observer, the ROC obtained from that model should go through the points of the empirical ROC.

Note that it does not matter where the observer places their confidence criteria in d-space – the ROC will be the same.

7.4.2 Sensitivity

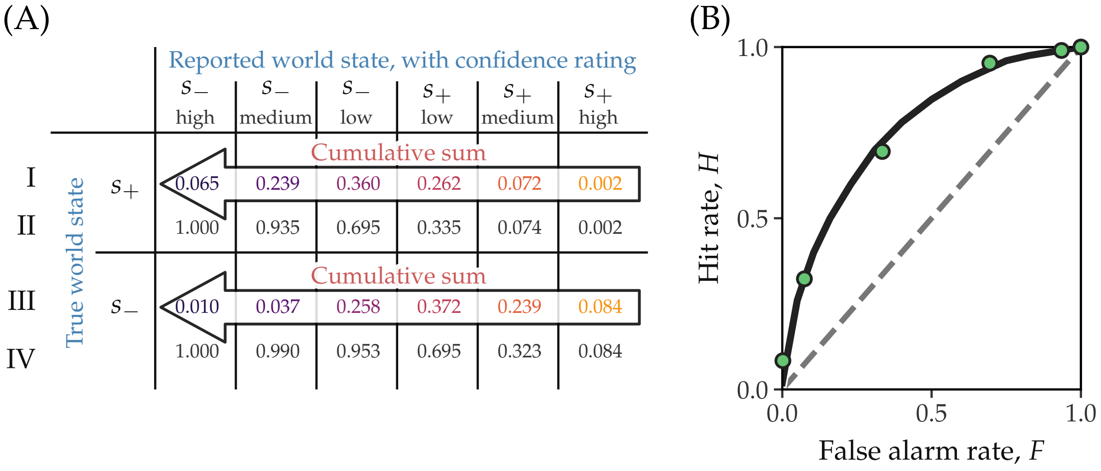
In previous sections, we used the decision rule $d > 0$, where d is the log posterior ratio. There are, however, scenarios in which the observer would use a criterion different from 0. One possibility is that the observer is suboptimal and makes a wrong assumption about the prior probabilities. In that case, the criterion will be replaced by an unknown number. A second scenario is that the observer may (rationally or not) attach greater importance to estimating one of the stimuli correctly than to estimating the other correctly (we will elaborate on this situation in a later chapter).

In these scenarios, the decision rule takes the form $d > k$. The hit and false-alarm rates with respect to this unknown criterion (which is sometimes also called *bias*), Eqs. (7.36), become

$$H(k) = \Phi_{\text{standard}}\left(\frac{s_+ - k}{\sigma}\right) \tag{7.37}$$

$$F(k) = \Phi_{\text{standard}}\left(\frac{s_- - k}{\sigma}\right). \tag{7.38}$$

Adopting a different k changes various measures of the observer's performance. For instance, the observer's proportion correct, which is $P(s = s_+)H + P(s = s_-)F$, now depends on k. However, the adoption of a different k does not change the observer's ROC. Rather, the effect of a change in k is simply to move the observer to a different point on the same ROC. The fact that the observer's ROC is unaffected by the criterion suggests that it should be possible to derive a numerical measure of performance that does not depend on k. One such criterion-invariant measure of performance is the *area under the ROC*, but the most common criterion-invariant measure is sensitivity. Sensitivity, also called discriminability and denoted d' (read "d prime") is a way to quantify how well-separated



		s_- high	s_- medium	s_- low	s_+ low	s_+ medium	s_+ high
I	s_+	0.065	0.239	0.360	0.262	0.072	0.002
II		1.000	0.935	0.695	0.335	0.074	0.002
III	s_-	0.010	0.037	0.258	0.372	0.239	0.084
IV		1.000	0.990	0.953	0.695	0.323	0.084

Figure 7.9: Empirical receiver-operating characteristic. **(A)** Rows I and III show the response proportions in a hypothetical experiment when the world state was s_+ (I) or s_- (III). The subject reported s_+ or s_- with low, medium, or high confidence rating (columns). We first take the cumulative sum of proportions from right to left; this produces rows II and IV. Then, we plot row II (hit rate) against row IV (false-alarm rate). This produces the green points in **(B)**. The black curve represents the theoretical ROC underlying these data.

the distributions of the decision variable under the two alternatives are. This measure is defined as

$$d' = \Phi^{-1}_{\text{standard}}(H) - \Phi^{-1}_{\text{standard}}(F). \tag{7.39}$$

where $\phi^{-1}_{\text{standard}}$ refers to the inverse function of ϕ_{standard}. This means that $\phi^{-1}_{\text{standard}}(y)$ is the value x for which $\phi_{\text{standard}}(x) = y$. You can think of this as a "reverse lookup". In some texts, z (z-score) is used to denote $\phi^{-1}_{\text{standard}}$. The sense of Eq. becomes clear when we choose $k = k_{\text{MAP}}$. Then, we find

$$d' = \frac{\Delta s}{\sigma}. \tag{7.40}$$

This remarkably simple expression does not depend on the criterion (or on the prior)! No matter how much or how little the observer may be biased, sensitivity only reflects the distributions of the decision variable conditioned on s, i.e., the sensory evidence. The more these distributions overlap, the lower d' is. As the ratio between the difference between the two stimuli to be discriminated and the level of sensory noise, d' can be interpreted as the observer's signal-to-noise ratio for the task.

Note that some text use Eq. (7.40) instead of Eq. (7.39) as the definition of d'. However, this would be much less general.

The introduction of performance measures that are invariant under changes in criterion was one of the main accomplishments of signal detection theory. However, an important caveat is that if the noise model is not equal-variance Gaussian, then Eq. (7.40) ceases to hold and d', as defined in Eq. (7.39), does become criterion-dependent.

Box 7.2 — Discriminability or accuracy?. There might seem to be a tension between discriminability, d', and accuracy. In signal detection theory, it is common to regard discriminability as a better measure of performance than accuracy, because it is independent of the criterion. By contrast, accuracy is maximized by the Bayesian MAP observer, so it would make sense to use accuracy as a measure of performance. This apparent tension is resolved by noting that the Bayesian MAP observer does not use just any criterion but rather the optimal one (the one that maximizes the posterior). Thus, accuracy is an entirely valid measure of performance. However, it might still be useful to split up accuracy into hit rate H and 1 minus false-alarm rate F. Discriminability is equivalent to accuracy (i.e., perfectly correlated with accuracy) when the distributions of the decision variable conditioned on s are Gaussian distributions with the same variance. In other cases, discriminability is of limited use. ■

7.5 The relation between Bayesian inference and signal detection theory

Signal detection theory has been widely applied in many domains, ranging from detecting objects on radar (for which the theory was originally developed), to characterizing the performance of clinical diagnostic tests, to studying recall of words from memory. These applications have been described in detail elsewhere. Each of these situations involves the observer using noisy information (a radar image, physiological measurements taken from a patient, or a sense of familiarity) to classify a stimulus into one of two categories (presence or absence of an object, presence or absence of the disease, having seen the word before or not).

As we've seen above, signal detection theory is in essence an application of Bayesian inference. Nevertheless, many users of signal detection theory are unaware of its direct connection to Bayesian inference. Indeed, signal detection theory studies typically are not described in the same texts as more recent studies of Bayesian perception, such as cue combination as described in the previous chapter. This might seem surprising, since both are based on computing the posterior distribution followed by MAP estimation. The reason for the separation is that historically, signal detection theory has mostly concerned itself with binary discrimination or detection tasks with a flat prior. As we have seen, the Bayesian MAP decision rule is then simply $x > (s_+ + s_-)/2$. Unlike the MAP estimate for cue combination, this decision rule does not require any knowledge of stimulus uncertainty, i.e. the width of the likelihood function, σ. (If the prior is not flat, then the decision rule does require such knowledge.) By contrast, more recent studies of Bayesian perception, like those discussed in Chapters 3-4, have focused on tasks that do require knowledge of stimulus uncertainty. There, we emphasized that the MAP estimate was given by an expression in which the measurement x is weighted by its reliability or precision, $\frac{1}{\sigma^2}$. In the current chapter, we have not emphasized this, but the same feature is present in Eq. (7.10), the log likelihood ratio in the discrimination task. In a visual orientation discrimination task, the precision could be manipulated for example through presentation time or stimulus contrast. We saw that when the prior is flat, the decision rule simplifies to an equation from which σ^2 disappears, $x > \frac{s_+ + s_-}{2}$. However, in the presence of a non-trivial prior, the weighting by reliability survives. Studies in which observers are optimal even when optimality requires knowledge of sensory uncertainty provide clues that neurons encode entire likelihood functions over stimuli, rather than only maximum-likelihood estimates. The dichotomy between the two branches of Bayesian modeling is therefore mostly a matter of whether the expression for the MAP estimate contains the stimulus uncertainty.

Nevertheless, many concepts from signal detection theory, such as the ROC curve, are generally useful and can also be applied to binary decision tasks in which the expression for the MAP estimate contains the stimulus uncertainty. In fact, we will do so in the next chapter, when discussing more complex generative models.

7.6 Summary and remarks

In this chapter we have introduced the Bayesian framework for binary decision-making, also known as signal detection theory. We have learned the following:

- The proportion correct is an impoverished way of representing performance, since it does not distinguish between the two types of correct responses, hits and correct rejections.
- Hit rate and correct rejection rate themselves depend on sensitivity and criterion or bias.

- The receiver-operating characteristic (ROC) is a curve obtained by varying the criterion. Both sensitivity and the area under the ROC are bias-independent measures of performance.
- Hit and false-alarm rates, as well as the ROC, extend to any binary decision.
- Sensitivity is specific to the Gaussian assumption. When the distributions of the decision variable are not Gaussian with equal variance as they were here, Eq. eq:7:dprime2 no longer follows from the definition, Eq. eq:7:dprime. Whatever its form, d' loses its significance if it depends on the criterion.
- Binary and continuous variables are two ends on a spectrum. A stimulus variable that is discrete but has a large number of possible values comes close to being continuous. An example would be choosing in which of 8 directions a cloud of dots is moving. All probability distributions in the Bayesian model would be probability mass functions rather than probability density functions. In that sense, all Bayesian inference on a discrete stimulus variable is very similar to binary decisions. However, many of the concepts introduced in this chapter, such as log posterior ratios, decision rules, and ROCs, are not natural concepts when there are multiple alternatives.
- The type of binary decisions considered in this chapter have been rather limited, namely only those where the class C uniquely specifies the stimulus (whose values we denoted s_+ and s_-). Much more general is the case where each class C determines a *distribution* over the stimulus. For example, in a typical orientation discrimination task, the subject is not asked to discriminate between 2° to the right and 2° to the left of the vertical, but between any leftward tilted and any rightward tilted stimulus. To treat this case properly, we need to introduce the concept of marginalization, which we will do in the next chapter.

7.7 Suggested readings

- George A Gescheider. *Psychophysics: the fundamentals*. Psychology Press, 2013
- David Marvin Green, John A Swets, et al. *Signal detection theory and psychophysics*. Volume 1. Wiley New York, 1966
- Michael J Hautus, Neil A Macmillan, and C Douglas Creelman. "Detection Theory: A User's Guide". In: (2021)
- Thomas D Wickens. *Elementary signal detection theory*. Oxford university press, 2001

7.8 Problems

Problem 7.1 In medicine, it is common to encounter the terms *sensitivity* and *specificity* for a diagnostic test for a disease; these are synonyms for the true-positive rate and the true-negative rate, respectively. In addition, the (objectively correct) prior probability of a disease is called its *prevalence*. The *positive predictive value* (PPV) is the probability that someone has the disease given that they test positive. Use Bayes' rule to show that

$$\text{PPV} = \frac{\text{sensitivity} \cdot \text{prevalence}}{\text{sensitivity} \cdot \text{prevalence} + (1 - \text{specificity}) \cdot (1 - \text{prevalence})}. \tag{7.41}$$

Problem 7.2 Suppose the prior distribution and the posterior distribution are as in **Fig. 7.2**.

a) Calculate the likelihood ratio.
b) Does the sensory evidence alone (without the prior) indicate that the stimulus was s_+ or s_-?
c) Do the prior and the likelihood favor the same alternative?

Problem 7.3 In this problem, we explore the relationship between posterior probabilities and log posterior ratio numerically.

a) Create a vector of 99 possible posterior probabilities of s_+, from 0.01 to 0.99 in steps of 0.01. For each value, calculate the log posterior ratio $d = \log \frac{p(s=s_+|x)}{p(s=s_-|x)}$. Then plot this ratio as a function of the posterior probability of s_+. This should show that every posterior probability corresponds to exactly one log posterior ratio and the other way round (we are dealing with monotonous functions). Knowing one is as good as knowing the other.

b) Why did we not include the posterior probabilities 0 and 1?

c) Suppose you know the log posterior ratio d. Express the posterior probability of $s = s_+$, $P(s = s_+|x)$, as a function of d only. Do the same for $p(s = s_-|x)$.

d) If the log posterior ratio is 0.1, what are the posterior probabilities of s_+ and s_-? What if the log posterior ratio is 1?

Problem 7.4 We formulated the decision rule as reporting one alternative $d > 0$ and the other when $d < 0$. Why does the case $d(x) = 0$ usually not have to be considered? What would the observer do when $d(x) = 0$?

Problem 7.5 Prove Eq. (7.10) for the log likelihood in the Gaussian measurement model.

Problem 7.6 Suppose the stimulus s can take two values: $s_+ = 1°$ and $s_- = -1°$. Suppose that the measurement is normally distributed around s with standard deviation 0.5°. On a given trial, the observer's measurement is $-0.1°$, and $s = s_+$ occurs on 80% of trials. Would an optimal observer report that the stimulus was s_+ or s_-? Provide all the steps in your reasoning.

Problem 7.7 We want to choose our criterion k for a decision making task so that the probability of being correct is maximized. Starting from Eq. (7.20), derive an expression for the criterion k.

Problem 7.8 Above we introduced two confidence measures confidence1 and confidence2. Prove Eq. (7.34) for the relation between the two confidence measures.

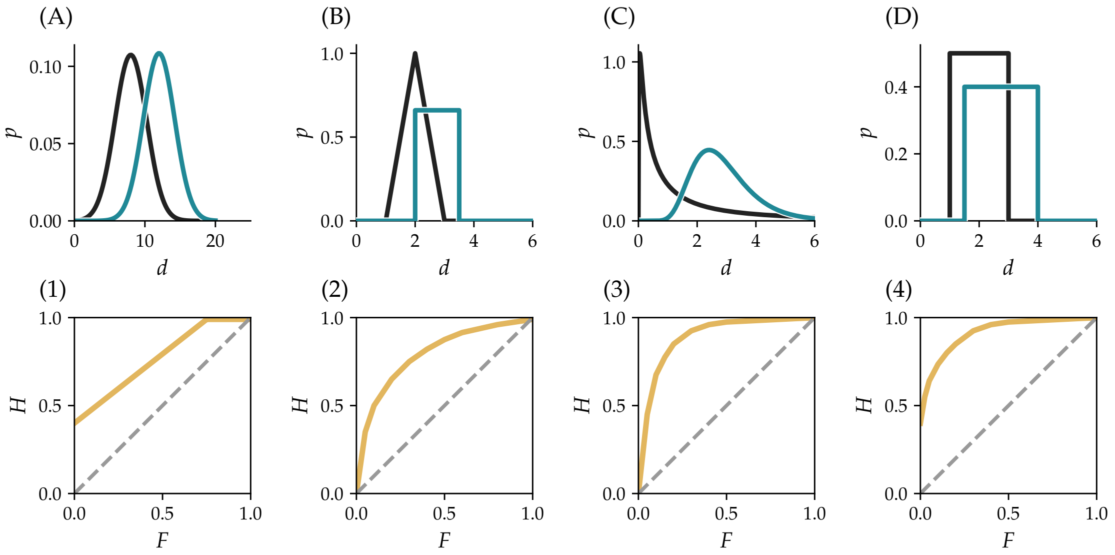
Problem 7.9 In the context of our discrimination task, assume a flat prior, so that the Bayesian decision variable d becomes the log likelihood ratio. We can think of d as a random variable that "inherits" its distribution from the distribution of x. Derive that the conditional distributions of the decision variable are

$$p_{d|s}(d|s_+) = \mathcal{N}\left(d; \frac{-(\Delta s)^2}{2\sigma^2}, \frac{(\Delta s)^2}{\sigma^2}\right) \tag{7.42}$$

$$p_{d|s}(d|s_-) = \mathcal{N}\left(d; \frac{-(\Delta s)^2}{2\sigma^2}, \frac{(\Delta s)^2}{\sigma^2}\right) \tag{7.43}$$

Problem 7.10 [4] In the first row of the figure below, each plot shows the distributions of the log posterior ratio under each of the two alternatives in a binary decision task. The second row displays receiver-operating characteristic (ROC) curves. Indicate for each ROC to which plot in the top row it belongs.

[4] Thanks to Ronald van den Berg for this problem.



Problem 7.11 We will simulate the ROC in a detection problem. An observer is trying to detect a signal of strength $s_+ = 3$ in noise ($s_- = 0$). The noise has a normal distribution with standard deviation $\sigma = 2$. On each trial, an experimenter presents noise (probability 0.4), or noise plus signal (probability 0.6). The task of the observer is to respond whether the signal is present or absent.

a) Simulate the stimulus (signal or noise) on each of the 100,000 trials. Save as a column vector.

b) Simulate the measurement on each trial.

c) Based on the measurements in part (b), calculate two measurement histograms: one for the trials when the signal was present and one for the trials when the signal was absent. Use as basis for your histograms a set of 50 bins, linearly spaced between -10 and 10. Normalize both histograms. Plot both in the same plot as lines (not as bars).

d) Based on the measurements in part (b), calculate the log posterior ratio on each trial. Calculate and plot the histograms of the log posterior ratio analogous to the histograms of the measurement in part (c).

e) Assume now that on each trial, the observer also provides a confidence rating by reporting "high confidence" when the absolute value of the log posterior ratio exceeds 2, "medium confidence" when it lies between 1 and 2, and "low confidence" when it is lies between 0 and 1. Create a 2-by-6 table of the two possible stimuli (signal present or absent) and the six possible responses. In each cell, put the frequency of the response (normalized by row).

f) Calculate the empirical ROC by cumulatively summing the response frequencies.

g) Plot the resulting points on top of the theoretical ROC based on Eqs. (7.38).

h) Simulate and describe what happens to the ROC when you reduce the signal strength to $s_+ = 2$?

i) Interpret the change. Why does it make intuitive sense?

Problem 7.12 This is a mathematical problem extending the formalism of this chapter to measurement noise with unequal (stimulus-dependent) variance. Consider our discrimination task with two possible stimulus values, s_+ and s_-. Assume equal probabilities (flat prior). In the chapter, we assumed that the distributions of the measurement, $p(x|s = s_+)$ and $p(x|s = s_-)$, were normal with equal variance. Now assume instead that their variances are different and have values σ_+^2 and σ_-^2, respectively.

a) Show that the log posterior ratio is given by

$$d = \log\frac{\sigma_-}{\sigma_+} - \frac{1}{2}\left(\frac{(x-s_+)^2}{2\sigma_+^2} - \frac{(x-s_-)^2}{2\sigma_-^2}\right). \tag{7.44}$$

b) Now assume $s_+ = 3$, $s_- = 0$, $\sigma_+ = 3$, and $\sigma_- = 1$. Plot the log posterior ratio as a function of x.

c) Interpret the shape of this function. Compare and contrast with **Fig. 7.3**

d) Without assuming specific values, simplify the Bayesian decision rule $d > 0$ to a set of inequalities for x. Why do you get two inequalities rather than one?

e) Derive expressions for the hit and false-alarm rates in terms of the standard cumulative normal distribution Φ_{standard}.

f) Numerically calculate the hit and false-alarm rates for the values in part (b).

Problem 7.13 We continue our extension to unequal variances but now treat the problem through simulations. Assume $s_+ = 3$, $s_- = 0$, $\sigma_+ = 3$, and $\sigma_- = 1$.

a) Simulate the stimulus (signal or noise) on each of the 100,000 trials. Save as a column vector.

b) Simulate the measurement on each trial.

c) Based on the measurements in part (b), calculate two measurement histograms: one for the trials when the signal was present and one for the trials when the signal was absent. Use as basis for your histograms a set of 50 bins, linearly spaced between -10 and 10. Normalize both histograms. Plot both in the same plot as lines (not as bars).

d) Based on the measurements in part (b) and the answer to Problem 7.12a,calculate the log posterior ratio on each trial. Calculate and plot the histograms of the log posterior ratio analogous to the histograms of the measurement in part (c).

e) Based on the measurements in part (b), calculate the hit and false-alarm rates. Compare to what you found in Problem 7.12f.

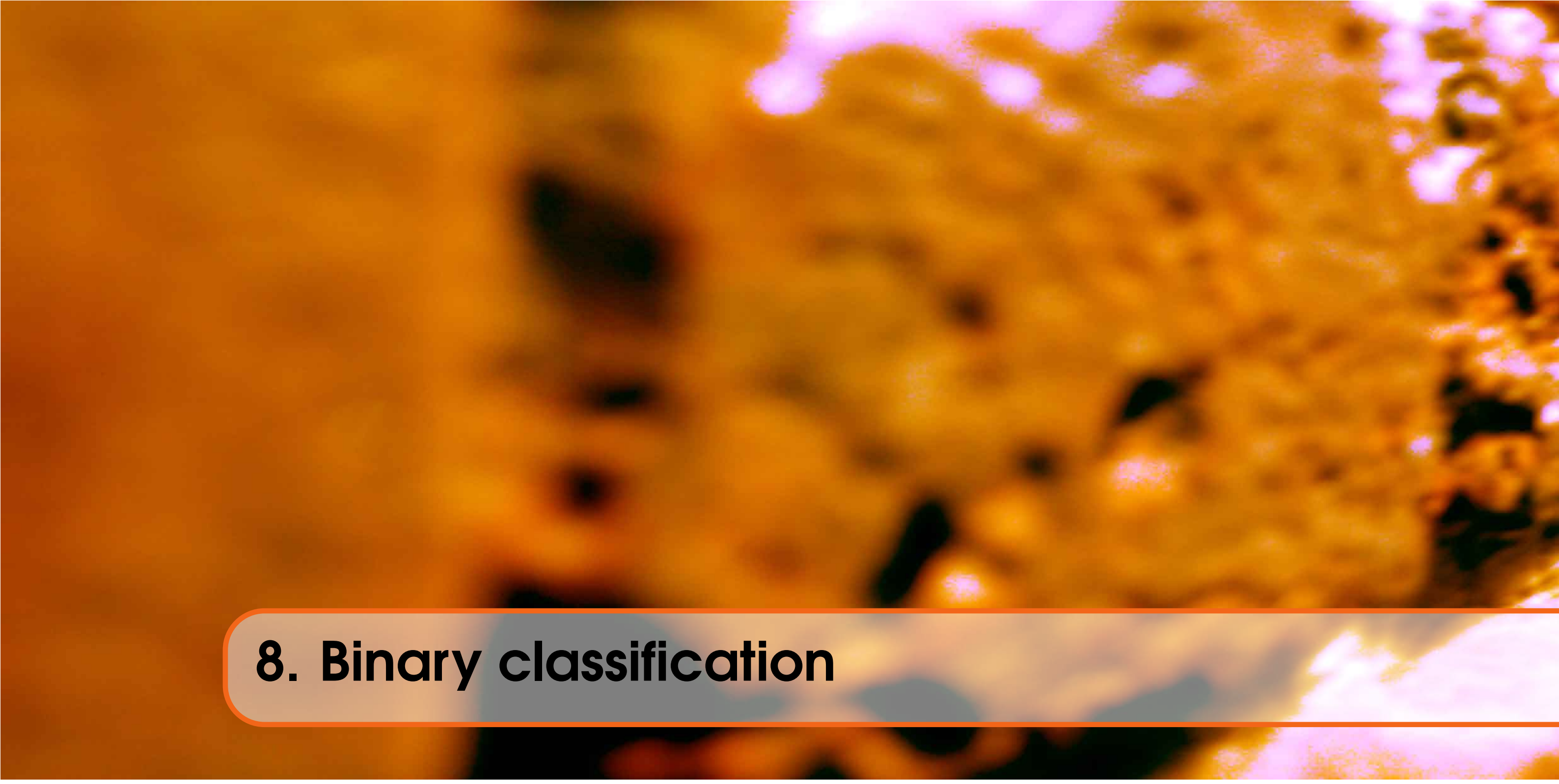
Problem 7.14 Here, we combine cue combination (Chapter 5) with the discrimination task of the current chapter. A judge in a trial is trying to determine whether a suspect is guilty. The juror's prior is 0.5. The judge has three conditionally independent pieces of evidence. If they had had only one of these pieces (any one of them), the posterior probability that the suspect is guilty would have been 60%. Now that they have all three pieces, what is the posterior probability that the suspect is guilty? Note that this problem only uses the conditional independence assumption Eq. (5.1) from Chapter 5, not the Gaussian noise assumption Eq. (5.3). We want to emphasize here as well that the use of Bayes rule is not entirely unchallenged in the world of jurisdiction but we would like you to answer this question as if this problem was already overcome.

Problem 7.15 Like the previous problem, this problem is about cue combination for a binary stimulus. However, we are now more specific about the measurement, and assume a Gaussian measurement model. Consider a stimulus s that takes two values, s_+ and s_-, with equal probability. The observer makes not one but two, conditionally independent measurements, x_1 and x_2, drawn from normal distributions with mean s and variance σ^2.

a) Derive an expression for the log posterior ratio.

b) What is the optimal decision rule in terms of the measurements?

c) Repeat parts (a) and (b) for N instead of 2 conditionally independent measurements.



8. Binary classification

How do we determine to which of two categories a stimulus belongs?

In Chapter 7, we introduced discrimination and detection tasks, tasks that require the observer to decide between two specific stimulus values, which we called s_+ and s_-. We concluded the chapter by noting that these tasks far from cover all binary decision tasks: in the real world as well as in the laboratory, the choice in a binary decision is often not between two specific stimulus values, but between two *categories* or *classes*, each of which comprise multiple, in some case infinitely many, stimulus values. Three real-world examples:

- Deciding whether a distant car on a country road is moving towards or away from you. You are categorizing the velocity vector by its sign (positive or negative).
- Deciding whether the person approaching you is the friend you are waiting for or not. You are categorizing the image as "friend" or "other".
- Deciding whether a cloud is a rain cloud or not.

Two laboratory examples:

- Deciding whether a noisy cloud of moving dots has a net motion direction to the right or left. You are categorizing the net motion vector by its sign (left or right).
- Deciding in a split second whether a natural scene contains an animal or not.
- Deciding if a stimulus that has both a size and an orientation belongs to Category 1 or Category 2, both of which are defined by the experimenter.

Plan of the chapter

In the present chapter, our focus is on such *binary classification* tasks. In these tasks, the observer is asked not to report the stimulus, only the stimulus category. Yet, the stimulus value is still unknown to the observer. Such tasks are richer than discrimination and detection, because the dependence of the observer's behavior on the stimulus can be studied.

In terms of the mathematical machinery, binary classification requires a Bayesian observer to integrate over all possible values of this unknown stimulus, an operation known as *marginalization*. Marginalization is a central operation in Bayesian models of virtually all tasks, except the very simplest ones. Thus, this chapter is the gateway to a large domain of applications of Bayesian models.

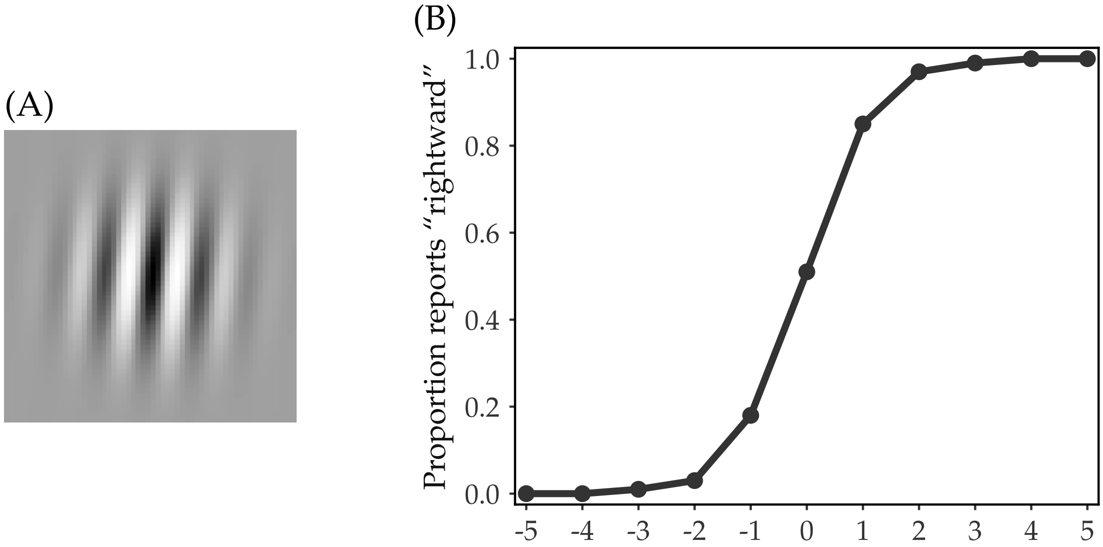


Figure 8.1: A binary classification experiment. **(A)** Example Gabor stimulus. **(B)** Psychometric curve.

8.1 A binary classification experiment

Although any of the examples above could be used to build a formal model, we will here use an extremely simple single-feature example, so that few additional assumptions are needed and we can focus on the essence of classification. Consider the following common visual task. You are briefly shown an oriented pattern like the one in **Fig. 8.1A**. The orientation of the pattern varies among many possible values, and your job is to report through a key press whether the pattern was tilted left (counterclockwise) or right (clockwise) of vertical. The data in this task consists of a set of pairs of stimulus and a category report. If the stimuli are discrete, for example from -5° to 5° in steps of 1°, then it is common to calculate for each presented stimulus the proportion of trials in which the subject reports one alternative, say "right". This proportion can be plotted as a function of the stimulus (**Fig. 8.1B**). The result is an example of a *psychometric curve*: a curve that has some summary of human behavior plotted against a physical quantity that is varied by the experimenter. It stands to reason that the more rightward the stimulus is, the greater will be the proportion of "rightward" responses. One outcome of the Bayesian modeling in this section will be a model for the psychometric curve. Let's first consider two caveats about the psychometric curve:

- If the stimulus can take a large number of values or is continuous, then to plot a visually useful psychometric curve, the stimulus values have to be binned or otherwise grouped together. While useful for visualization, such grouping loses information. Therefore, any quantitative data analysis is ideally based on the raw data.
- Even if the stimulus can take only a small number of values, as in **Fig. 8.1B**, then the psychometric curve does not capture the entirety of the raw data: what is missing is the number of trials on which each stimulus is presented. In many experiments, this number would be the same for all possible stimuli and for all subjects, so it only needs to be reported once. It is always an option to plot within-subject error bars; their magnitude will reflect the number of trials.

Even though a task like this is often called a "discrimination" task, that terminology is inaccurate; it is not discrimination, but falls under the broader umbrella of *classification* or *categorization*. What is the difference? In discrimination, the number of values the stimulus can take and the number of possible responses are both 2. In classification, the stimulus may take more than 2

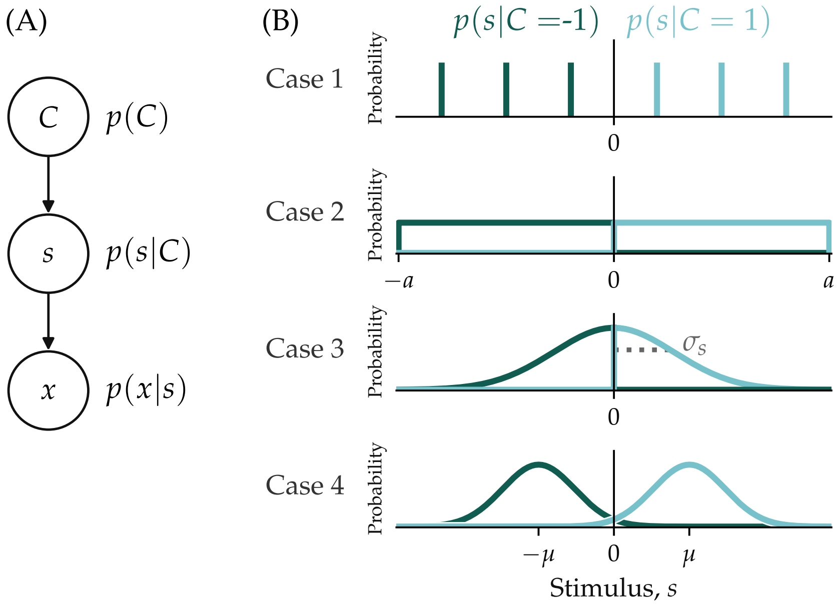


Figure 8.2: (A) Generative model diagram. **(B)** Examples of mirror-imaged class-conditioned stimulus distributions.

values, but the number of possible responses (usually 2) is smaller than the number of stimulus values. Thus, discrimination and the restricted form of detection in Chapter 7 are special cases of classification.

Task type	Number of distinct stimuli	Possible responses
Discrimination	2	2 stimulus identifiers
Detection	2 (absent and present)	2 (absent and present)
Identification	$n > 2$	n stimulus identifiers
Search (detection)	n	2 (absent and present)
Search (localization)	n	n locations
Classification / categorization	n	$< n$ categories

Table 8.1: Task types.

For our classification experiment, we will follow the same recipe as in earlier chapters: generative model, inference, and distribution of estimates. However, the generative model will now have an interesting extra ingredient, namely *class-conditioned stimulus distributions*.

8.2 Generative model

The generative model diagram is shown graphically in **Fig. 8.2A**. It has three nodes: class C, stimulus s, and measurement x. The observer is asked to report C, the world state of interest. As in earlier chapters, each node is associated with its own probability distribution. We will now discuss these one by one.

Class. We designate the two possible values of C by 1 (in the example: rightward) and -1 (in the example: leftward). We could have chosen any two values here, but in the example of **Fig. 8.1**, C is naturally equal to the sign of s. Associated with C is a distribution $p(C)$, which is specified by two values, $p(C=1)$ and $p(C=-1)$, which have to sum to 1. These probabilities reflect the believed prevalence of rightward and leftward tilted stimuli in the experiment. In many experiments, a class is chosen randomly with probability 0.5, so that, if the observer knows this, $p(C=1)$ and $p(C=-1)$=0.5; however, we will not restrict ourselves to that case.

Stimulus. When the class C equals -1, the experimenter draws the stimulus randomly from one set of values; when $C=1$, the experimenter draws the stimulus randomly from the other set of values. We denote the corresponding stimulus distributions by $p(s|C=-1)$ and $p(s|C=1)$, respectively; these are *class-conditioned stimulus distributions* (CCSDs). Here, we assume that the observer's believed CCSDs are the same as the experimental CCSDs.

Mirror-symmetric class-conditioned stimulus distributions. In binary classification, an important type of CCSDs are ones that are mirror images *of each other*. Four specific examples of this type of CCSD are (**Fig. 8.2B**):

- Case 1 ("Method of constant stimuli"): The stimulus is discrete and chosen with equal probability from n possible values: for $C=1$, they are $s_1, s_2, \ldots, s_n$ (all positive numbers), and for $C=-1$, they are mirror-symmetric: $-s_1, -s_2, \ldots, -s_n$ (all negative numbers). This procedure is called the *method of constant stimuli*.
- Case 2: The stimulus is continuous and drawn from a uniform distribution on an interval $[-a,0]$ when $C=-1$, and from a uniform distribution on $[0,a]$ when $C=1$, where a is a positive number.
- Case 3: The stimulus is continuous and drawn from a Gaussian distribution with mean 0, but then designated to be class -1 or class 1 based on its sign.
- Case 4: The stimulus is continuous and drawn from a Gaussian distribution (variance σ_s^2) with mean $-\mu$ when $C=-1$ and mean μ when $C=1$.

Measurement. The final step of the generative model is the same as in previous chapters: we assume that the observer's measurement x is drawn from a normal distribution centered at the stimulus, with standard deviation σ:

$$p(x|s)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-s)^2}{2\sigma^2}}. \tag{8.1}$$

8.3 Marginalization

So far, we have specified the distribution of the measurement conditioned on the stimulus, $p(x|s)$, and the distribution of the stimulus conditioned on the class, $p(s|C)$. We do not, however, have the distribution of the measurement conditioned on the class, $p(x|C)$, which is the distribution needed to do the inference. To obtain an expression for $p(x|C)$, we will first introduce an important general mathematical identity called *marginalization*. We will then apply marginalization to the generative model of **Fig. 8.2A**.

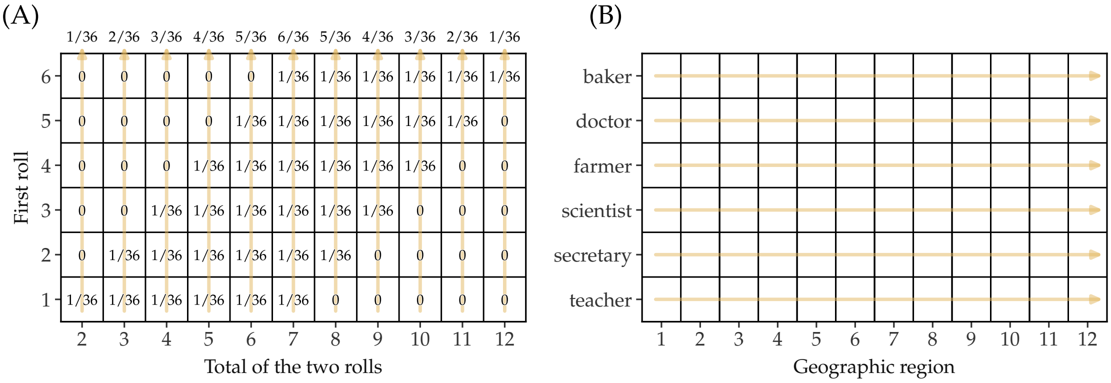
8.3.1 The sum of two die rolls

In probability theory, marginalization is the operation of turning a probability distribution over multiple variables into a distribution over one of them. For example, if a and b are discrete random variables, and $p(a,b)$ is their joint distribution, then summing the joint distribution over b produces the distribution over a:

$$p(a)=\sum_b p(a,b). \tag{8.2}$$

By using the definition of conditional probability, we can also write this as

$$p(a)=\sum_b p(a|b)p(b). \tag{8.3}$$



First roll \ Total of the two rolls	2	3	4	5	6	7	8	9	10	11	12
	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36
6	0	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36
5	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0
4	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0
3	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0
2	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0
1	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	0

Figure 8.3: Marginalization. Each panel shows a joint probability distribution over two variables. The brown lines represent marginalization over the nuisance variable, a procedure that reduces the two-dimensional distribution to a one-dimensional distribution over the variable of interest. **(A)** Dice example. The value in each square is the probability of that particular (first roll value, total value) pair. Marginalization over the first roll value results in a probability distribution over the total value (top numbers). **(B)** Farmer example. Values within each square (not shown) would represent the proportion of citizens characterized by the corresponding (occupation, geographic region) pair. Marginalization over region results in a probability distribution over occupation. (Only a small subset of occupations is shown).

As an example, suppose we roll two fair dice, one at a time. The game we are playing rewards us if the total score from the two rolls is 10. What is the probability that this will occur? To find out, we can consider the probability of every value resulting from the first roll, and the probability of a total of 10 given that first value:

$$p(\text{total} = 10) = \sum_{i=1}^{6} p(\text{total} = 10|\text{first roll} = i)p(\text{first roll} = i) \tag{8.4}$$

We are *marginalizing* over the value of the first roll. To express the marginalization formula in words, we replace each product with "and" and each addition with "or". We are stating that the probability of a total of 10 is the probability that the first die lands 1 AND the total will be 10 given that the first lands 1, OR that the first lands 2 AND the total will be 10 given that the first lands 2, and so on. To compute the marginalization sum, we note that if the first die lands 1, 2, or 3, it is impossible for the total of the two dice to reach 10; if the first die lands 4, 5, or 6, then the second die would need to land 6, 5, or 4, respectively, and each of these events occurs with probability $\frac{1}{6}$. Thus, we have:

$$p(\text{total} = 10) = 0 \cdot \frac{1}{6} + 0 \cdot \frac{1}{6} + 0 \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} + \frac{2}{6} \cdot \frac{1}{6} + + \frac{1}{6} \cdot \frac{1}{6} = \frac{3}{36} = \frac{1}{12}. \tag{8.5}$$

Note that we only care about the probability of the total value, but in order to calculate this, we must nevertheless consider all possible values of the first roll. Because we must take the first roll, which we don't really care about, into consideration, the value of the first roll is called a nuisance variable.

For a second example, suppose we want to know the probability that a randomly selected citizen of a particular country is a farmer. The country has 12 geographic regions. Suppose we find an almanac that reports the proportion of the population that lives within each region, and also the proportion of farmers within each region. To obtain the desired probability, we multiply those two proportions for every region, and then sum over all regions. Here, the region of residence is the nuisance variable:

$$p(\text{farmer}) = \sum_{i=1}^{12} p(\text{farmer}|\text{region}_i)p(\text{region}_i) \tag{8.6}$$

We are stating that the probability of randomly selecting a farmer is the probability that we will randomly select a person from region 1 AND that a randomly selected person from region 1 is a farmer, OR that we will randomly select a person from region 2 AND that a randomly selected person from region 2 is a farmer, and so on. Equivalently, we can conceptualize our procedure as first randomly selecting a region, with a probability proportional to the population of the region, and then randomly selecting a person from within that region.

Box 8.1 — Etymology. Where does the name "marginalization" come from? In its most basic form, marginalization is captured by Eq. (8.2), $p(a) = \sum_b p(a,b)$. Thus, we can think of marginalization as the sum over one dimension – the dimension we are not interested in estimating – of a joint probability distribution. When the sum is repeated for each value of the relevant dimension (e.g., not just for dice totals of 10, but for all totals; not just for farmers, but for all occupations), then marginalization reduces the joint distribution to a distribution over just the dimension of interest. Represented graphically, the summation occurs towards the "margin" of the joint distribution, giving marginalization its name (**Fig. 8.3**). ■

8.3.2 Continuous variables

In many cases we have to deal with latent variables that we need to marginalize that are continuous. If b is a continuous variable, marginalization consists of an integral:

$$p(a) = \int p(a,b)db \tag{8.7}$$

$$= \int p(a|b)p(b)db \tag{8.8}$$

In such cases, everything about marginalization stays the same, the only difference is that sums are replaced with integrals.

One example of such a continuous marginalization is if we want to calculate the probability distribution of the sum of two continuous variables a and b. This can be thought of as a continuous analog of the "sum of two dice" example. We assume that both a and b have Gaussian distributions:

$$p(a) = \mathcal{N}(a;\mu_a,\sigma_a^2) \tag{8.9}$$

$$p(b) = \mathcal{N}(b;\mu_b,\sigma_b^2) \tag{8.10}$$

Denoting the sum variable by c, we have

$$p(c) = \mathcal{N}(c;\mu_a+\mu_b,\sigma_a^2+\sigma_a^2) \tag{8.11}$$

Exercise 8.1 Show this, either by solving the integral or by referring to an equation derived earlier. ■

8.3.3 Conditioned marginalization

The basic form of marginalization, Eqs. (8.2) and (8.3), still holds if all probabilities are already conditioned on other variables, e.g. c:

$$p(a|c) = \sum_b p(a,b|c) = \sum_b p(a|b,c)p(b|c). \tag{8.12}$$

For example, suppose that we want to know the probability that a randomly selected middle-aged citizen (e.g., a person 45-65 years old) from the country mentioned above is a farmer. If we define C=1 to be young adult, C=2 to be middle-aged, and C=3 to be older-aged, then conditioning on C=2, we could calculate:

$$p(\text{farmer}|C=2) = \sum_{i=1}^{12} p(\text{farmer}|\text{region}_i, C=2)p(\text{region}_i|C=2) \tag{8.13}$$

The probability that a randomly selected middle-aged person is a farmer is the probability that a randomly selected middle-aged person is from region 1 AND that a randomly selected middle-aged person from region 1 is a farmer, OR that a randomly selected middle-aged person is from region 2 AND that a randomly selected middle-aged person from region 2 is a farmer, and so on.

As another example of conditioned marginalization, let's consider the transmission of a particular virus. Suppose we are interested in the probability $p(x=1|C=1)$ that an unvaccinated person exposed to an infected unvaccinated person (we denote this exposure by $C=1$) themselves gets infected by a virus ($x=1$). A complication is that the virus comes in many variants, which we will denote by s_i, with $i=1,\dots,n$ (here, n is the number of variants); the variable s_i is the nuisance variable in the problem. Each variant has its own transmission rate for unvaccinated people, $p(x=1|s_i,C=1)$, and its own prevalence among infected unvaccinated people, $p(s_i|C=1)$. To obtain the answer to our question, we first multiply those two proportions for every variant, yielding $p(x=1|s_i,C=1)p(s_i|C=1)$ for every i. This is interpreted as the probability that an infected unvaccinated person has variant s_i *and* transmits it. Finally, we sum over all variants. This calculation implements the equation

$$p(x=1|C=1) = \sum_{i=1}^{n} p(x=1|s_i,C=1)p(s_i|C=1). \tag{8.14}$$

Exercise 8.2 This book was finalized in its current version during the COVID 19 Pandemic. Discuss how this phrasing of the problem would apply to pandemics and which of the modeling assumptions are justified. Specifically, what can you say about the relations between people? ■

8.3.4 Using the generative model

Let's now return to the case study of our chapter. Recall that the generative model specifies the distribution of the stimulus conditioned on the class, $p(s|C)$, and the distribution of the measurement conditioned on the stimulus, $p(x|s)$. Despite its central role in the generative model, the stimulus s is a *nuisance variable*; it is neither the observation (measurement, x) nor the variable of primary interest (C). In order to make an inference, we must calculate the distribution of the measurement conditioned on the class, $p(x|C)$, which we obtain by marginalizing over s:

$$\text{When } s \text{ is discrete: } p(x|C) = \sum_{i=1}^{n} p(x|s_i,C)p(s_i|C) \tag{8.15}$$

$$\text{When } s \text{ is continuous: } p(x|C) = \int p(x|s,C)p(s|C)ds \tag{8.16}$$

See appendix C for a problem where this is derived.

Eqs. (8.15)-(8.16) are completely general; they are mathematical identities that hold regardless of context. We now incorporate the structure of the generative model. In the generative model, the distribution of x only depends on s, and not directly on C. In **Fig. 8.2**, this is graphically understood by the fact that the only arrow pointing to x comes from s; there is no arrow from C to x. In other words, when s is known, knowledge of C is redundant when one is interested in the distribution of x. Mathematically, this is expressed by the fact that the conditional distribution $p(x|s,C)$ is identical to $p(x|s)$. Substituting this into Eqs. (8.15)-(8.16), we arrive at the following expressions for the class likelihood:

$$\text{When } s \text{ is discrete: } p(x|C) = \sum_{i=1}^{n} p(x|s_i)p(s_i|C) \tag{8.17}$$

$$\text{When } s \text{ is continuous: } p(x|C) = \int p(x|s)p(s|C)ds \tag{8.18}$$

These equations act as a kind of chain rule to link the class, C, to the measurement, x, by way of the intermediate variable, the stimulus, s.

8.4 Inference

On a given trial, the observer makes a measurement x. Since the observer is interested in class, C, the posterior distribution we want to calculate is now $P(C|x)$, not $p(s|x)$. This is the first time in the book that the stimulus, s, does not appear in the posterior: the stimulus is not directly of interest, only class is. Nevertheless, the logic of inference is exactly the same as in previous chapters. Just as in Chapter 7, the Bayesian observer decides based on the log posterior ratio, but now over class:

$$d \equiv \frac{P(C=1|x)}{P(C=-1|x)} \tag{8.19}$$

$$= \frac{P(C=1)}{P(C=-1)} + \frac{p(x|C=1)}{p(x|C=-1)} \tag{8.20}$$

The likelihood of class 1, $p(x|C=1)$, and the likelihood of class -1, $p(x|C=-1)$, can both be obtained from the corresponding distributions in the generative model, $p(x|C=1)$ and $p(x|C=-1)$. Because s is continuous, these distributions are given by the marginalization equation Eq. (8.18). The resulting likelihoods are:

$$\mathcal{L}(C=1;x) \equiv p(x|C=1) \qquad = \int p_{x|s}(x|s)p(s|C=1)ds \tag{8.21}$$

$$\mathcal{L}(C=-1;x) \equiv p(x|C=-1) \qquad = \int p_{x|s}(x|s)p(s|C=-1)ds. \tag{8.22}$$

From an inference perspective, these equations can be interpreted in terms of the “propagation" of uncertainty information: $p_{x|s}(x|s)$ as a function of s is the likelihood over the stimulus and represents sensory uncertainty. By contrast, $p_{x|C}(x|C)$ as a function of C is the likelihood over the class, which represents class uncertainty:

$$\underbrace{\mathcal{L}(C;x)}_{\text{likelihood of class}} \equiv p_{x|C}(x|C) \qquad = \int \underbrace{p_{x|s}(x|s)}_{\substack{\text{likelihood of } s,\\ \mathcal{L}(s;x)}} \underbrace{p(s|C)}_{\substack{\text{CCSD}\\ \text{(task-dependent)}}} ds \tag{8.23}$$

(and similarly for discrete s). Thus, the uncertainty about the “lower-level” variable of s gets transformed into or propagated to uncertainty about the “higher-level” variable of class. This transformation is mediated by the learned, “top-down” knowledge of the CCSDs. Unlike the likelihoods, the CCSDs depend only on the task (to be precise, on the observer’s beliefs about the task) and do not change from trial to trial.

Graphically, the likelihood of Class 1 is the “overlap” between the likelihood over s and the CCSD for Class 1: first multiply, then take the area (**Fig. 8.4**). If the resulting area is small, then it means that there was not much overlap.

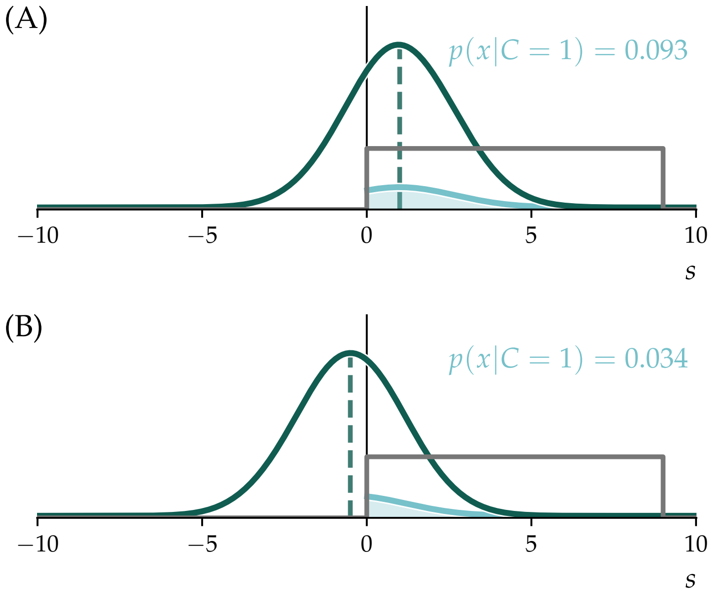


Figure 8.4: Graphical explanation of the calculation of the likelihood of a class.

Substituting Eqs. (8.17)-(8.18) back into Eq. (8.20), we find for the log posterior ratio:

$$\text{When } s \text{ is discrete: } d = \frac{P(C=1)}{P(C=-1)} + \frac{\sum_{i=1}^{n} p_{x|s}(x|s_i)p(s_i|C=1)}{\sum_{i=1}^{n} p_{x|s}(x|s_i)p(s_i|C=-1))} \tag{8.24}$$

$$\text{When } s \text{ is continuous: } d = \frac{P(C=1)}{P(C=-1)} + \frac{\int p_{x|s}(x|s)p(s|C=1)ds}{\int p_{x|s}(x|s)p(s|C=-1)ds} \tag{8.25}$$

The optimal decision rule is

$$\text{“report } \hat{C} = 1 \text{ if } d > 0 \text{ ".} \tag{8.26}$$

Like in Chapter 7, the prior over C biases the observer's decision, and its effect is stronger when the sensory evidence, as expressed by the log likelihood ratio, is weaker.

Outside of a few special cases, it is impossible to obtain an analytical decision rule out of Eq. (8.26). The best strategy is typically to numerically solve the inequality $d > 0$ for x.

Flat prior over class. Analytical progress *can* be made in an important special case, namely that the observer uses a flat prior over class, $P(C=1) = P(C=-1) = 0.5$. This would be the cases when the observer – correctly or incorrectly – believes that the classes are equally frequent (see the section on suboptimal Bayesian observers in Section 3.5). Since the prior is flat, MAP estimation is equivalent to reporting $C = 1$ when the likelihood $p(x|C=1)$ exceeds the likelihood $p(x|C=-1)$. This makes the problem completely symmetric, and the only sensible candidate for the optimal decision rule is $x > 0$. Proving this, however, is not that easy and we leave it to a problem [which one?].

We can, however, draw an important lesson from this special case. The decision rule $x > 0$ does not depend on the CCSDs $p(s|C)$, even though the general rule, $d > 0$, does (see Eq. (8.25)). That means that when the prior is flat, the observer might have a completely wrong belief about the shape of the CCSDs, yet make the optimal decision, simply because the wrong belief is irrelevant to the decision rule. Thus, wrong beliefs about the generative model do not always cause suboptimal behavior. (Keep in mind that we are still assuming $p(C=1) = 0.5$. If the prior over class were *not* flat, then the observer's beliefs about the CCSDs *do* matter both for the decision rule and for performance.) In short, in Bayesian modeling, an observer's wrong assumptions about the generative model do not always affect the decision rule.

> **Box 8.2 — Problem with the method of constant stimuli.** While Case 2 (discrete) is probably the most common CCSD in binary classification experiments, it is not ideal from the point of view of Bayesian model. The reason is that it is unlikely that the subjects learns exactly the experimental distribution, because it is so "spiky" and the locations of the "spikes" would have to be learned. Thus, the observer would most likely approximate the distribution by a continuous distribution, but it is not clear which one. This does not matter if the observer has a class prior of 0.5, but it will matter if you consider the possibility that the class prior is different from 0.5. It might still not matter *very much*, but you do not want to blindly count on that. ■

8.5 Response distribution

When the stimulus is s, the response distribution is given by the probability of reporting either $\hat{C}=1$ or $\hat{C}=-1$ given s. We evaluate the former probability:

$$P(\hat{C}=1|s) = P(d>0|s) \tag{8.27}$$

$$= P(x>k|s), \tag{8.28}$$

where k is the criterion that we can numerically compute in Step 2. Eq. (8.28) is the probability that a measurement x drawn from $p(x|s)$ will be greater than k. We can further evaluate this probability using Eq. (8.1):

$$P(\hat{C}=1|s) = \int_k^\infty \mathcal{N}(x;s,\sigma^2)dx \tag{8.29}$$

$$= \Phi_{\text{standard}}\left(\frac{s-k}{\sigma}\right), \tag{8.30}$$

where we used the same steps that gave us Eq. (7.20), and $\Phi_{standard}$ is the cumulative standard normal distribution (see Box 7.1). Eq. (8.30) has sensible properties. First, it is a monotonic function: when s increases (e.g. the orientation becomes more rightward tilted), then the probability of reporting $C=1$ increases as well. The curve has the characteristic sigmoid shape seen in **Fig. 8.1B**). Second, when there is more noise or the stimulus is closer to the criterion, the quantity $\frac{s-k}{\sigma}$ is smaller in absolute value (closer to 0), and the probability of reporting class 1 will be closer to 0.5; this makes sense because in both scenarios, the task will be harder.

Psychometric curve. Now we are finally ready to plot the psychometric curve predicted by the Bayesian model. It is given by Eq. (8.30) as a function of s. The predicted psychometric curve is a cumulative normal distribution that crosses 0.5 when $s=k$. In psychophysics parlance, C is the *point of subjective equality* (PSE): the value of the stimulus for which the subject (in this case, the Bayesian observer), reports the two classes equally often. The *slope* of the psychometric curve is usually defined as the inverse of the standard deviation of the cumulative normal, i.e. $\frac{1}{\sigma}$. Since in general, k depends on sensory noise level σ, the parameter(s) of the CCSDs, and the log prior ratio over class, the final psychometric curve also depends on all of those parameters.

Proportion correct. In Chapter 7, we computed hit rate H and false-alarm rate F in a detection task. Knowing H and F is equivalent to knowing H and the correct rejection rate, $1-F$. Here, we can similarly compute the probability of a correct report of Class 1, denoted by PC_1, and the probability of a correct report of Class -1, denoted by PC_{-1}. These probabilities are similar to the one in Eq. (8.30) for the response distribution, except that the conditioning is on a (true) class C

instead of on the stimulus s. Let's examine PC_1:

$$\begin{aligned}
\mathrm{PC}_1 &= p(\hat{C}=1|C=1) && (8.31)\\
&= p(d>k|C=1) && (8.32)\\
&= \frac{1}{n_{\text{trials },1}} \sum_{\substack{\text{trials } t:\\ C_t=1}} p(d>k|s_t) && (8.33)
\end{aligned}$$

where the sum is over all experimentally presented trials, labeled t, for which C_t was 1, $n_{\text{trials },1}$ is their number, and s_t is the stimulus on the t^{th} trial. The sum is almost always calculated numerically, as we will do in Problem X. The expression for PC_{-1} is analogous.

Overall proportion correct according to the model is a weighted sum of PC_1 and PC_{-1}:

$$\mathrm{PC} = p(C=1)\mathrm{PC}_1 + p(C=-1)\mathrm{PC}_{-1}, \tag{8.34}$$

where $p(C=1)$ and $p(C=-1)$ represent the true frequencies of $C=1$ and $C=-1$ trials.

Predictions for proportion correct, whether overall or split up by true C, are an impoverished description of predicted behavior. They involve a grand average over stimuli, whereas the psychometric curve predicts the probability of reporting class 1 for each value of the stimulus. More generally, when evaluating a model (after fitting its parameters) through its predictions for summary statistics of behavior, the summary statistics that are chosen are typically the result of a trade-off of ease of visualization (or the ease of numerical reporting) and the granularity of the statistic, with more granular meaning more informative. In this trade-off, trial-level predictions are usually hard to visualize but most informative, whereas a proportion correct is easiest to report but least informative. It is good practice to report the match between data and model using summary statistics of different levels of granularity.

Exercise 8.3 Think about how you could, for your favorite kind of experiment and model, evaluate the match between data and model. ■

8.6 Beyond mirror-image class-conditioned stimulus distributions

Although in many experiments that use yes/no binary classification, the CCSDs are mirror images of each other, this is unnecessarily restrictive. A world of new possibilities opens up if we consider other situations. For example, the friend and animal examples in the introduction do not feature symmetric classes. In both cases, one class is a restricted set of stimuli (images of friend/animal), whereas the other class is broad, encompassing class (images of random people, or images not containing animals). To treat such cases, we simplify them to their essence: a narrow class "embedded in" a wide class. While images are complex stimuli, we can define such classes this even for single-feature (one-dimensional) stimuli, and indeed experimenters have done so. The math does not fundamentally change from the previous section.

8.6.1 Generative model

We denote the classes by $C=1$ and $C=2$; since their CCSDs are not mirror images of each other, the notation $C=-1$ makes less sense for the second category. Class 1 stimuli are drawn from a Gaussian distribution with mean 0 and variance σ_1^2. Class 2 stimuli are drawn from a Gaussian distribution with mean 0 and variance σ_2^2, which is greater than σ_1^2. The distributions are illustrated in **Fig. 8.5A**, with example stimuli in panel B.

Ambiguity. When the CCSDs overlap (regardless of whether they are mirror images of each other), the observer cannot reach perfect performance even in the absence of sensory noise. This situation produces ambiguity: the same stimulus could have come from more than one class, though

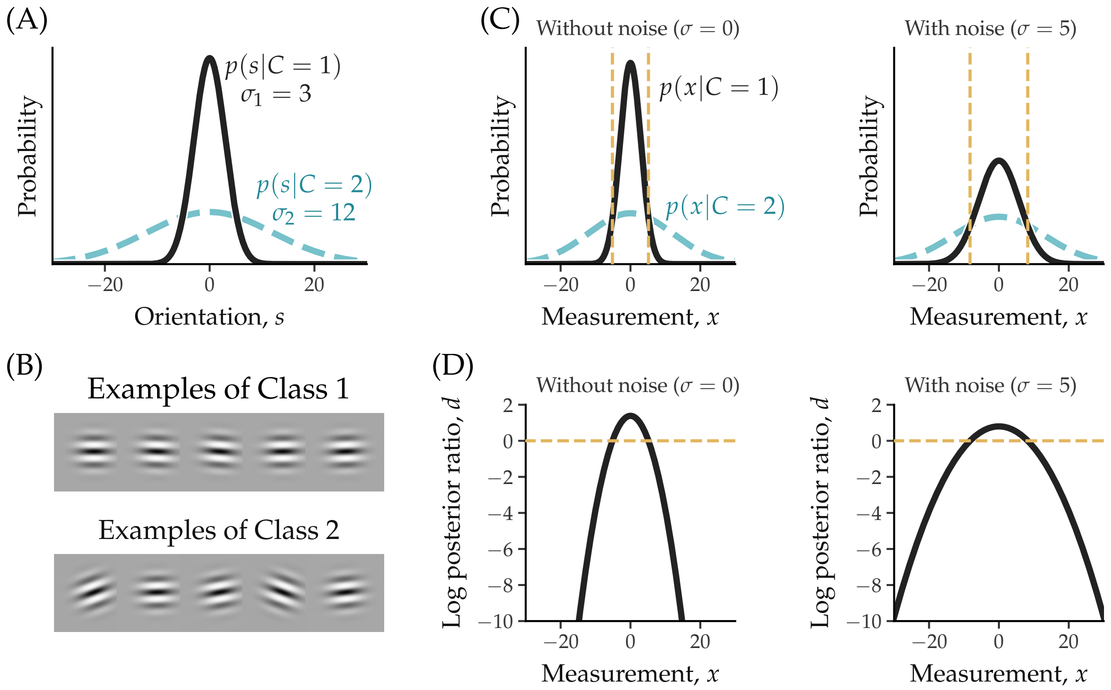


Figure 8.5: Embedded class task. **(A)** CCSDs over orientation (in °). The distributions have the same mean, but different standard deviations and are therefore not mirror images of each other. **(B)** Representative examples of stimuli in each class. **(C)** Class-conditioned *measurement* distributions at two different noise levels. The higher the noise level, the wider the CCMDs. The vertical dashed lines indicate the intersection points of the CCMDs of $C = 1$ and $C = 2$. **(D)** Corresponding log posterior ratio over C as a function of the measurement, assuming equal priors. When noise is higher, the evidence in favor of $C = 1$ falls off more slowly. The horizontal dashed line is at $d = 0$, the value that corresponds to the intersection points in **(C)**. The higher the noise, the larger the region for which $d > 0$ and the observer reports "$C = 1$".

usually not with the same probability. As we have noted previously (Section 1.3), ambiguity is common in perception, one example being that when viewing a visual scene with one eye, the retinal image could have been produced by many three-dimensional scenes. Here, the CCSD is the 3D scene-conditioned distribution of the retinal image.

8.6.2 Inference

The observer infers class C based on a measurement x. As in Eqs. (8.17)-(8.18), the likelihood of class C is given by an integral over s:

$$\mathcal{L}(C;x) = p(x|C) = \int p(x|s)p(s|C)ds. \tag{8.35}$$

This integral has a closed-form solution, which we already encountered in Eq. (8.11):

$$\mathcal{L}(C;x) = \mathcal{N}(x;0,\sigma^2 + \sigma_C^2), \tag{8.36}$$

where σ_C is either σ_1 or σ_2. The intuition is that the measurement x results from two independent noise processes: one external noise process with variance σ_C^2, and one internal noise process with variance σ^2. Per Box 4.2, the overall variance is the sum of these variances.

If the prior were flat, the MAP decision would be the ML decision. The ML decision rule, in turn, one would be able to graphically deduce from a plot of the class-conditioned measurement distributions (CCMDs) $p(x|C)$ at a given noise level σ (**Fig. 8.5C**): for any x, the maximum-likelihood decision is to pick the class for which the CCMD has a higher value. Thus, the decision switches at the intersection points of the two CCMDs.

If the prior is not flat, this is no longer the case. For a general prior, the log posterior ratio takes the form (see Problem):

$$d = \log \frac{p(C=1)}{p(C=2)} + \frac{1}{2} \log \frac{\sigma^2+\sigma_2^2}{\sigma^2+\sigma_1^2} - \frac{x^2}{2}\left(\frac{1}{\sigma^2+\sigma_1^2} - \frac{1}{\sigma^2+\sigma_2^2}\right) \tag{8.37}$$

We plotted this expression in **Fig. 8.5D**

This is the first time we have encountered a log posterior ratio that is *quadratic* in the measurement; previous log posterior ratios were all linear in the measurement. This makes perfect sense, however. It is clear from **Fig. 8.5D** that the probability of x is higher under Class 1 than under Class 2 when x falls in a narrow region around 0. To make this more precise: the MAP observer reports class 1 when $d > 0$. This inequality can be rewritten as one for the measurement, x. First, we observe that if $\log \frac{p(C=1)}{p(C=2)} + \frac{1}{2} \log \frac{\sigma^2+\sigma_2^2}{\sigma^2+\sigma_1^2} < 0$, then d is negative and the observer reports Class 2 regardless of the value of x.

Exercise 8.4 Why? ■

Second, if that condition is not satisfied, then the Bayesian MAP decision rule becomes

$$|x| < \sqrt{\frac{2\log \frac{p(C=1)}{p(C=2)} + \log \frac{\sigma^2+\sigma_2^2}{\sigma^2+\sigma_1^2}}{\frac{1}{\sigma^2+\sigma_1^2} - \frac{1}{\sigma^2+\sigma_2^2}}}. \tag{8.38}$$

In other words, there are two decision criteria on the measurement, one on each side of 0. The observer reports Class 1 when the measurement falls between those criteria. As in Section 8.4, criteria depend on the sensory noise level, the CCSD parameters σ_1 and σ_2, and the log prior ratio.

8.6.3 Response distribution

It is possible to express the probability that a Bayesian observer reports Class 1 in terms of cumulative normal functions. We will do this in Problem 8.5.

8.7 "Following the arrows"

As we will be exploring generative models of increasing complexity, it is helpful to have a clear and straightforward recipe for deriving an expression for a posterior of our choice based on the information provided by the generative model. Recall that the generative model specifies exactly which distributions are given in the problem. Each variable that does not have any arrows pointing to it follows a regular probability distribution. Each variable that does have arrows pointing to it follows a conditional distribution, where the conditioning is on the variable(s) from which the arrows originate; its distribution does not depend on any other variables in the problem. This gives us the recipe we are looking for:

1. Evaluate the joint distribution over all variables in the generative model by "following the arrows". Start at the top with the variables that have no arrows pointing to them. Working your way down, write down the prior or conditional probability distribution associated with each node, and multiply all distributions obtained this way.

2. Compute any conditional distribution by writing out its definition and marginalizing the joint distribution (i.e. summing or integrating over the variables not in the conditional).

When a posterior over a state-of-the-world variable is computed, a marginalization is done over every variable in the generative model other than the observations and the state-of-the-world variable. Note that in this recipe, the joint distribution is central, not the likelihood or the prior. In fact, Bayes' rule, which expresses the joint in terms of the likelihood and the prior, is simply the first step in the recipe of evaluating the joint distribution!

In the present chapter, the joint distribution is $p(C,s,x)$. Following the arrows in **Fig. 8.2A**, we find $p(C,s,x) = p(C)p(s|C)p(x|s)$. The posterior we are interested in is $p(C|x)$, which we obtain from marginalization. For the discrete case:

$$\begin{aligned}
p(C|x) &\propto p(C,x) && (8.39)\\
&= \int s p(C,s,x) ds && (8.40)\\
&= \int p(C)p(s|C)p(x|s) ds && (8.41)\\
&= p(C) \int p(s|C)p(x|s) ds. && (8.42)
\end{aligned}$$

If the generative model looks like a "string" of variables, each of which only receives an arrow from the previous one, such as here $C \to s \to x$, it is called a *Markov chain*. We will encounter other Markov chains in Chapter 12. As in Section 3.3.3, the proportionality sign in the first line is saying "we will calculate the *unnormalized* posterior (what we call the protoposterior); to get the posterior, normalize at the end". When the MAP estimate is the only quantity of interest, normalization is not even necessary.

8.8 Summary and remarks

In this chapter, we have analyzed how binary classification can be phrased in terms of Bayesian statistics. We have learned:

- Psychometric curves can characterize the dependency of the choice on the stimulus.
- Even for flat priors the observer calculation can be very complicated.
- In classification, the stimulus may take more than 2 values, but the number of possible responses (usually 2) is smaller than the number of stimulus values.
- If subjects learn about the stimulus distribution then the popular approach for measuring psychometric curves, the method of constant stimuli may become problematic.
- What makes these problems hard is the need to marginalize.
- Ultimately, Bayesian models are about representing information in probability distributions – but these do not have to be priors.

8.9 Suggested readings

- Rachel N Denison et al. "Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence". In: *Proceedings of the National Academy of Sciences* 115.43 (2018), pages 11090–11095
- Thomas L Griffiths and Joshua B Tenenbaum. "Theory-based causal induction." In: *Psychological review* 116.4 (2009), page 661
- Daniel Kersten, Pascal Mamassian, and Alan Yuille. "Object perception as Bayesian inference". In: *Annu. Rev. Psychol.* 55 (2004), pages 271–304
- Zili Liu, David C Knill, and Daniel Kersten. "Object classification for human and ideal observers". In: *Vision research* 35.4 (1995), pages 549–568

- Gregory L Murphy, Stephanie Y Chen, and Brian H Ross. "Reasoning with uncertain categories". In: *Thinking & Reasoning* 18.1 (2012), pages 81–117
- Ahmad T Qamar et al. "Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization". In: *Proceedings of the National Academy of Sciences* 110.50 (2013), pages 20332–20337

8.10 Problems

Problem 8.1 What would the psychometric curve of a noiseless observer look like? (Any class-conditioned stimulus distributions (CCSDs), and either non-Bayesian or Bayesian.)

Problem 8.2 In the context of Section 8.5, an observer's *75%-threshold* can be defined as the value of the stimulus for which the observer (who has observation noise of σ) reports "right" 75% of trials, minus the value of the stimulus for which the observer reports "right" 50% of trials. How many standard deviations σ does the 75%-threshold correspond to? As a reminder:

$$P(\hat{C}=1|s) = \Phi_{\text{standard}}\left(\frac{s-k}{\sigma}\right)$$

Problem 8.3 Consider binary classification with a general class prior $p(C)$ and a Gaussian measurement distribution. Please refer to the Class Conditioned Stimulus Distributions (CCSDs) in Section 8.2.

a) Derive the decision rule of the Bayesian observer in Case 2 (CCSDs uniform on an interval). The answer will involve more than one cumulative standard normal distribution (Φ_{standard}), and it is not a pretty rule.

b) Repeat for Case 4 (CCSDs are equivariant Gaussians).

Problem 8.4 Consider Case 3 in Section 8.2, where the stimulus is continuous and drawn from a Gaussian distribution with mean 0, where the class is -1 if the stimulus is below zero and 1 otherwise.

a) Show that the log likelihood ratio is equal to

$$\text{LLR} = \log \Phi_{\text{standard}}\left(\frac{x\sigma_s}{\sigma\sqrt{\sigma^2+\sigma_s^2}}\right) - \log\left(1-\Phi_{\text{standard}}\left(\frac{x\sigma_s}{\sigma\sqrt{\sigma^2+\sigma_s^2}}\right)\right). \quad (8.43)$$

b) Set $\sigma = 1$. Plot the LLR as a function of x in two cases: $\sigma_s = 1$ and $\sigma_s = 10$ (two curves in one plot). You should see that σ_s has a big effect.

c) Set $\sigma = 1$ and log prior ratio to 0.2. Numerically solve $d = 0$ for x in two cases: $\sigma_s = 1$ and $\sigma_s = 10$. You may do this by choosing a fine grid for x, then finding the zero-crossing of d using interpolation.

d) Use the two numbers found in (d) to plot the psychometric curves of the optimal observer for $\sigma_s = 1$ and $\sigma_s = 10$ (two curves in one plot). You should see that σ_s does not have much of an effect. Thus, CCSD parameters sometimes greatly affect confidence (log posterior ratio) while only minimally affecting choice (decisions).

Problem 8.5 Consider the binary classification (C_1 vs C_2) task of Section 8.6 where both classes are equally probable ($p(C=1) = p(C=2) = 0.5$.), both stimuli are drawn from Gaussian distributions with a mean of 0, and where the variance C_1 stimuli is $\sigma_1^2 = 9$ and that of C_2 stimuli is $\sigma_2^2 = 144$.

a) Plot the distributions $p(s|C=1)$ and $p(s|C=2)$. Use the plot to explain why even an optimal observer cannot be 100% correct on this task.

b) For general σ_1 and σ_2, derive Eq. (8.37), the equation for the log posterior ratio.

c) Derive an expression for the probability that the optimal observer reports Class 1 when the true stimulus is s. Use the standard cumulative normal distribution, Φ_{standard}, defined in Box 7.1.

d) Plot this probability as a function of s (between -30 and 30) for $\sigma = 10$. This is the psychometric curve of the optimal observer with $\sigma = 10$. Do the same for $\sigma = 1$. Plot both cases in the same plot.

e) Interpret the differences between the two curves.

f) Derive expressions for "hit rate" (probability of reporting Class 1 when the true class was 1), "false-alarm rate" (probability of reporting Class 1 when the true class was 2), and proportion correct.

g) Plot all three expressions as a function of σ in the same plot.

h) Interpret the plot.

Problem 8.6 In this problem, we discuss unmodeled errors. Suppose that we perform an experiment with binary responses ($r = 0$ or $r = 1$) and that $p(r|s)$ expresses the predicted probability of the observer's response – under an arbitrary model, Bayesian or non-Bayesian – when the stimulus is s.

a) Suppose that the observer accidentally presses the wrong key on a proportion λ of all trials. How does this change the predicted probability of the observer's response? Remark, much of the human psychophysics literature makes such an assumption and calls the relevant effect lapse.

b) Suppose that the observer makes a random guess on a proportion g of all trials (e.g. because he sometimes did not pay attention and didn't see the stimulus). How does this change the predicted probability of the observer's response?

Problem 8.7 Consider binary classification with a flat prior, mirror-image CCSDs (i.e., $p(s|C = 1) = p(-s|C = -1)$), and a measurement distribution $p(x|s)$ that is symmetric around s (though not necessarily Gaussian). Show that the Bayesian MAP observer has the decision rule $x > 0$, where x is the observed measurement.

Problem 8.8 Consider an observer performing binary classification with a class distribution $p(C)$ and mirror-image, non-overlapping CCSDs $p(s|C = -1)$, nonzero for $s < 0$, and $p(s|C = 1)$, nonzero for $s > 0$. In the chapter, we described the decision strategy of a Bayesian observer in this task. A student suggests an alternative decision strategy, namely that the observer first calculates the overall stimulus distribution,

$$p(s) = p(s|C = 1)p(C = 1) + p(s|C = -1)p(C = 1) \tag{8.44}$$

then uses this as a prior to compute a posterior over s, then compares the mean of this posterior, which is an estimate of s, to 0 (the midpoint of our two CCSDs).

a) Show using equations that the resulting decision rule is equivalent to the decision rule of a Bayesian observer who assumes incorrect CCSDs when doing inference, namely

$$q(s|C = -1) \propto -sp(s|C = -1) \tag{8.45}$$
$$q(s|C = 1) \propto sp(s|C = 1). \tag{8.46}$$

b) Characterize for the CCSDs of Cases 1 and 2 the differences between the student's strategy and the optimal strategy in terms of psychometric curve and proportion correct. You have to make your own choices for the parameters. Exploration includes examining whether those choices matter.

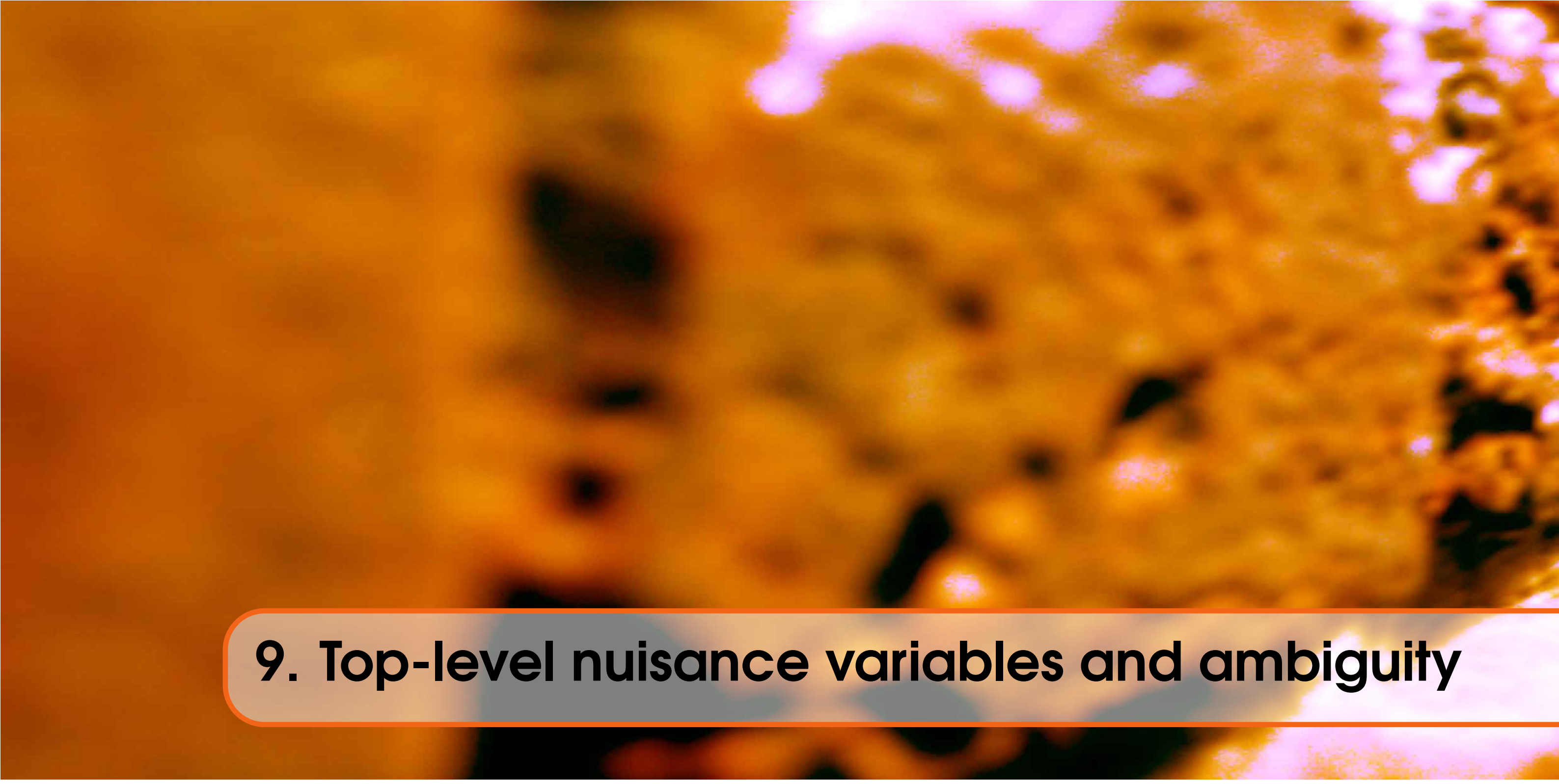
Problem 8.9 This problem is about model mismatch. Experimental psychologists often assume that the observer's beliefs about the CCSDs do not matter for behavior, as long as the CCSDs (whether true or assumed) are mirror images of each other. However, that is only guaranteed if $p(C = 1) = 0.5$. Here, we will show this concretely in Case 3 from Section 8.2. CCSDs are half-Gaussian with $\sigma_s = 1$. For the prior probability of Class 1, consider $p(C = 1) \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. For the sensory noise, consider $\sigma \in \{0.5, 1, 2\}$.

a) Consider a *suboptimal* observer who instead of the MAP decision rule uses the rule "report $C = 1$ when $x > 0$" (which as we know is only optimal if $p(C = 1) = 0.5$). Calculate proportion correct of this suboptimal observer for every combination of $p(C = 1)$ and σ.

b) Repeat for the optimal observer. You may use the expression for the LLR from Eq. (8.43).

c) Plot all results in the same plot of proportion correct as a function of $p(C=1)$, with different lines (in different colors) corresponding to the different values of σ. Use solid lines for suboptimal, dashed lines for optimal.

There is a lesson for experimental design here: if there is any possibility that the observer has a non-flat prior (and usually this possibility exists), then it is good practice to perform simulations to determine whether the assumption that the observer makes about the CCSDs might matter.



9. Top-level nuisance variables and ambiguity

How can we deal with aspects of the world that affect our observations but are not directly related to the question we want to ask?

We have seen that Bayesian inference can model many perceptual tasks, ranging from binary estimation to the estimation of continuous world states, and from tasks involving a single observation to ones involving cue combination. Bayesian inference can model such a wide variety of tasks because it is flexible: the terms entered into Bayes' formula represent the statistical structure of the task at hand, as expressed in the generative model.

In Chapter 8, we encountered the concept of *nuisance variables*, which are world state variables that are not of primary interest to the decision-maker, yet must be taken into account during inference because they affect the observations. In other words, nuisance variables constitute an essential part of the formulas that link the world state variable of interest with the observations.

The nuisance variable showed up in a "chain"-shaped generative model, depicted in **Fig. 8.2A**: the nuisance variable is the intermediate variable s, whereas the variable interest is the class, C. However, there are also nuisance variables that are not intermediate between the variable of interest and observations, but that are themselves top-level variables. The distribution of a top-level variable does not depend on any other variables in the problem. Examples include:

- When inferring the distance to an object, its size is a top-level nuisance variable, because the observation – the retinal size of the object – depends not only on one's distance to it but also on object size.
- Conversely, when inferring the size of an object, one's distance to it is a nuisance variable. This illustrates that what is a nuisance variable depends on the task.
- When inferring the color of a surface, the color of the incident light is a top-level nuisance variable.
- When classifying an object based on a visual image, the viewing angle is a top-level nuisance variable.
- When an airplane pilot tries to infer the approach angle of their plane, the slope of the runway is a top-level nuisance variable.
- When inferring a person's intent from their spoken words, the amount of experience that

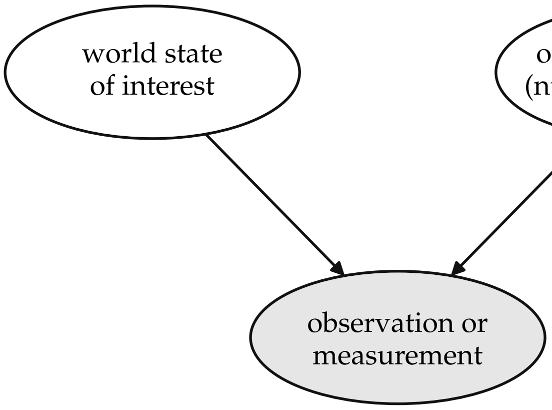


Figure 9.1: Generative model with a top-level nuisance variable. All cases in this chapter have this structure.

the speaker has with the language is a nuisance variable, because people are less adept at expressing important nuances when they are early-stage learners of a language.

In all these examples, two world state variables together give rise to the observation or measurement. Thus, the generative model can graphically be represented as in **Fig. 9.1**. It has a characteristic V-shape.

Top-level nuisance variables cause ambiguity even in the absence of noise, Namely, if the value of the nuisance variable were known, it might be possible to know the value of the variable of interest exactly. However, since the value of the nuisance variable is not known, the same observation is consistent with multiple (and often infinitely many) values of the variable of interest.

Plan of the chapter

We will work through two meaningful examples of top-level nuisance variables: depth perception and color perception. These examples are mathematically similar even though they come from two entirely different domains of perception. We discuss how top-level nuisance variables cause ambiguity, and show how the observer has to marginalize in order to compute the likelihood function over the world state variable of interest.

9.1 Size as a top-level nuisance variable in depth perception

A nuisance variable can introduce ambiguity where none would otherwise be present. A famous example is size-depth ambiguity. Consider a defensive driver who wants to maintain a safe distance between her car and the one ahead (**Fig. 9.2A**). To do so, she must accurately perceive the distance to the next car (distance away, in the three-dimensional world, is known as *depth*). Under good visual conditions, a driver has many cues to aid depth perception. However, when visual conditions are poor, as for instance in darkness or fog, the number of distance cues is diminished. One cue a driver can use provided simply that she can see the taillights of the car ahead, even if with just a single eye, is the size of the image of the car on her retina. The observer's task, then, is to estimate the distance to the next car, D, from the width of the retinal image of that car, x. For simplicity, we assume that the retinal image is flat rather than slightly curved.

Step 1: Generative model (**Fig. 9.2B**). The generative model contains three variables: width w, distance D, and measured width x. In view of the geometry of the problem, retinal width is

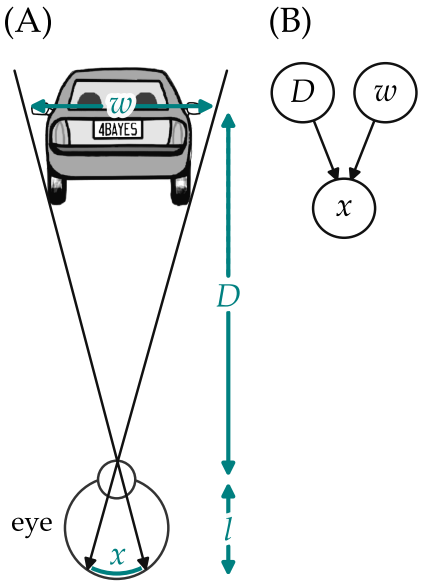


Figure 9.2: Depth perception from retinal size. **(A)** At distance D from the observer, a car of width w produces a retinal image of width x. In the absence of measurement noise, we know from trigonometry that $\frac{x}{l} = \frac{w}{D}$, where l is the distance from the observer's pupil to their retina; a smaller car, closer to the observer, would subtend the same visual angle and produce the same retinal image. The observer can therefore infer the distance to the car only if they have prior beliefs about the car's size. **(B)** The generative model. The nuisance variable, w, and the world state variable of interest, D, are both needed to generate the observation, x. We assume no noise in the observation.

completely determined by w and D,

$$x = \frac{lw}{D}, \tag{9.1}$$

where l is the diameter (length) of the eye (distance from the pupil to the fovea), assumed to be a fixed and known constant. To focus on the ambiguity induced by the top-level nuisance variable, we assume that the measurement is noiseless.

We have specified x in terms of w and D, but we yet have to provide w and D with prior distributions. We first assume that these variables are independent:

$$p(D,w) = p(D)p(w) \tag{9.2}$$

Beyond that, we observe that w and D are restricted to the positive real line: none of them can take negative values. For such positive-valued variables (sometimes called "magnitude variables"), a normal distribution is not appropriate, because that choice would imply that the variable can take negative values. A common correct solution is to assume a *lognormal distribution* instead *(see Section 3.6).* We assume a lognormal distribution over w:

$$p(w) = \text{Lognormal}(w;\mu_w,\sigma_w^2) \tag{9.3}$$

This is equivalent to saying that $\log w$ follows a normal distribution with mean μ_w and variance σ_w^2:

$$p(\log w) = \mathcal{N}(\log w;\mu_w,\sigma_w^2) \tag{9.4}$$

For $\log D$, we assume a flat (improper) prior for simplicity:

$$p(\log D) = \text{constant}. \tag{9.5}$$

It is helpful to rewrite Eq. (9.1) as

$$\log x = \log l + \log w - \log D, \tag{9.6}$$

> **Exercise 9.1** Come up with another example that has the same general computational structure in the generative model that is outside of the depth perception area. ■

Step 2: Inference. The observer infers D from a given measured width x. The ambiguity in this problem consists of the fact that for a given x, there are infinitely many combinations of $\log w$ and $\log D$ that satisfy Eq. (9.1).

To calculate the posterior distribution over $\log D$, we apply Bayes' rule:

$$p(\log D|\log x) \propto p(\log D) p_{\log x|\log D}(\log x|\log D) \tag{9.7}$$
$$\propto p_{\log x|\log D}(\log x|\log D), \tag{9.8}$$

where we have used the assumption from Step 1 that the prior over $\log D$ is flat. We have so far not yet specified $p(\log x|\log D)$. To so so, we start with Eq. (9.6). Since we are conditioning on $\log D$ in $p(x|\log D)$, we treat $\log D$ as a constant. Moreover, $\log l$ is a constant. As a result, $\log x$ is equal to $\log w$ plus a constant. Since $\log w$ has a normal distribution with mean μ_w and variance σ_w^2 (Eq. (9.4), we can use the properties in Box 4.2 to find that $\log x$ conditioned on $\log D$ has a normal distribution with a shifted mean:

$$p(\log x|\log D) = \mathcal{N}(\log x; \mu_w + \log l - \log D, \sigma_w^2). \tag{9.9}$$

Thus, Eq. (9.8) becomes

$$p(\log D|\log x) \propto \mathcal{N}\left(\log x; \mu_w + \log l - \log D, \sigma_w^2\right) \tag{9.10}$$
$$= \mathcal{N}\left(\log D; \mu_w + \log l - \log x, \sigma_w^2\right) \tag{9.11}$$

where we made use of the rule that $\mathcal{N}(a; b, \sigma^2) = \mathcal{N}(a+c; b+c, \sigma^2)$ for any a, b, and c.

> **Exercise 9.2** Let us try to get an intuition for this equation.
> a) Why is this rule true?
> b) Show we applied this rule to obtain Eq. (9.11).
> ■

Eq. (9.11) implies that D itself (without the log) follows a lognormal distribution:

$$p(D|x) = \text{LogNormal}(D; \mu_w + \log l - \log x, \sigma_w^2). \tag{9.12}$$

We have plotted several example posteriors in **Fig. 9.3**.

The last part of Step 2 is to read out the posterior. If the observer minimizes the expected squared error in the log domain (which makes more sense than in the domain of D itself), then they will report the mean of the posterior over $\log D$ which is

$$\log \hat{D} = \mu_w + \log l - \log x. \tag{9.13}$$

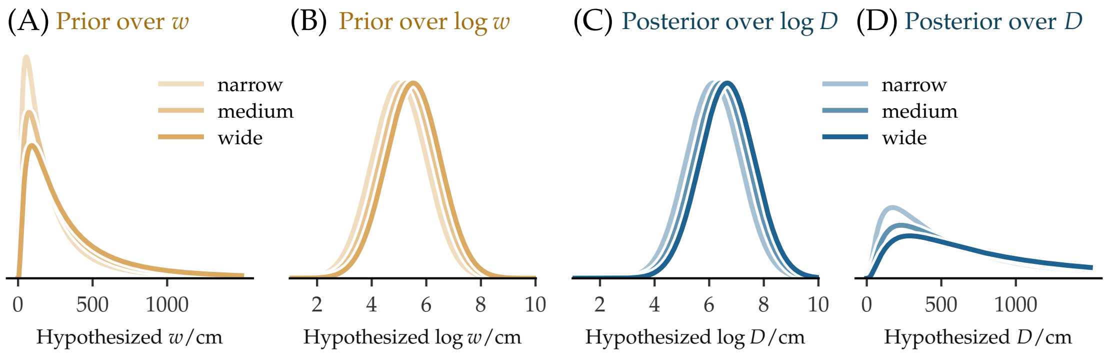


Figure 9.3: Priors and posteriors over w and D in the car example.

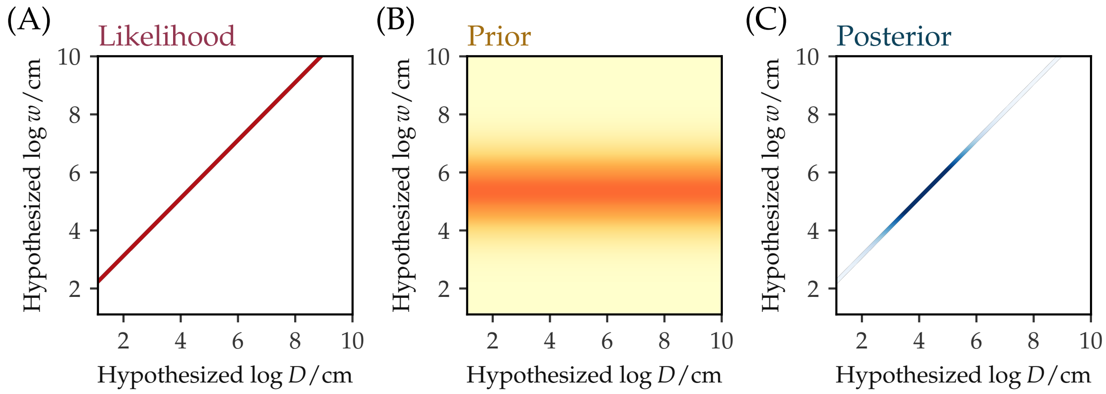


Figure 9.4: Two-dimensional prior, likelihood, and posterior in the car example.

Transforming back to the original space, this means that

$$\hat{D} = \frac{l}{x} e^{\mu_w}. \tag{9.14}$$

This equation implies that the Bayesian observer in this problem uses prior knowledge of the width of the car, through the parameter μ_w, which is the mean of the logarithm of the width. At first, one might think that e^{μ_w} is simply Mean$[w]$. However, this is not the case. The mean of a lognormally distributed variable is not the exponentiated first parameter. Instead, the exponentiated first parameter is the *median* of the variable (see Section 3.6). Thus, e^{μ_w} is the median of w (computed using the prior over w). Finally, since we assumed that the observation is noiseless, $\frac{l}{x}$ is equal to the ratio of true distance to true width, $\frac{D}{w}$, which is known from the experimenter's point of view. As a result,

$$\hat{D} = D \frac{\text{Median}[w]}{w} \tag{9.15}$$

Although this relation is simple, it demonstrates an interesting and deep point: that estimation of the world state of interest (here distance) can be biased by the prior over the nuisance parameter (here width). The bias manifests as follows: when a car is wider than median, then the observer's estimate of distance will be lower than the true distance: you think the car is closer than it truly is, because it is bigger than you expect. Conversely, when a car is narrower than median, then the observer will overestimate its distance. However, the origin of the bias is different than in previous chapters, where it arose when measurement noise was high. Here, there is no measurement noise, but the bias originates from marginalizing over the nuisance parameter.

Beyond the exact equation, the take-home message is that the likelihood over the variable of interest (here D) is "inherited" from a prior over the nuisance variable (here w). This is typical for inference in generative models of the type in **Fig. 9.1**: the prior over the nuisance variable affects the likelihood over the variable of interest, as the two variables are "coupled" through the observation. From the point of view of the generative model, a transformation of variables has occurred from w to x, turning the prior over w into a conditional distribution over x (given D); the latter is then used for the likelihood over D.

We have seen that the prior over the nuisance variable, w, narrows down the posterior over the world state of interest, D: the prior over car width resolves the ambiguity about distance. This prior has to be learned from experience. Most of us have a great deal of familiarity with cars, and so we have considerable knowledge, gained from years of experience, regarding the distribution of the sizes and shapes of cars on the road. Priors over sizes of objects are sometimes referred to as a form of monocular cue about depth (distance).

So far, we have seen how our understanding of an object's size affects our perception of its distance. Conversely, our understanding of an object's distance also affects our perception of its size. **Fig. 9.5** illustrates three examples of the *Ponzo illusion*. Within each panel, the figures have the same size on the page (and therefore on the retina), but the topmost figure appears larger. This occurs because the brain interprets the two-dimensional picture as a three-dimensional scene, where the context suggests that the topmost figure is farther away. The only way one object can have the same retinal image size as another and yet be farther away is if the farther object is larger. Your thumb, viewed at arm's length, may occupy the same retinal area as a distant mountain (for this reason, you can use your thumb to block your view of the mountain). Size-depth ambiguity occurs because objects with an infinite set of possible physical sizes can produce the same retinal image size, depending on their distance from the observer. The more confident we are of the distance to an object, the more confident we can be of its size.

9.2 Marginalization formulation

In Section 8.3, we saw that the Bayesian decision-maker deals with a nuisance variable by considering each possible value for the variables according to their probability, and then averaging over these possibilities. This procedure is called *marginalization*. Marginalization always involves a sum or integral. Marginalization is common in Bayesian models and inevitable in all but the simplest problems. The combination of Bayes' rule and marginalization powers virtually all of Bayesian modeling.

In the previous section, we avoided explicit mention of marginalization. However, marginalization is still taking place in the background. In this section, we explain this; this explanation comes with a helpful two-dimensional visualization. As a first step, we write down a *two-dimensional* posterior distribution over both the variable of interest, D, and the nuisance variable, w:

$$p(\log D, \log w | \log x) \propto p(\log D, \log w) p_{\log x | \log D, \log w}(x | \log D, \log w). \tag{9.16}$$

Using the independence equation, Eq. (9.2), this becomes

$$p(\log D, \log w | \log x) \propto p(\log D) p(\log w) p_{\log x | \log D, \log w}(\log x | \log D, \log w). \tag{9.17}$$

In Eq. (9.17), $p(D)p(w)$ is a two-dimensional prior distribution. An example is shown in **Fig. 9.4A**. But what is the likelihood function over $\log D$ and $\log w$, $p_{\log x | \log D, \log w}(\log x | \log D, \log w)$? We know from Eq. (9.6) that x is a deterministic function of D and w. That means that a particular combination of D and w is either fully compatible with x or not at all; there is no in-between.

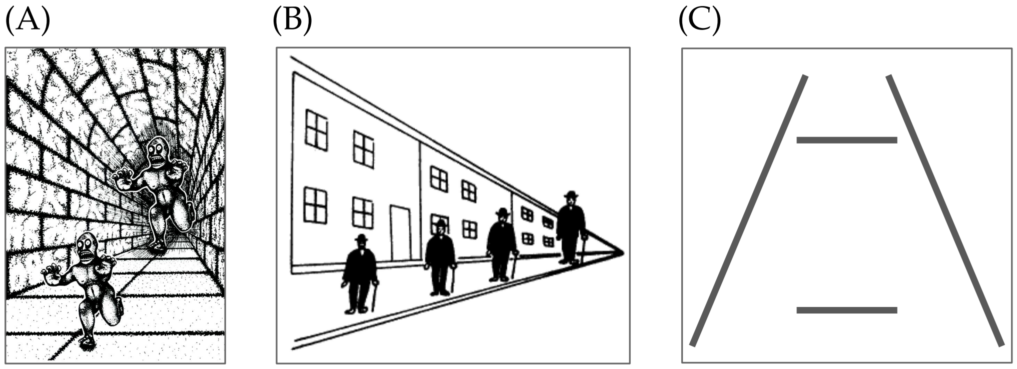


Figure 9.5: Three variants of the Ponzo illusion. In each case, the topmost figure (line segment) looks taller (wider) even though it is physically equally tall/wide in the drawing; *figure inclusion pending permissions*.

Pictorially, this means that the *two-dimensional likelihood function* is a sharp line (**Fig. 9.4B**): all combinations of D and w on this line have a non-zero likelihood (which we can think of as being infinite), and all combinations off the line have zero likelihood. Formally, we can write such a deterministic relationship as a *delta function*:

$$p(\log x|\log D, \log w) = \delta\left(\log x - (\log l + \log D - \log w)\right). \tag{9.18}$$

Multiplying the two-dimensional prior with the two-dimensional likelihood produces a two-dimensional protoposterior, which we can numerically normalize to obtain a two-dimensional posterior (**Fig. 9.4C**). The effect of the sharp likelihood is to take a “slice” out of the two-dimensional prior.

The final step is then to marginalize over the nuisance variable:

$$p(\log D|\log x) = \int p(\log D, \log w|\log x)d(\log w). \tag{9.19}$$

This means “collapsing” the two-dimensional posterior distribution into a one-dimensional distribution over D, by averaging over the second dimension, w. The word “marginalization” in fact refers to “collapsing into the margin”. We will show in Problem 9.8 that the marginalization formulation gives rise to the same posterior over $\log D$ as in Eq. (9.11).

Exercise 9.3 In Eq. 9.19, we integrate over $\log w$. Ask yourself under which circumstances we can do that analytically and under which circumstances we will need to solve this numerically. ■

9.3 In color perception

We now move to an entirely different domain of perception, namely color vision. Although color vision is not typically discussed under the same header as depth perception, we do so in this book because the observer's inference process is nearly identical. We will only sketch the ingredients of this example here, and further work it out in the problems.

We consider surfaces that are a shade of gray (i.e. anywhere in between white and black). We see a surface when there is a light source. A light source emits photons (light particles), each of which carries an amount of energy. The surface absorbs some proportion of photons and reflects the rest. Some of the reflected photons reach our eye and trigger the process of vision.

Step 1: Generative model (see **Fig. 9.6**). The world states are surface shade and illuminant intensity. The *shade* of a surface is the grayscale in which a surface has been painted. It is a form of

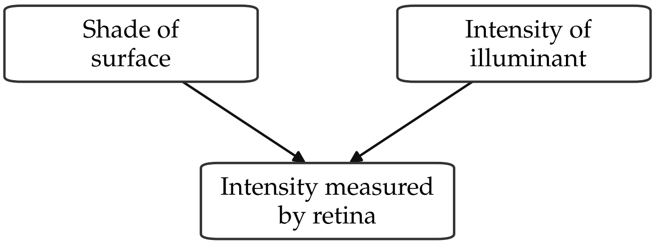


Figure 9.6: Generative model of color perception.

color of a surface. (Note: Shade has nothing to do with shadow.) Technically, shade is reflectance: the proportion of incident light that is reflected. The shade of the paper determines its *reflectance* ρ: black paper might absorb 90% of the incident light and reflect only 10%, while white paper might absorb only 10% and reflect 90%. *The intensity of a light source* (illuminant), denoted by I, is the amount of light it emits; it is measured in "number of photons" or power. The intensity of the light source is a nuisance parameter when inferring the shade of a surface. Both reflectance and intensity will have probability distributions associated with them. We will choose examples in a problem.

The measurement is the amount of light measured by the retina, which we will also refer to as *retinal intensity* and denote by x. Assuming no measurement noise, the retinal intensity is $x = \rho I$. In other words, if you make a surface twice as reflectant, it has the same effect on your retina as doubling the intensity of the light source. Since intensity and reflectance are magnitude variables, we immediately move to the log domain and write

$$\log x = \log \rho + \log I \tag{9.20}$$

Although the variables have entirely different meanings, Eq. (9.20) parallels Eq. (9.6) in the car example. Computation can unify disparate domains!

Step 2: Inference. The observer's inference problem is to infer the surface shade (reflectance) ρ from retinal intensity x. From Step 1, we can understand that retinal intensity provides ambiguous information about the shade of a surface. A particular retinal light intensity is consistent with multiple (in fact infinitely many) possible combinations of true shade and illumination. For example, the same retinal intensity can be produced by dark paper in sunlight and by white paper in dim light. The main challenge of the inference process is to resolve the ambiguity. We use the term *lightness* for the perceived shade of a surface. It is the result of inference, not a physical state of the world.

So far, we have considered grayscale surfaces. However, analogous arguments apply to colored surfaces: the intensity of the illuminant is replaced by the set of intensities at different wavelength, also called the *power spectrum* of the illuminant. The shade of the surface is replaced by the color of the surface, or technically the *reflectance curve*, which specifies what proportion of photons at each wavelength is reflected by the surface. Finally, the retinal intensity is replaced by the power spectrum of the light incident on the retina. The power spectrum of the illuminant is the nuisance parameter when inferring the color of the surface.

Besides the prior, other ways of resolving ambiguity about surface shade/color are to remove the uncertainty about the intensity of the illumination, e.g. by using a light source of known intensity, or – as in the Ponzo illusion – to use contextual cues: other clues in the visual scene that tell you the intensity or power spectrum of the light. Then, the generative model is as in **Fig. 9.7A**. Then, we can infer the intensity of the illumination from the other observations, and using that information, we can compute a posterior distribution over surface shade. This process is called *discounting the illuminant*. For example, if the other observations tell us that the surface is in the shadow, then we

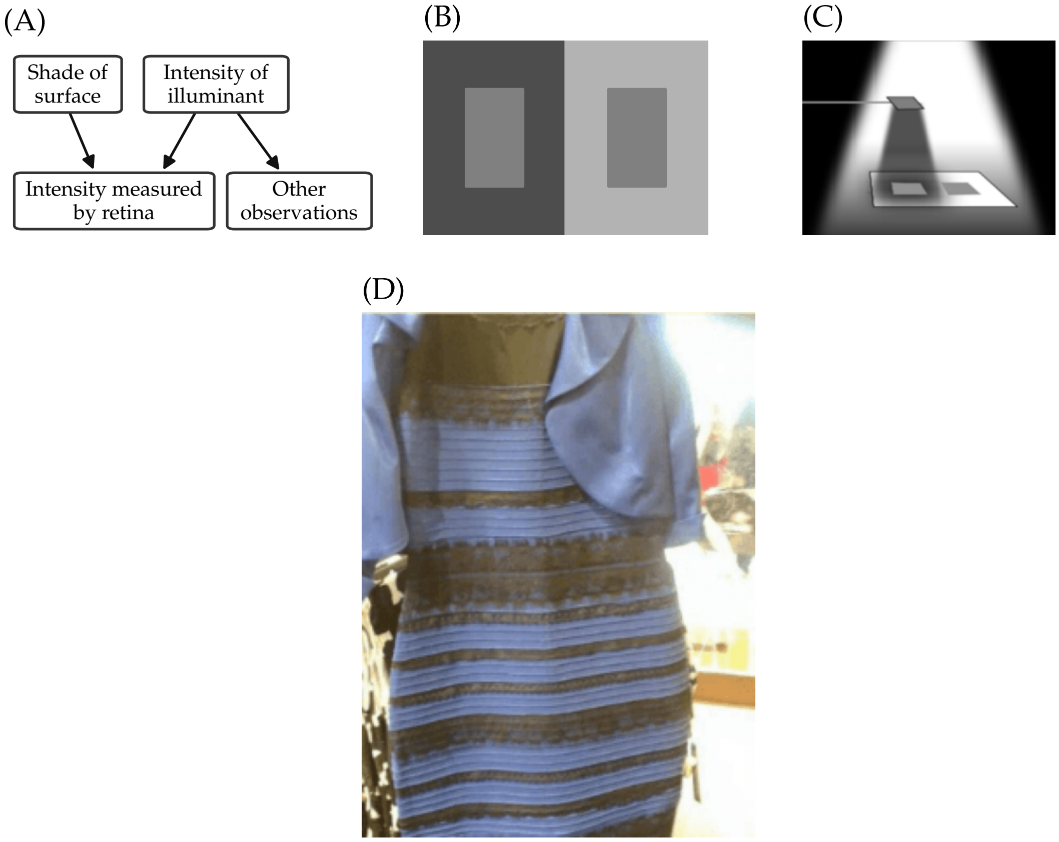


Figure 9.7: Context to help resolve ambiguity. **(A)** Generative model. **(B)** Simultaneous-contrast illusion. **(C)** Edward Adelson's explanation of the simultaneous-contrast illusion. **(D)** The dress. What colors do you perceive? *Figure inclusion pending permissions.*

will believe that the shade of the surface is whiter than when we get the same retinal intensity but the other observations tell us that the surface is in sunlight.

This inference process implies that it is possible to perceive different shades (i.e. report different *lightness*) based on identical retinal intensities, as long as we are made to believe that the illumination differs. One illusion in which this seems to happen is the *simultaneous contrast illusion* (**Fig. 9.7B**). Although both central squares have the same shade, the right one looks darker. Edward Adelson has proposed that this happens because the surrounding big squares suggests different illumination for the two halves of the picture: for example, the right half might be in the light and the left half might be in the shadow (**Fig. 9.7C**). In this explanation, the brain uses other observations (here the surrounding big squares) to infer the illuminations, and based on the result of this inference, proceeds to infer the shades of the small squares.

Similarly, it is possible to perceive different colors based on identical retinal color, as long as we are made to believe that the illumination differs. A famous example of this phenomenon is the dress illusion (**Fig. 9.7D**). Some individuals perceive the stripes to be black and blue, whereas others perceive them to be white and gold. This may reflect inter-individual variability in the brain's assumptions regarding the color (more precisely, the wavelength power spectrum) of the ambient light.

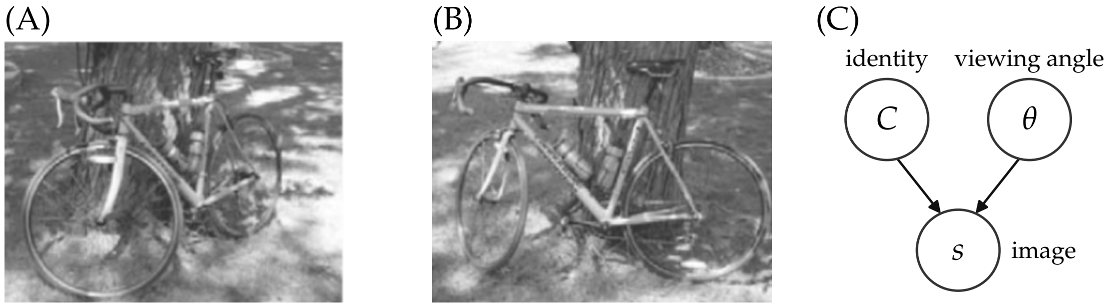


Figure 9.8: Object recognition and nuisance parameters. **(A, B)** The same object when viewed from a different angle produces an image that is pixel by pixel very different. Viewing angle is a nuisance parameter. Picture and example from Kersten and Yuille (2003). **(C)** Generative model of the object recognition task; *figure inclusion pending permissions*.

9.4 In object recognition

Marginalization over nuisance parameters is common in all forms of perception. Here, we consider full-fledged object recognition. Even though this example will not allow us to work out the mathematical model in detail, it is relevant to natural perception. Suppose you want to identify the object in a photograph (**Fig. 9.8A**). You care only about the object's identity, not the angle from which it was photographed. Yet, the camera angle does help determine the image and therefore the visual information received by your retina. In making the identification, your brain has to somehow *discount* camera angle, and infer only the value of the state-of-the-world variable of interest to you, object identity. In other words, your brain needs to realize that you could be viewing the object from any angle, and take into account how each object (e.g., a bicycle, a car, etc) would look from each angle. If you are able to identify the bicycle from any angle, your identification ability is *viewpoint-invariant*.

> Exercise 9.4 Ask yourself how you would build a technical system if you want it to be viewpoint invariant. Why do people working on artificial neural networks not explicitly marginalize? Do a search on the term data augmentation. ■

The generative model of this task is given in **Fig. 9.8B**. Besides a node for object identity, it has a node for each irrelevant variable, such as viewing angle. The observation is in this case the image. We are assuming zero sensory noise; if sensory noise were present, there would be an additional node in the generative model, representing the noisy internal representation of the image. We denote class by C, viewing angle by θ, and the image by s. We assume class and viewing angle are independent; this means that it is not the case that certain objects are photographed more often from particular angles than other objects are. The probability distributions in the generative model are the distribution over class, $p(C)$, the distribution over viewing angle, $p(\theta)$, and the distribution of the stimulus conditioned on both class and viewing angle, $p(s|C,\theta)$. Optimal inference in more complex generative models frequently requires knowledge of multiple distributions over world state variables, and properly incorporating those as priors. The posterior distribution over C is $p(C|s)$. This is obtained by first applying Bayes' rule:

$$p(C|s) \propto p(s|C)p(C) \tag{9.21}$$

and then writing out the class likelihood as a marginalization over θ:

$$p(s|C) = \int p(s|C,\theta)p(\theta|C)d\theta \tag{9.22}$$

$$= \int p(s|C,\theta)p(\theta)d\theta \tag{9.23}$$

In the last equality, we have used the information that C and θ are independent, so that $p(\theta|C) = p(\theta)$. To appreciate the meaning of Eq. (9.23), let's consider a particular class, C: bicycle. The equation states that the probability of the visual image, given the object is a bicycle, is the probability the photographer chose to shoot at an angle 0° relative to the object AND that a bicycle shot at that angle would produce the visual image we are seeing, OR that the photographer chose to shoot at an angle 1° AND that a bicycle shot at that angle would produce the visual image we are seeing, and so on, for all angles. By computing equation Eq. (9.23) for many different classes of objects, C (bicycle, car, person, etc), the observer can in principle generate a class likelihood function and therefore a posterior probability distribution, whose mode is the most probable object identity.

9.5 Summary and remarks

In this chapter, we have introduced top-level nuisance parameters. We have learned:

- In depth perception, the size of an object often appears as a top-level nuisance variable.
- In color perception, the color of the light illuminating an object is a top-level nuisance variable.
- In object perception, the direction of view appears as a top-level nuisance variable.
- In general, most real world problems are affected by countless top-level nuisance variables.
- For all nuisance variables, top-level nuisance variables require the Bayesian observer to consider all values that it might take. The observer does this through marginalization, as an integration over all possible values weighted with the relevant probabilities.
- For top-level nuisance variables in particular, marginalization involves transforming the prior over the nuisance variable into a likelihood function over the variable of interest.
- For magnitude-type variables, one should avoid using normal distribution as it assigns nonzero probability to negative values. A good distribution to use is often the lognormal distribution.
- Marginalization produces integral expressions that are not analytically solvable except in the simplest cases.

9.6 Suggested readings

- David H Brainard and William T Freeman. “Bayesian color constancy”. In: *JOSA A* 14.7 (1997), pages 1393–1411
- Daniel Kersten, Pascal Mamassian, and Alan Yuille. “Object perception as Bayesian inference”. In: *Annu. Rev. Psychol.* 55 (2004), pages 271–304
- David C Knill. “Mixture models and the probabilistic structure of depth cues”. In: *Vision research* 43.7 (2003), pages 831–854
- Rosa Lafer-Sousa, Katherine L Hermann, and Bevil R Conway. “Striking individual differences in color perception uncovered by ‘the dress’ photograph”. In: *Current Biology* 25.13 (2015), R545–R546
- Pascal Wallisch. “Illumination assumptions account for individual differences in the perceptual interpretation of a profoundly ambiguous stimulus in the color domain:“The dress””. In: *Journal of Vision* 17.4 (2017), pages 5–5

9.7 Problems

Problem 9.1 In each of the following forms of behavior, mention one nuisance parameter:

a) (Vision) Estimating the time it will take for an approaching car to reach you.
b) (Vision) Estimating how heavy an object is that you are about to pick up.
c) (Olfaction) Determining whether food has gone bad.
d) (Cognition) Estimating how someone will respond to your criticism.
e) (Cognition) Estimating your capabilities based on your success in a particular task.

Problem 9.2 As you read the text of this book, you ultimately consume words. But the input to your brain is the visual scene you see. What do you marginalize over as you convert the picture of the page of the book to words and meanings?

Problem 9.3 Someone tells you, "I slept only a total of 10 hours in the past two nights." This immediately makes you wonder how much they slept in each of those nights.

a) The variables are the number of nights slept last night and the number of nights slept the night before. Based on the person's statement, what is the two-dimensional likelihood over both variables?
b) What would your two-dimensional prior look like? Explain.
c) As a result, what would the two-dimensional posterior look like?
d) If you got a single guess of how many hours the person slept last night, what would it be? Explain.

Problem 9.4 In Section 9.1, we used a lognormal distribution over the width w (i.e., $p(w) = \text{Lognormal}(w; \mu_w, \sigma_w^2)$). Using 1 meter as the implicit unit in all cases, let's choose $\mu_w = 2$ (cars may be 2 meters wide), $\sigma_w = 0.3$ (most cars are between 1.7 and 2.3), and eye diameter $l = 0.025$ (the eye is roughly 2.5 cm in diameter).

a) Consider three retinal observations of width, $x = 1.6$, $x = 2$, and $x = 2.4$. For each retinal width, plot the posterior probability density over distance (single plot, three curves, color-coded).
b) Discuss how these posteriors compare to each other, and why so.

Problem 9.5 In Section 9.1, we discussed an observer who inferred car distance with car width being a nuisance parameter. Now consider the opposite: an observer who infers car width w, with car distance D being a nuisance parameter. Assume a flat distribution over log w, and a normal distribution over $\log D$ with mean $\mu_{\log D}$ and standard deviation $\sigma_{\log D}$.

a) Step 2: Derive an expression for the estimate of w in terms of x.
b) Step 3: Derive an expression for the estimate of w in terms of true w and true distance D.
c) Interpret this expression in a way analogous to our interpretation in Section 9.1.

Problem 9.6 In Section 9.1, we assumed a flat prior over $\log D$. We will now consider a more realistic extension in which this prior is a normal distribution,

$$p(\log D) = \mathcal{N}(\log D; \mu_D, \sigma_D^2) \tag{9.24}$$

We already know that this means that D is lognormally distributed:

$$p(D) = \text{Lognormal}(D; \mu_D, \sigma_D^2) \tag{9.25}$$

a) Show that the posterior over $\log D$, from Eq. (9.11), now becomes

$$p(\log D | \log x) = \mathcal{N}(\log D; \mu_{\text{post}}, \sigma_{\text{post}}^2), \tag{9.26}$$

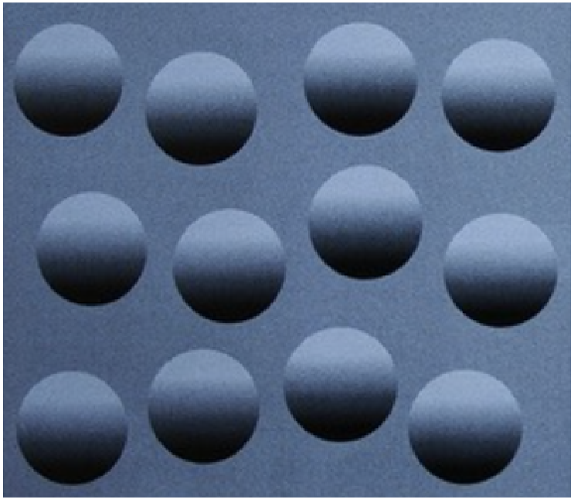
where

$$
\begin{aligned}
J_D &\equiv \frac{1}{\sigma_D^2} \\
J_w &\equiv \frac{1}{\sigma_w^2} \\
\mu_{\text{post}} &= \frac{J_D\mu_D + J_w(\mu_w + \log l - \log x)}{J_D + J_w} \\
\sigma_{\text{post}}^2 &= \frac{1}{J_D + J_w}.
\end{aligned}
$$

b) Show that the estimate of D, from Eq. (9.15), now becomes

$$\hat{D} = \text{Median}[D]^{\frac{J_D}{J_D+J_w}} \left(D \frac{\text{Median}[w]}{w} \right)^{\frac{J_w}{J_D+J_w}} \tag{9.27}$$

Problem 9.7 In the following image, you probably see protruding (convex) half-spheres; however, if you flip the image upside down, you see hollow (concave) half-spheres. This can be explained using a prior favoring light coming from above.



Problem 9.8 By combining Eqs. (9.18) and (9.19), find the posterior over D. The answer should be Eq. (9.11).

a) Draw the generative model. For the observation/measurement, simply write “image” (I).

b) Write down the log posterior ratio for convex relative to concave given I, and apply Bayes' rule. Assume equal priors.

c) Evaluate the log likelihood ratio by marginalizing over light direction. Assume for simplicity that light can only come from above or below. This should reduce the marginalization of each likelihood to two terms, for a total of four terms in the log likelihood ratio.

d) Each of the four terms contains a likelihood of the form $p(I|\ldots)$. Which two of these four are nearly zero, and why?

e) Simplify the log likelihood ratio accordingly.

f) How does the simplified expression explain the percept described above?

Problem 9.9 We consider the problem of estimating surface reflectance ρ when light intensity I is unknown. We start from Eq. for the log of the retinal intensity, $\log x = \log \rho + \log I$. We assume that the observer has a prior over light intensity, which can be described by a lognormal distribution with parameters $\mu_{\log I}$ and $\sigma_{\log I}$.

a) The distribution of $\log x$ given $\log \rho$ is normal with mean $\log \rho + \mu_{\log I}$ and standard deviation $\sigma_{\log I}$. Explain why.

b) Show that the likelihood of ρ based on a measured retinal intensity x is proportional to $e^{-\frac{\log\rho-\log x+\mu_I)^2}{2\sigma_I^2}}$.

The surface reflectance ρ is a proportion and therefore a number between 0 and 1. Therefore, we assume a prior over ρ that is uniform between 0 and 1, and 0 elsewhere.

c) Suppose $x = 10$, $\mu_{\log I} = 3$, and $\sigma_{\log I} = 1$. Choose a grid for ρ from 0 to 1 in steps of 0.001. Plot the posterior mass function over ρ, making sure that the distribution is normalized.

d) Numerically compute the posterior mean of ρ.

e) Vary $\mu_{\log I}$ from 1 to 5 in steps of 0.1. For each value of $\mu_{\log I}$, repeat part (d). Plot the posterior mean of ρ as a function of $\mu_{\log I}$.

f) Interpret the plot and relate it to the simultaneous-contrast illusion.

10. Same-different judgment

How do we tell whether two stimuli are the same or different?

In this chapter, we consider for the first time how observers infer the relationships between objects. We focus on a fundamental and very useful aspect of relatedness judgement: sameness-difference judgment. To understand the importance of such judgments, consider that, in order to accurately segment a visual scene, an observer can rely on knowledge that an individual object will more often have the same color or orientation, whereas different objects typically differ in these features. While judging sameness sounds simple, it often requires considerable calculation. When two stimuli are affected by noise, they may seem the same despite being different or different despite being the same. Furthermore, the observer does not know the actual stimulus that occurred. Hence, marginalization is often necessary. Examples of same-different judgments include:

- You are a prehistoric hunter-gatherer. You encounter a berry bush. You want to determine whether its berries are the same as those that you normally eat, which you know to be non-toxic.
- A generalization of cue combination (Chapter 5): two cues do not necessarily have the same source, as we assumed in Chapter 5. They could instead have two different sources. You want to know whether they are from the same source or not. The inference in this problem, within perception and sensorimotor research is also called *causal inference*.
- *Contour integration*: do two line segments belong to the same continuous contour, or are they independent? Many other forms of perceptual organization can be framed as a similar inference: do elements belong together in a coherent whole, or not?
- *Change detection*: are two images, separated in time, identical or not? A related working memory task is *delayed match-to-sample*.
- At a more cognitive level, judging whether two quantities are the same underlies judgments of fairness and also forms the basis of mathematics.

Plan of the chapter

We first discuss same-different judgment using stimuli that can only take two values (analogous to Chapter 7). We then consider stimuli that are drawn from continuous distributions (analogous to Chapter 8). Thus, this chapter will synthesize material from multiple previous chapters, but with

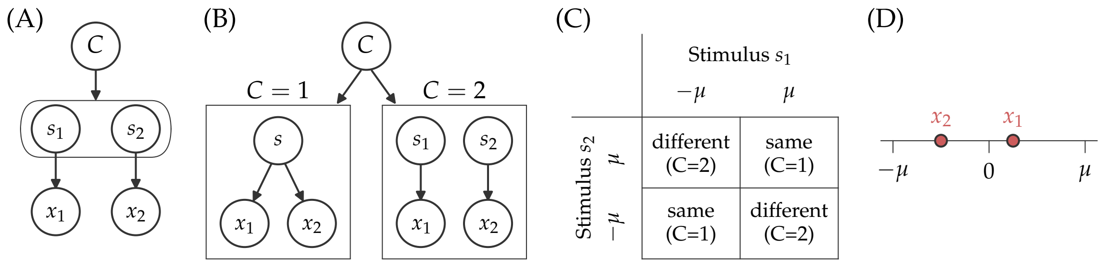


Figure 10.1: Same-different judgment. **(A)** Graphical depiction of the generative model. **(B)** Equivalent depiction in which the $C = 1$ and $C = 2$ scenarios are made explicit. **(C)** Desired classification table. The first stimulus s_1 and the second stimulus s_2 can each take on the values of μ and $-\mu$ giving us four possible situations. **(D)** Example trial. The possible stimulus values are $-\mu$ and μ, and the measurements on this trial are x_1 and x_2.

a new twist: central is the *relation between two stimuli*, rather than the identity or category of a single stimulus.

10.1 Binary stimuli

The inference of this chapter involves determining whether two stimuli, which we will denote by s_1 and s_2, are the same or different. In this section, we consider stimuli that can only take two values, as in Chapter 7. We choose the values to be $-\mu$ and μ. We will now ask how an observer can estimate if the stimuli are the same or different.

10.1.1 Step 1: Generative model

The generative model is shown in two equivalent ways in **Fig. 10.1A-B**. The top-level variable is a binary variable C, which is 1 for "same" and 2 for "different"; the two depictions differ by whether or not the two possibilities for C are explicitly distinguished. For the distribution of C, we write

$$p(C=1) = p_{\text{same}} \tag{10.1}$$
$$p(C=2) = 1 - p_{\text{same}} \tag{10.2}$$

Since s_1 and s_2 each can be only equal to $-\mu$ and μ, there are four possible combinations of s_1 and s_2. The class-conditioned stimulus distributions can simply be specified explicitly for these four combinations (**Fig. 10.1C**). On "same" trials:

$$p(s_1 = -\mu, s_2 = -\mu) = 0.5 \tag{10.3}$$
$$p(s_1 = -\mu, s_2 = \mu) = 0 \tag{10.4}$$
$$p(s_1 = \mu, s_2 = -\mu) = 0 \tag{10.5}$$
$$p(s_1 = \mu, s_2 = \mu) = 0.5 \tag{10.6}$$

And on "different" trials:

$$p(s_1 = -\mu, s_2 = -\mu) = 0 \tag{10.7}$$
$$p(s_1 = -\mu, s_2 = \mu) = 0.5 \tag{10.8}$$
$$p(s_1 = \mu, s_2 = -\mu) = 0.5 \tag{10.9}$$
$$p(s_1 = \mu, s_2 = \mu) = 0 \tag{10.10}$$

Finally, we assume that the measurements x_1 and x_2 are noisy versions of the stimuli s_1 and s_2, with the noise being independent between the measurements, and normally distributed with standard deviation σ:

$$p(x_1,x_2|s_1,s_2) = p(x_1|s_1)p(x_2|s_2) \tag{10.11}$$

$$p(x_1|s_1) = \mathcal{N}(x_1;s_1,\sigma_1^2) \tag{10.12}$$

$$p(x_2|s_2) = \mathcal{N}(x_2;s_2,\sigma_2^2) \tag{10.13}$$

This concludes Step 1.

Exercise 10.1 Can you think of a very different problem which has a similar generative model? ■

10.1.2 Step 2: Inference

An example trial is shown in **Fig. 10.1D**: the observer measures x_1 and x_2, and has to determine whether they both came from the same stimulus. If they did, there are still two possibilities: that stimulus was $-\mu$, or it was μ. There are also two possibilities if the measurements were produced by different stimuli: x_1 could have come from $-\mu$ and x_2 from μ, or the other way around.

As in previous classification tasks (Chapters 7 and 8), the world state variable of interest is a high-level categorical variable, C, rather than a physical stimulus. The posterior distribution over C is $p(C|x_1,x_2)$. Since C is binary, we consider for convenience the log posterior ratio, denoted by d, which is the sum of the log likelihood ratio and the log prior ratio:

$$d = \log\frac{p(C=1)|x_1,x_2)}{p(C=2|x_1,x_2)} \tag{10.14}$$

$$= \log\frac{p(x_1,x_2|C=1}{p(x_1,x_2|C=2)} + \log\frac{p(C=1)}{p(C=2)} \tag{10.15}$$

We evaluate the likelihood of "same" and "different", $p(x_1,x_2|C=1)$ and $p(x_1,x_2|C=2)$, by marginalizing over s_1 and s_2 (see Chapter 8 if you need a reminder why). Since together, these variables van take only four combinations of values, the marginalization takes the form of a sum:

$$d = \log\frac{\sum_{s_1,s_2} p(x_1,x_2|s_1,s_2)p(s_1,s_2|C=1)}{\sum_{s_1,s_2} p(x_1,x_2|s_1,s_2)p(s_1,s_2|C=2)} + \log\frac{p(C=1)}{p(C=2)} \tag{10.16}$$

Now, we can substitute what we know from the generative model, Eqs. 10.1 through 10.10. We do not yet substitute the normal distributions. This gives

$$d = \log\frac{p(x_1|s_1=-\mu)p(x_2|s_2=-\mu)+p(x_1|s_1=\mu)p(x_2|s_2=\mu)}{p(x_1|s_1=-\mu)p(x_2|s_2=\mu)+p(x_1|s_1=-\mu)p(x_2|s_2=\mu)} + \log\frac{p_{\text{same}}}{1-p_{\text{same}}} \tag{10.17}$$

Although this expression is long, it is intuitive: the large numerator and the large denominator correspond directly to the four possibilities described at the beginning of this section. Their ratio is the sensory evidence that the measurements came from the same stimulus, relative to that they came from different stimuli. Finally, substituting the normal distributions from Eq. 10.13, we find

$$d = \log\frac{e^{-\frac{\mu}{\sigma^2}(x_1+x_2)}+e^{\frac{\mu}{\sigma^2}(x_1+x_2)}}{e^{-\frac{\mu}{\sigma^2}(x_1-x_2)}+e^{\frac{\mu}{\sigma^2}(x_1-x_2)}} + \log\frac{p_{\text{same}}}{1-p_{\text{same}}} \tag{10.18}$$

(see Problem). The Bayesian MAP decision rule is to report "same" ($\hat{C}=1$) when $d>0$. In general, this rule cannot be further simplified. However, when $p_{\text{same}}=0.5$, the inequality $d>0$ takes an extraordinarily simple form,

$$\text{sign}(x_1)=\text{sign}(x_2) \tag{10.19}$$

(see Problem 10.6). In other words, the observer reports that the two measurements come from the same stimulus if they have the same sign. More generally, the decision rule is hard to interpret but can still be visualized, which we will do in a problem 10.4.

10.1.3 Step 3: Estimate distribution

We would like to compute the probability for the Baysian MAP observer to report "same" ($\hat{C}=1$) for given true stimuli: $p(\hat{C}=1|s_1,s_2)$. After computing the observer's response probabilities as predicted by the model, the experimenter can compare them with empirical data. This estimate distribution can be written as a marginalization over the measurements:

$$p(\hat{C}=1|s_1,s_2)=\iint p(\hat{C}=1|x_1,x_2)p(x_1,x_2|s_1,s_2)dx_1dx_2, \tag{10.20}$$

or, since the mapping from x_1 and x_2 to $\hat{C}$ is deterministic, as

$$\begin{aligned} p(\hat{C}=1|s_1,s_2) &= \iint_{d(x_1,x_2)>0} p(x_1,x_2|s_1,s_2)dx_1dx_2 & (10.21)\\ &= \Pr(d(x_1,x_2)>0|s_1,s_2), & (10.22)\end{aligned}$$

the probability that the decision variable d, which is a function of x_1 and x_2, is positive, when x_1 and x_2 follow their respective distributions conditioned on s_1 and s_2. Geometrically, this probability is the volume under the two-dimensional measurement distribution in the region defined by the condition $d(x_1,x_2)>0$. Since we know the measurement distribution from Eq. 10.13, we can write Eq. (10.22) as

$$p(\hat{C}=1|s_1,s_2)=\iint_{d(x_1,x_2)>0} \mathcal{N}(x_1;s_1,\sigma_1^2)\mathcal{N}(x_2;s_2,\sigma_2^2)dx_1dx_2 \tag{10.23}$$

In the case of $p_{\text{same}}=0.5$, when $d(x_1,x_2)>0$ reduces to Eq. (10.19), we can make some progress: when the stimuli are s_1 and s_2, what is the probability that the generated measurements have the same sign? We will do this in a problem.

In general however, because of the specific form of the decision variable in Eq. (10.18), we cannot make any further analytical progress. This is a new situation, since in all Bayesian models discussed so far in this book, the estimate distribution could be calculated analytically or expressed in terms of a standard non-elementary function (the cumulative standard normal). In the present situation – and in fact in most Bayesian models – we instead have to resort to numerical methods. There are two general classes of numerical methods for computing integrals such as:

Numerical integration. The simplest technique is Riemann integration: we discretize the continuous variable (here x_1 and x_2) in steps of a certain size. We then approximate the integral as a sum multiplied by that step size. In the sum, we only retain the terms that satisfy the condition (here $d(x_1,x_2)>0$). Care should be taken to use a sufficiently fine grid. An improved version of Riemannian integration is the trapezoidal method. Both methods are only practical when the number of dimensions in the integral (here 2) is relatively low (typically fewer than 5).

Monte Carlo simulation. The second technique is simulation. For example, in Eq. (10.23), we can randomly draw a large number of pairs of measurements from their respective stimulus-conditioned distributions (the normal distributions inside the integral). Each pair represents a

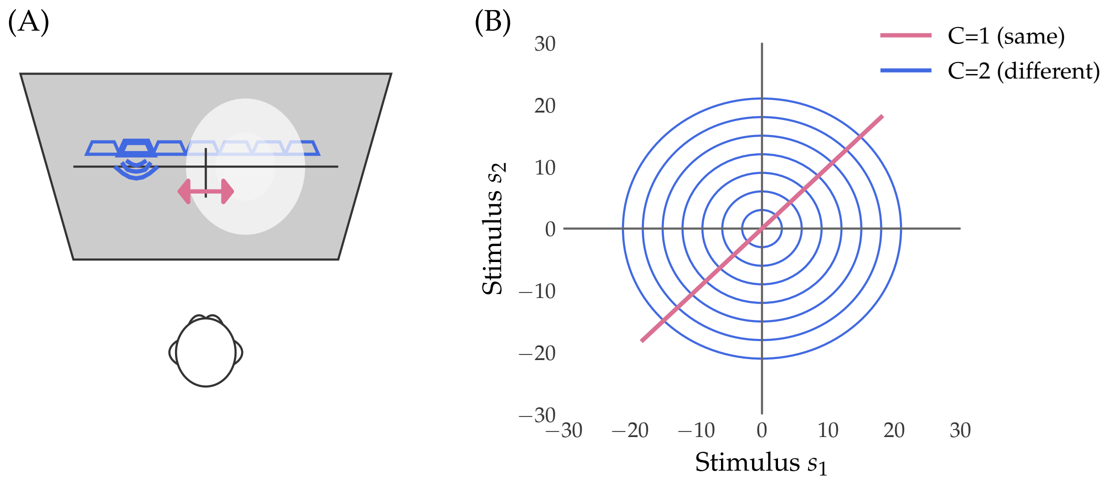


Figure 10.2: Cue combination experiment. **(A)** Sounds are presented from speakers mounted in a horizontal row behind a screen on which visual flashes are projected. The visual stimulus is a spot of light with its center on the same horizontal line as the speakers. Both auditory and visual stimuli are very brief. Visual reliability is manipulated through the size of the spot. Trials are either unisensory (auditory or visual) or multisensory. On multisensory trials, a visual and an auditory stimulus are presented simultaneously. In different blocks, the observer either localizes the auditory stimulus using a cursor on the horizontal meridian (vertical black line), or reports whether the visual and auditory stimuli shared the same location (using a key press). **(B)** Class-conditioned stimulus distributions. "Same" trials are represented by a one-dimensional Gaussian distribution on the diagonal. "Different" trials are represented by a two-dimensional Gaussian distribution.

simulated trial. For each simulated trial, we apply the decision rule to determine the simulated observer's response. The proportion of the simulated trials for which the response is "same" is an approximation of the underlying probability of a "same" response given the stimuli. This technique, of approximating a probability distribution by its samples, is a specific case of a method called *Monte Carlo simulation*. In a sense, a Monte Carlo simulation creates an "empirical" distribution using a computer subject.

Exercise 10.2 Thinking about the previous chapters, what kinds of assumptions would have prevented us from analytically solving the relevant equations? ■

10.2 Continuous stimuli

In standard cue combination (Chapter 5), the observer has two measurements and knows that they were generated by the same stimulus, it is unclear whether the measurements came from the same stimuli or from different stimuli. In this case, the observer has to infer the probability that there was a single stimulus from the measurements. This probability will subsequently play a role in estimating the values of the stimuli. This inference is completely analogous to the same-different judgment in Section 10.1, but with continuous variables; in the context of cue combination, the problem is also been called *causal inference*, whereas the special case in Chapter 5 is called *forced fusion*.

In a laboratory experiment, we might simultaneously present an auditory tone and a visual flash (**Fig. 10.2A**). Intuitively, when the location of the tone and the location of the flash are close to one another, observers may conclude they come from the same source; when the two stimuli are farther apart, the conclusion may be that they come from different sources. Using this as an example, we will see how a Bayesian model will come to the same conclusion. The causal inference model was developed in 2007 by two of the authors of this book and collaborators (Kording, Beierholm et al. 2007) as well as simultaneously by an independent group (Sato, Toyoizumi et al. 2007).

10.2.1 Step 1: Generative model

The generative model is the same as in **Fig. 10.1A-B**. We again start with the "sameness" variable, C. If $C = 1$, then there is only a single stimulus s. If $C = 2$, then there are two stimuli, s_1 and s_2, which we assume to be drawn independently:

$$p(s_1, s_2|C = 2) = p(s_1|C = 2)p(s_2|C = 2). \tag{10.24}$$

We assume that all three stimulus variables, s, s_1, and s_2, follow the same normal distribution with mean 0 and standard deviation σ_s:

$$p(s|C = 1) = \mathcal{N}(s; 0, \sigma_s^2) \tag{10.25}$$

$$p(s_1|C = 2) = \mathcal{N}(s_1; 0, \sigma_s^2) \tag{10.26}$$

$$p(s_2|C = 2) = \mathcal{N}(s_2; 0, \sigma_s^2) \tag{10.27}$$

A visualization of the CCSDs $p(s|C = 1)$ and $p(s_1, s_2|C = 2)$ (**Fig. 10.2B**) makes clear that this inference problem is conceptually similar to nested classification in Chapter 8: one narrow category ($C = 1$) that is "embedded" inside a broad category ($C = 2$), except now in two dimensions.

We denote the two measurements by x_1 and x_2. As always, we model them as conditionally independent and normally distributed, but we allow for them to have different levels of noise, σ_1 and σ_2, as in Eq. (10.13):

$$p(x_1, x_2|s_1, s_2) = p(x_1|s_1)p(x_2|s_2) \tag{10.28}$$

$$p(x_1|s_1) = \mathcal{N}(x_1; s_1, \sigma_1^2) \tag{10.29}$$

$$p(x_2|s_2) = \mathcal{N}(x_2; s_2, \sigma_2^2) \tag{10.30}$$

This concludes Step 1.

10.2.2 Step 2: Inference

The observer infers whether the two measurements come from the same stimulus ($C = 1$) or from different stimuli ($C = 2$). Thus, the posterior of interest is $p(C|x_1, x_2)$, the probability over C given the measurements x_1 and x_2. The log posterior ratio d is given by Eq. (10.15). As in Chapter 8 and in Section 10.1, we evaluate the likelihoods over C, $p(x_1, x_2|C)$, by marginalizing over the stimulus or stimuli. The likelihood of "same" is

$$\mathcal{L}(C = 1) = p(x_1, x_2|C = 1) \tag{10.31}$$

$$= \int p(x_1|s)p(x_2|s)p(s|C = 1)ds \tag{10.32}$$

The likelihood of "different" is

$$\mathcal{L}(C = 2) = p(x_1, x_2|C = 2) \tag{10.33}$$

$$= \left(\int p(x_1|s_1)p(s_1|C = 1)ds_1\right)\left(\int p(x_2|s)p(s|C = 2)ds_2\right) \tag{10.34}$$

Plugging in the distributions from the generative model, the log posterior ratio is equal to

$$d = \log\frac{p_{\text{same}}}{1 - p_{\text{same}}} + \frac{1}{2}\log\left(1 + \frac{J_1 J_2}{J_s(J_1 + J_2 + J_s)},\right) \tag{10.35}$$

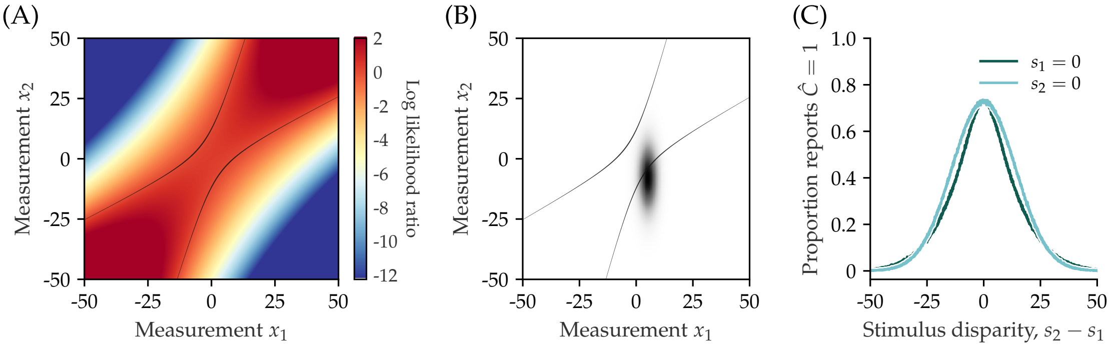


Figure 10.3: **(A)** The strength of the evidence in favor of a common cause, as expressed by the log likelihood ratio, as a function of the measurements x_1 and x_2. The $d = 0$ contour lines are shown in black. Two aspects of interest are the band around the diagonal and the structure within this band. Parameters were $p_{\text{same}} = 0.5$, $\sigma_1 = 3$, $\sigma_2 = 10$, $\sigma_s = 10$. **(B)** Same decision boundaries as in **(A)** but with overlaid the two-dimensional measurement distribution when $s_1 = 5$ and $s_2 = -8$. The volume under this distribution in between the two decision boundaries is equal to the probability of reporting "right" given the specific stimuli. **(C)** Proportion reports "same" as a function of stimulus disparity (the difference between s_2 and s_1), when $s_1 = 0$ (so disparity $= s_2$) or when $s_2 = 0$ (so disparity $= s_1$).

where we introduced the notation $J_1 \equiv \frac{1}{\sigma_1^2}$, $J_2 \equiv \frac{1}{\sigma_2^2}$, and $J_s = \frac{1}{\sigma_s^2}$. The final part of Step 2 is for the observer to report "same" when

$$d > 0 \tag{10.36}$$

When faced with a complicated expression such as Eq. (10.35), it is always useful to try plotting and interpreting it. This is useful both to detect mistakes and to gain an intuition for the equation. In **Fig. 10.3A**, we plotted a heat map of the log posterior ratio d against the observations, x_1 and x_2, for a particular parameter combination. The black contours indicate $d = 0$. The diagonal corresponds to trials on which the measurements happen to be identical to each other. The hypothesis $C = 1$ becomes more likely relative to $C = 2$ the closer to the diagonal a set of measurements lies. This is intuitive: when two measurements are similar, they are likely to have come from the same stimulus. *It would be too coincidental for the two to come from different sources.* This is the same logic that we used in Chapter 2 for the simultaneously moving dots. Formally, the current problem is more complex, but the essence is the same. Note that the prior does not play any role in this argument.

Moreover, the further from 0 such a pair of measurements lies, the more likely the stimuli were the same. This is because we chose a prior that peaks at the origin. Since stimuli are drawn from this prior, even when the causes are different, the two stimuli and therefore the two measurements tend to lie close to each other near 0. When the measurements lie close to each other but far from 0, this is harder to explain away as a consequence of the prior, and it is therefore more likely that the stimuli were the same.

10.2.3 Step 3: Response probabilities

The goal is to compute the probability that the observer will report "same" for a given combination of stimulus values (s_1, s_2), or in other words, $p(\hat{C} = 1|s_1, s_2)$. Analogous to Section 10.1.3, this probability is equal to the area under the two-dimensional measurement distribution given those stimulus values that lies in between the two boundaries (**Fig. 10.3B**). There are three methods:

- **Method 1: Analytical derivation.** This is always preferred if it is possible. Here, it is not. Therefore, we have to use numerical methods to calculate the response probabilities.

- **Method 2: Numerical integration on a grid.** Under this method, we first define a two-dimensional measurement distribution $p(x_1, x_2|s_1, s_2)$ on a fine grid. We then numerically normalize this distribution on the grid. We then calculate d for all values on the grid. Finally, we calculate the total probability mass under the measurement distribution for which $d > 0$.
- **Method 3: Monte Carlo simulation.** The third method is Monte Carlo simulation. This means that we randomly draw a very large number (e.g. 1 million) measurement pairs (x_1, x_2) from the same stimulus combination of interest s_1, s_2. For each pair, we count how often the inequality $d > 0$ is satisfied, i.e. the observer reports "same". As a proportion of the total number of measurement pairs drawn, this is the probability we are looking for.

Exercise 10.3 Can you see how Monte Carlo simulation is really a way of calculating an integral? ■

Continuing our case study, we apply any of these methods to stimulus pairs s_1, s_2 where s_1 is equal to 0 and s_2 is varied, with $p_{\text{same}} = 0.5$. The resulting probability of reporting that both cues came from the same position ($\hat{C} = 1$) is plotted as a function of stimulus disparity in **Fig. 10.3C**. We see that the larger the spatial disparity between the two stimuli, the less frequently the observer reports that there is a common cause. This makes sense!

Exercise 10.4 Stimulus disparity is $s_2 - s_1$, so there are many combinations of s_1 and s_2 that produce the same disparity. In **Fig. 10.3C**, we remove this ambiguity by either setting s_1 or s_2 to 0. These choices produce slightly different curves. Why? ■

When to use which method? Whenever an analytical approach is possible, use it. When it is not possible, your decision depends on your computational time available. In a given amount of time, you can only do a finite number of evaluations of the decision rule, so you have to spend these wisely. Method 2 is inefficient (it spends many evaluations in low-probability regions of measurement space) but Method 3 is stochastic (so your result might vary from run to run). To illustrate this, if you have 2 measurements and define a 1000 by 1000 grid in Method 2, you need 1 million evaluations. In Method 3, 1 million evaluations would also give an accurate estimate of the desired probability. However, if you had 3 measurements, Method 2 would take 1000 times as long (1 billion evaluations) and might be infeasible in practice (since the response probabilities need to be computed for many stimulus pairs and parameter combinations). If you had a budget of only 1 million evaluations, Method 3 would usually be preferable over Method 2 with a coarse grid of 100 by 100 by 100.

10.2.4 Step 2 revisited: Inferring the stimuli

So far we have discussed inference of the number of causes. One might also be interested in the posterior distribution over s_1 and s_2, the stimulus values. This posterior can be written as

$$p(s_1, s_2|x_1, x_2) \propto p(x_1|s_1)p(x_2|s_2)p(s_1, s_2) \tag{10.37}$$

So far, nothing special; we could have done this in Chapter 5. However, in the current generative model, the prior $p(s_1, s_2)$ is not directly available. All we know is the distribution of s_1 and s_2 conditioned on the number of causes, C. To find the "prior" $p(s_1, s_2)$, we marginalize over C:

$$p(s_1, s_2) = \sum_{C=1}^{2} p(s_1, s_2|C)p(C). \tag{10.38}$$

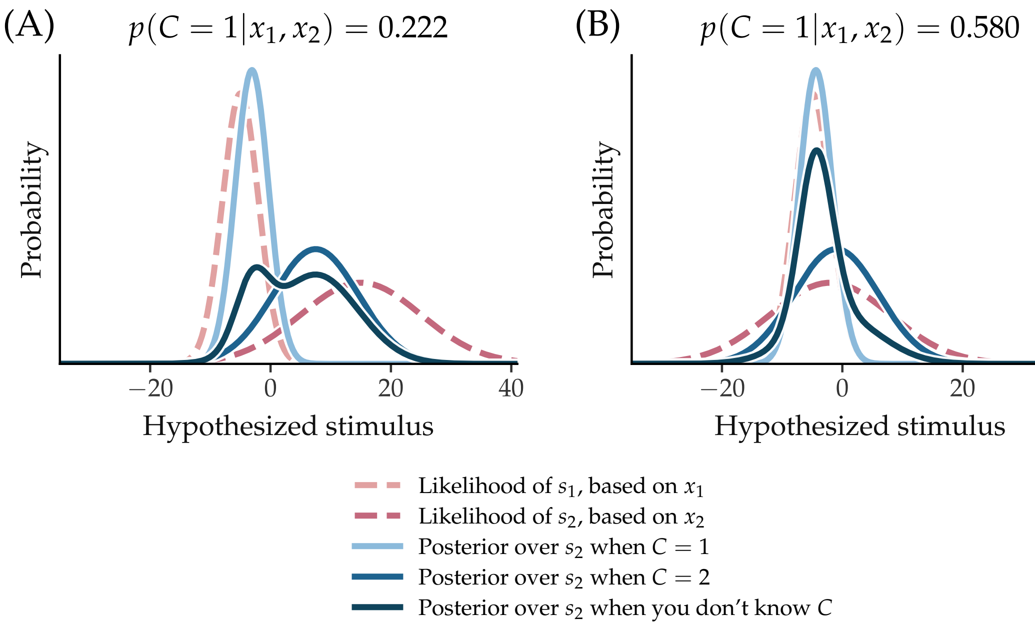


Figure 10.4: Posteriors and conditional posteriors in causal inference. **(A)** In causal inference, the posterior distribution over a stimulus can be bimodal (two-peaked). We are considering a trial on which $x_1=-5$ and $x_2=15$. Shown are the likelihoods over s_1 and s_2m the posterior over s_2 if C were known to be 1, the posterior over s_2 if C were known to be 2, and the posterior over s_2 when there exists uncertainty about C. The latter posterior is a weighted average of the conditioned posteriors, with the weights being the posterior probabilities of $C=1$ and $C=2$, respectively. Here the posterior probability of $C=1$ is only 22.2%, so the overall posterior is dominated by the posterior conditioned on $C=2$. **(B)** Same but with $x_1=-5$ and $x_2=-2$. Now, the posterior probability of $C=1$ is higher, so the overall posterior is more of an equally weighted average of the two conditional posteriors.

Substituting in Eq. (10.37), we find

$$p(s_1,s_2|x_1,x_2) \propto p(x_1|s_1)p(x_2|s_2)\sum_{C=1}^{2} p(s_1,s_2|C)p(C) \tag{10.39}$$

$$= \sum_{C=1}^{2} p(C)p(s_1,s_2|C)p(x_1|s_1)p(x_2|s_2) \tag{10.40}$$

Thus, this posterior before normalization is a weighted average of the likelihood function under the hypothesis $C=1$, $p(x_1|s_1)p(x_2|x_2)p(s_1,s_2|C=1)$, and the likelihood function under the hypothesis $C=2$, $p(x_1|s_1)p(x_2|x_2)p(s_1,s_2|C=2)$, These likelihoods are weighted by the prior probabilities of $C=1$ and $C=2$, respectively. This type of weighted average always appears when marginalization over a discrete variable (here C) is needed.

Exercise 10.5 Show that an alternative way to write the posterior over s_1 and s_2 is as a weighted average of the *posterior distributions* conditioned on C, with weights given by $p(C|x_1,x_2)$. This is pictured in **Fig. 10.4**. This is also known as "Bayesian model averaging", with the understanding that each value of C is interpreted as a "model" the observer has about the world. ■

This is the first time in this book that we encounter a posterior distribution that does not have

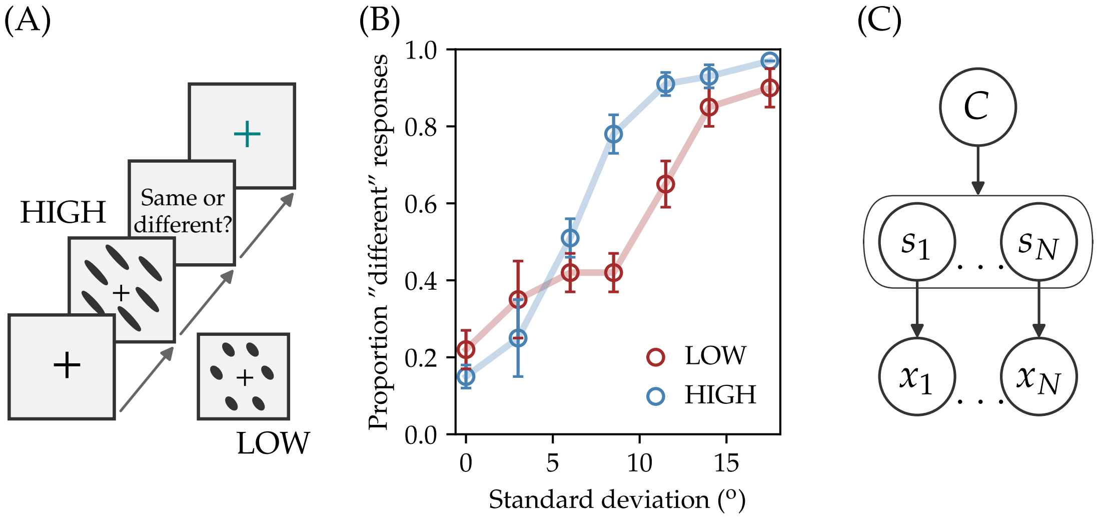


Figure 10.5: Sameness judgment (reproduced from Van den Berg, Vogel, Josic, & Ma, 2011). **(A)** Experimental procedure. Subjects fixated at the cross and a display containing 6 ellipses was shown for 100 ms. Stimulus reliability was controlled by ellipse elongation. In the LOW condition, all ellipses had low reliability. In the HIGH condition, all had high reliability. **(B)** Proportion "different" responses as a function of the standard deviation of the presented set, for the three conditions. **(C)** Generative model of this task.

a single local maximum (unimodal). This posterior has two peaks (is bimodal). For bimodal posteriors, MAP estimation and posterior mean estimation are not equivalent; the latter makes more sense, since it is minimizes the expected squared error.

Causal inference is an important generalization of cue combination. Kording et al. (2007) showed that the causal inference accurately describes human data in an auditory-visual localization task. When observers are asked to report whether both stimuli have the same cause, their reports follow the prediction illustrated in **Fig. 10.3C**. Observers also report the location of the auditory stimulus according to the posterior shown in **Fig. 10.4**.

10.3 Multiple-item sameness judgment

So far we have discussed judging whether two stimuli are the same or not. This idea can immediately be extended to any number of stimuli, say N. When the stimuli are the same ($C = 1$), their common value s is drawn from a distribution $p(s)$. When the stimuli are different ($C = 2$), their values s_i, where the index i now takes values from 1 to N, are drawn independently from the same distribution $p(s)$. William James, one of the founding fathers of psychology, called the sense of sameness "the keel and backbone of our thinking". Judging sameness plays a role in recognizing textures, which tend to consist of elements with the same orientation. Judging sameness is also said to underlie higher cognitive concepts, such as equality and equivalence in mathematics. Many animal species, from honeybees to pigeons to dolphins, can detect sameness at a rather abstract level, suggesting that the concept has had substantial evolutionary importance.

As a concrete example, we consider the case that $p(s)$ is Gaussian with mean 0 and standard deviation σ_s. An example of a sameness judgment experiment is shown in **Fig. 10.5A**. Note that the stimuli can have different elongation, chosen randomly. Here, elongation controls the quality of the orientation information: the more elongated the ellipse, the lower the orientation noise will be.

Example data are shown in **Fig. 10.5B**: the greater the standard deviation of the sample of stimuli shown on a given trial, the more often people would respond “different". Moreover, there was an effect of reliability condition. The generative model of the task is shown in **Fig. 10.5C**. The variables are as follows: C is a binary variable that denotes sameness (1 for same, 0 for different), $\mathbf{s}$ denotes the vector of N orientations presented, and $\mathbf{x}$ denotes the corresponding vector of N measurements. Each x_i is drawn from a Gaussian distribution with mean s_i and standard deviation σ.

The Bayesian observer bases their decision (same or different) on the posterior probability distribution over C given the measurements $\mathbf{x} \equiv (x_1, ..., x_N)^{\mathrm{T}}$. Since C is a binary random variable, we express that posterior as a log posterior ratio:

$$d = \log \frac{p(\mathbf{x}|C=1)}{p(\mathbf{x}|C=2)} + \log \frac{p(C=1)}{p(C=2)} \tag{10.41}$$

Evaluating the likelihoods in this expression, $p(\mathbf{x}|C)$, requires marginalization over the stimulus orientations, $\mathbf{s} \equiv (s_1, ..., s_N)$:

$$p(\mathbf{x}|C) = \int p(\mathbf{x}|\mathbf{s})p(\mathbf{s}|C)d\mathbf{s}. \tag{10.42}$$

As usual, we assume that the standard deviation, σ, of the noise associated with a stimulus is known to the observer for each stimulus and each trial. Therefore, we do not need to marginalize over σ, but can treat it as a known parameter.

When $C = 1$, all elements of the vector $\mathbf{s}$ have the same scalar value s. Then the integral reduces to an integral over this scalar value. Moreover, we assumed that the measurements are conditionally independent, which means that

$$p(\mathbf{x}|\mathbf{s}) = p(x_1|s)p(x_2|s)\cdots p(x_N|s) \tag{10.43}$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-s)^2}{2\sigma^2}}, \tag{10.44}$$

where $\prod$ is notation for a product. Then the likelihood of “same” is

$$\mathscr{L}(C=1) = p(\mathbf{x}|C=1) \tag{10.45}$$

$$= \int \left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-s)^2}{2\sigma^2}} \right) \frac{1}{\sqrt{2\pi\sigma_s^2}} e^{-\frac{s^2}{2\sigma_s^2}} ds \tag{10.46}$$

Although this integral seems daunting, it can be evaluated using a standard equation for the product of normal distributions.

We can similarly evaluate the likelihood of the hypothesis that the stimuli are different, that is, $C = 0$. In that case, all measurements are completely independent from each other, since they do not share a common s. Thus, the N-dimensional integral in Eq. (10.42) reduces to a product of N one-dimensional integrals, one for each measurement (see Box 10.1):

$$p(\mathbf{x}|C=0) = \prod_i \int p(x_i|s_i)p(s_i|C=0)ds_i \tag{10.47}$$

$$= \prod_i \left(\int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-s)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma_s^2}} e^{-\frac{s^2}{2\sigma_s^2}} \right) \tag{10.48}$$

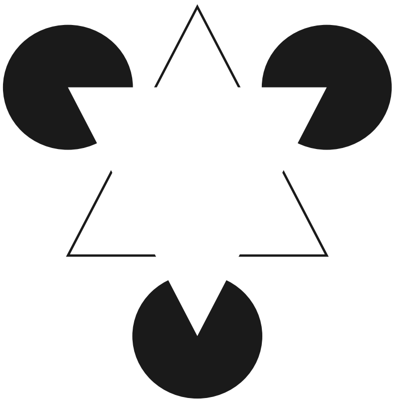


Figure 10.6: Kanizsa triangle. It is easy to perceive a white triangle lying on top of three black discs.

Box 10.1 — Factorizing a multi-dimensional integral. Suppose you have a function of two variables, x and y. The function has the special property that it can be written as the product of a function $f(x)$ that only depends on x, and a function $g(y)$ that only depends on y. Then its integral over x and y can be simplified as

$$\iint f(x)g(y)dxdy = \int f(x)\left(\int g(y)dy\right)dx \tag{10.49}$$

$$= \left(\int f(x)dx\right)\left(\int g(y)dy\right) \tag{10.50}$$

We used the fact that the expression $(\int g(y)dy)$ is simply a constant number (in particular, not a function of x) and can be taken out of the integral over x. Note that this argument only works if x and y are different variables. ■

Using the expressions for $p(\mathbf{x}|C=1)$ and $p(\mathbf{x}|C=0)$, we can evaluate the log posterior ratio in Eq. (10.41) and obtain a decision rule. The decision rule will be a quadratic function of the measurements. We do not complete the derivation here but refer to the Problems. Human subjects judge sameness in a way that is close to the predictions of this model.

This example shows how inference of a relatively abstract quality like "sameness" can be modeled in a Bayesian way using the exact same procedure we used for inferring a physical stimulus.

10.4 Perceptual organization

Extracting information about structure in the world from sensory input is an important part, if not the ultimate goal, of perception. The world is highly structured: the shape of an object consists of a string of small line elements, objects are ordered in depth, and a musical piece consists of delicately arranged sequences of tones. As the famous Kanizsa illusion shows (**Fig. 10.6**), the brain perceives visual structure even if there is only indirect sensory evidence for its existence. Indeed, at some level, all of our perception of structure is based on indirect information. When we walk on a busy street, our brain seemingly effortlessly separates the multiple sound sources from a single continuous stream. Arguably, object recognition, whether visual or auditory, consists of the detection of structures at multiple levels.

Since the birth of psychology, its practitioners have been intrigued by the ways in which the

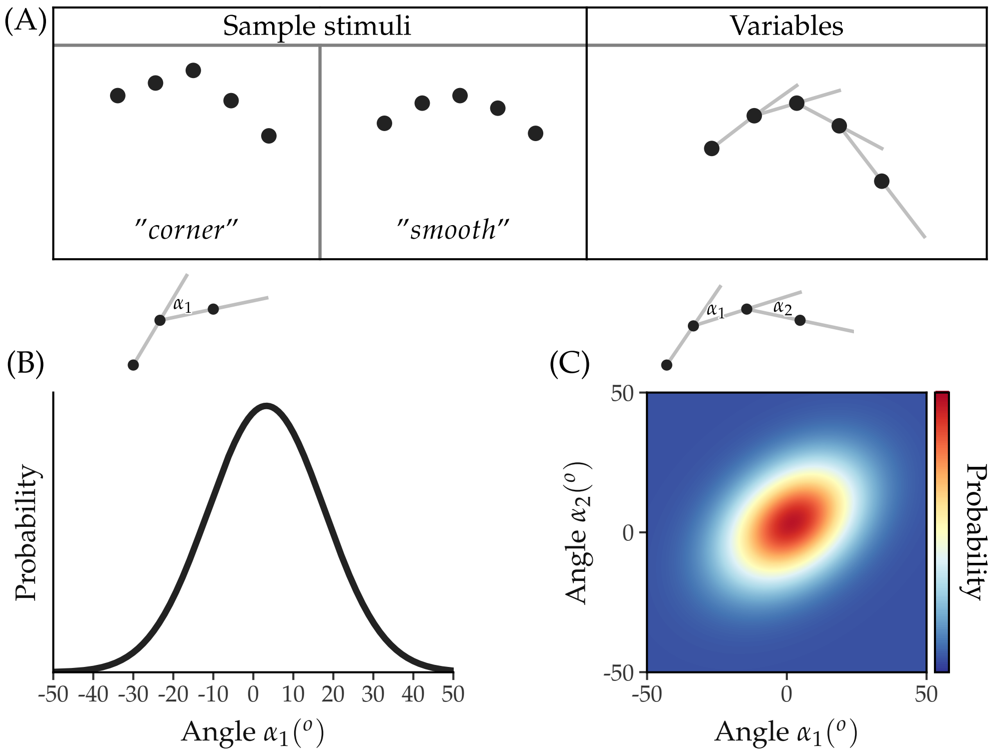


Figure 10.7: Corner or smooth? Figure reproduced from Feldman (2001) **(A)** Task: subjects judged whether or not a corner was formed by five dots. **(B)** Assumed distribution of angle formed by three dots on a smooth contour. **(C)** Assumed distribution of pair of angles formed by four dots on a smooth contour. The two angles are correlated.

brain perceives structure among sets of constituent elements. For the most part, the explanations given for structure perception phenomena have been qualitative and descriptive. Most notable are the so-called Gestalt laws, which describe under which circumstances elements are perceived as belonging to a whole (as in the Kanizsa illusion). However, the Bayesian framework can in many cases provide a more principled and precise account of structure perception, and that is the focus of this chapter.

The same-different inference discussed in previous section is central in Bayesian models of structure perception and perceptual organization: for example, are a set of stimuli the same or not, or do two line elements belong to a single contour?

10.4.1 Simple contours

Objects are defined in part by their boundaries. Therefore, a key part of object recognition is to identify which line elements belong to the same boundary or contour. Contour integration is an example of a task where natural statistics are likely to play an important role in shaping the prior (see Chapter 2, discussion about types of priors). Feldman (Feldman 2001) had human subjects do the task in **Fig. 10.7**: they judged whether five dots formed a corner or a smooth curve.

The top variable in the generative model is a binary variable C. $C = 1$ indicates that the corner is present, $C = 0$ that it is absent. If a corner is present, there are two smooth contours. If no corner is present, there is one smooth contour. Feldman then parameterized the probability of observing an angle α among three dots given that they lie on a smooth contour, using a Gaussian distribution about 0 (**Fig. 10.7B**):

$$p(\alpha|\text{smooth contour}) = \mathcal{N}(\alpha; 0, \sigma^2). \tag{10.51}$$

Similarly, he defined the probability of a pair of angles, α_1 and α_2, being formed by four dots on a smooth contour. This was done using a two-dimensional correlated Gaussian distribution as depicted in **Fig. 10.7C**. The idea is that α_1 and α_2 each have their own variance, but they might also be correlated. For example, if α_1 is positive (rightward deviation), then the smoothness of the contour would make it likely that α_2 is positive. This type of dependence can be captured by a correlation coefficient. Using these building blocks and some additional assumptions, one can compute the likelihood that a corner is present among the five dots. To compute this, one has to marginalize over the location of the corner: it could be at the second, third, or fourth dot.

10.4.2 Natural contours

A different approach to the same problem makes use of natural statistics to specify the generative model. Geisler and Perry (Geisler and Perry 2009) tested the hypothesis that natural statistics combined with a Bayesian model can predict human contour integration judgments. They first formalized the problem by introducing the relevant variables (**Fig. 10.8**). The state of the world of interest is whether two line elements belong to the same contour. This is a binary variable: $C = 0$ (no) or $C = 1$ (yes). The stimuli are two edge elements, and we assume that only their position relative to each other matters. The parameters used to describe this relative position are shown in **Fig. 10.8A**: distance between midpoints (d), the angle between the reference element and the line connecting the midpoints (ϕ), the angle between the reference element and the orientation of the other element (θ), and finally the contrast polarity (ρ): if one were to connect the two elements using a contour, would which side is darker change between the elements?

The generative model of this task is described by the probability distributions $p(d, \phi, \theta, \rho|C = 1)$ and $p(d, \phi, \theta, \rho|C = 0)$. The authors estimated these probabilities by analyzing natural scenes. A photograph of an outdoors, natural scene, such as leaves lying on the forest floor, was first analyzed automatically by an algorithm that extracted the edges. The image was then presented to a human observer with one pixel marked. The observer indicated which other pixels in the image belonged to the same contour. Humans were highly consistent in making these judgments. In this way, the generative model was estimated. Unlike the previous examples in this book, in this case, the generative model was purely specified numerically, that is, through histograms indicating the frequency of occurrence of every combination of parameters. Another difference with most generative models discussed so far is that sensory noise was assumed to be negligible. All uncertainty in the task derives from ambiguity. Finally, a difference is that human observers were used to construct the generative model; the generative model was not constructed by the experimenters.

In a subsequent experiment, different human observers judged whether two edge elements passing under an occluder belonged to the same or to different contours. “Same” and “different” each occurred 50% of the time. A Bayesian observer would make this judgment by computing the posterior ratio. When the prior is flat, reflecting the frequencies of “same” and “different”, the posterior ratio is equal to the likelihood ratio.

$$\frac{p(C = 1|d, \phi, \theta, \rho)}{p(C = 0|d, \phi, \theta, \rho)} = \frac{p(d, \phi, \theta, \rho|C = 1)}{p(d, \phi, \theta, \rho|C = 0)}. \tag{10.52}$$

The modelers were able to make predictions for human judgments based on the generative model. An illustrative example of these predictions is shown in **Fig. 10.8B**. As one might expect, this shows that contours tend to be smooth. Human observers performed close to the Bayesian observer, with similar patterns of errors.

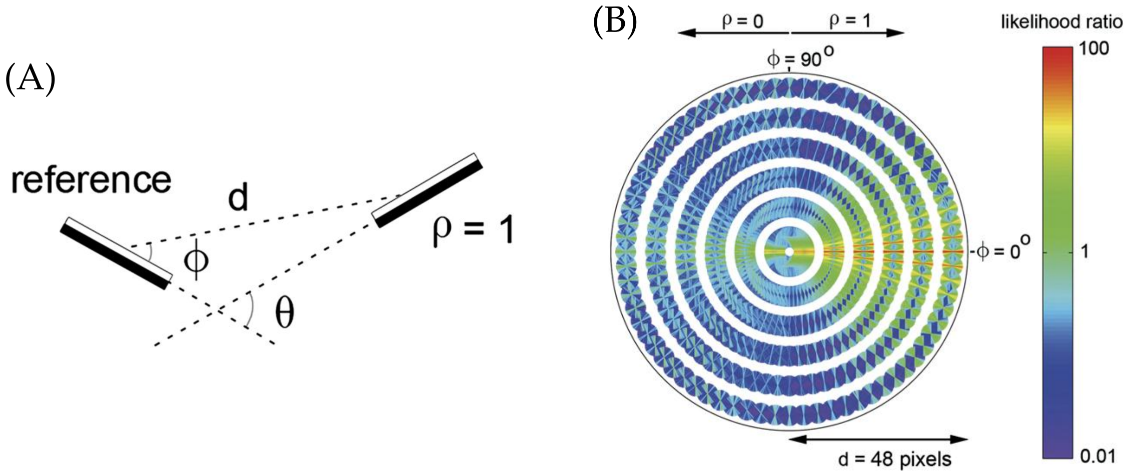


Figure 10.8: **(A)** Parameterization of pairs of edge elements. **(B)** Likelihood ratio as a function of d, ϕ, θ, and ρ. A reference edge element is in the center, oriented horizontally. Likelihood ratios higher than 1 mean that the pair of elements shown is more probable than not to belong to the same contour. From Geisler and Perry (2009); *figure inclusion pending permissions*.

This study is closely related to the notion of "good continuation" in Gestalt psychology. This Gestalt principle says that elements that suggest a continued visual line will tend to be grouped together. By examining the statistics of natural scenes, this rather vague principle can be quantified: elements are grouped together if they have a higher probability of belonging to the same contour than not. This illustrates how Bayesian models can improve on qualitative observations in psychology.

It is instructive to compare the approaches of these two sections. Geisler and Perry used a generative model obtained numerically from natural statistics. This is the more constrained approach, as no parametric assumptions are needed. They also modeled the elements of the contour in greater detail: elements had an orientation and a contrast polarity, instead of being dots. On the other hand, simplifying the problem to dots and using analytical expressions for the generative model allows for easier experimental manipulations and a more concise formulation of the model.

10.4.3 Gestalt laws

Gestalt principles or Gestalt laws have been the leading description of structure perception in psychology. We have mentioned the Gestalt principle of good continuation. There are many additional Gestalt principles, including:

- The law of closure: objects such as shapes, letters, pictures, etc., are perceived as being whole even when they are not complete (**Fig. 10.9A**).
- The law of similarity: elements within an assortment of objects are perceptually grouped together if they are similar to each other (**Fig. 10.9B**).
- The law of common fate: objects are perceived as lines that move along the smoothest path.
- The law of proximity: objects that are close to each other are perceived as forming a group (**Fig. 10.9C-D**).
- The law of good gestalt: objects tend to be perceptually grouped together if they form a pattern that is regular, simple and orderly.

Almost since their conception, Gestalt laws have been criticized for being vague and descriptive, instead of quantitative and explanatory. This is especially evident in the law of good Gestalt, where "regular", "simple", and "orderly" are not defined. Bayesian models have the potential to improve

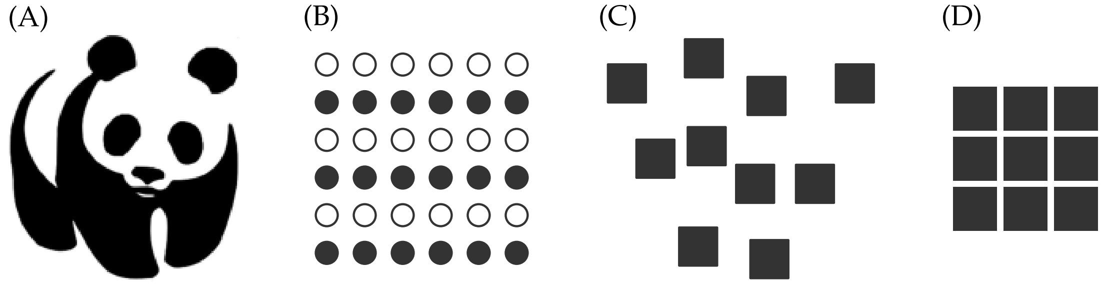


Figure 10.9: (A) Law of closure. **(B)** Law of similarity. **(C-D)** Law of proximity: the left set is seen as nine disjoint squares, and the right set as a single square.

on these laws. The basic idea in every case is that the observer considers two hypotheses – for example, the elements belong together or they don't – and evaluates their posterior probabilities. In Chapter 2, we considered one such Gestalt example, involving moving dots (section 2.5). We now examine another specific case.

Consider the picture in **Fig. 10.10A**. Most people will see this as two intersecting lines, instead of as two angles touching, though either interpretation is possible (see **Fig. 10.10B**; there are in fact more possible world states). The law of continuity would state in this case that individuals tend to perceive the two objects as two single uninterrupted entities, because elements tend to be grouped together when they are aligned.

Here, we don't need to measure natural statistics to be able to make the general argument why Hypothesis 1 is more common. The generative model is shown in **Fig. 10.10C**. The top variable refers to the world state, corresponding to two intersecting lines ($C = 1$) or to two touching angles ($C = 2$). We have to parameterize the stimuli. We start with $C = 1$. A line is parameterized by two numbers, as one can see by writing an equation of a line: $y = ax + b$. Thus, four parameters specify the two lines in Hypothesis 1. Now consider $C = 2$. An angle is parameterized by four numbers: two coordinates for the origin of the angle, one angle for the first leg, and one angle for the second leg. Thus, eight parameters are needed for Hypothesis 2. Finally, in both interpretations, the image is uniquely determined by the parameters.

The Bayesian observer performs inference by computing the posterior ratio of the world states, based on the given image I:

$$\frac{p(C=1|I)}{p(C=0|I)} = \frac{p(I|C=1)}{p(I|C=0)} \frac{p(C=1)}{p(C=0)} \tag{10.53}$$

We cannot say much about the prior, but we can evaluate the likelihood ratio. We denote the parameters in each hypothesis by a vector θ. Thus, θ is four-dimensional when $C = 1$ and eight-dimensional when $C = 2$. Since θ acts as a nuisance parameter, each of the likelihoods is computed by marginalizing over θ. To simplify the argument, let's assume that all parameters take on discrete values. The marginalization is then the sum

$$p(I|C) = \sum_{\theta} p(I|\theta, C) p(\theta|C) \tag{10.54}$$

We know that the image is uniquely determined by the parameters and the hypothesis. Therefore, $p(I|C,\theta)$ equals 0 for all parameter combinations θ except the one that produces the given image I. We denote this parameter combination by θ_I. For this combination, $p(I|\theta,C)$ equals 1. The integral then simply becomes

$$p(I|C) = p_{\theta|C}(\theta_I|C) \tag{10.55}$$

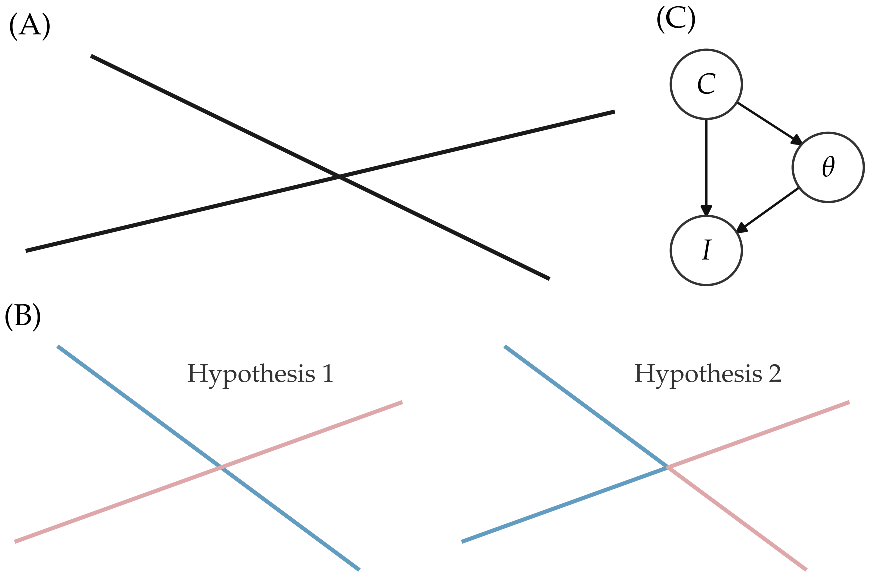


Figure 10.10: **(A)** Image. **(B)** Two possible interpretations of this image ($C = 1$ and $C = 2$). **(C)** Generative model. C is the world state, θ are the parameters, I is the image.

All that remains now is to evaluate the probability of θ_I under each hypothesis. For illustration, let's assume that all parameters are independent and each parameter takes on 100 possible values (this and the following argument are due to MacKay (MacKay 2003)). Then the probability of θ_I (or any other parameter combination) under hypothesis 1 is $\left(\frac{1}{100}\right)^4$, whereas the probability of θ_I (or any other parameter combination) under hypothesis 2 is $\left(\frac{1}{100}\right)^8$. That means that the likelihood ratio is $\left(\frac{1}{100}\right)^{-4} = 10^8$. In other words, Hypothesis 1 is 100 million times as likely as hypothesis 2.

This explains why humans observe the image as two intersecting lines rather than as two touching angles. Intuitively, the hypothesis $C = 1$ requires that the opposite angles in the image share a common vertex and be equal, whereas the hypothesis $C = 2$ permits this configuration but also a vast number of other configurations. The fact that the image conforms to the restricted features predicted by $C = 1$ therefore favors that hypothesis.

Of course, the precise numerical value of the likelihood ratio will depend on our assumptions regarding the priors over the parameters within each hypothesis. However, any Bayesian observer who begins with broad prior distributions will favor hypothesis $C = 1$ when shown the image in **Fig. 10.10A**. The essence of the argument is if two hypotheses can account for the observations equally well, a Bayesian observer will favor the hypothesis that has the lowest number of parameters. In this sense, Occam's razor is an emergent property of Bayesian inference: simpler models are better.

Often, more complex hypotheses (ones with more parameters) can account better for the data. A specific parameter combination within a complex hypothesis may (unlike the example considered here) fit the data more precisely than any parameter combination allowed by a simpler hypothesis. Thus, there is a trade-off between complexity and power. This trade-off is also captured in Eq.10.54, since $p(I|\theta, C)$ is an indication of the power of the hypothesis.

Incidentally, the Bayesian observer who selects between two perceptual hypotheses is mathematically identical to a Bayesian experimenter who analyzes data in order to select between two competing models. Thus, Bayesian model comparison follows the same equations that are here discussed to describe the human brain, including the trade-off between complexity and power. This is discussed in detail in Appendix C.

10.5 Summary and remarks

In this chapter we have introduced the marginalization steps needed to decide if two stimuli are the same or if they are different. We have learned:

- We have structured generative models that differ in complexity, i.e. where the models have different numbers of parameters.
- We have priors over the potential structures.
- Bayes rule can equally be used to calculate a posterior over structures instead of parameters.
- This means that the system uses Bayes rule both to calculate how likely each model is and also about the internal parameters. Bayes is the new Borg.
- The resulting integrals need to be solved.
- The problem of structure estimation occurs for discrete and also continuous variables.
- The problem occurs for Gestalt and contour problems.
- Arguably, most problems in life induce uncertainty about the complexity of the resulting models.

10.6 Suggested readings

- Jacob Feldman. “Bayesian contour integration”. In: *Perception & Psychophysics* 63 (2001), pages 1171–1182
- Wilson S Geisler and Jeffrey S Perry. “Contour statistics in natural images: Grouping across occlusions”. In: *Visual neuroscience* 26.1 (2009), pages 109–121
- Daniel Goldreich and Mary A Peterson. “A Bayesian observer replicates convexity context effects in figure–ground perception”. In: *Seeing and Perceiving* 25.3-4 (2012), pages 365–395
- Konrad P Körding et al. “Causal inference in multisensory perception”. In: *PLoS one* 2.9 (2007), e943
- David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003
- Yoshiyuki Sato, Taro Toyoizumi, and Kazuyuki Aihara. “Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli”. In: *Neural computation* 19.12 (2007), pages 3335–3355
- Ronald Van den Berg et al. “Optimal inference of sameness”. In: *Proceedings of the National Academy of Sciences* 109.8 (2012), pages 3178–3183
- Yanli Zhou, Luigi Acerbi, and Wei Ji Ma. “The role of sensory uncertainty in simple contour integration”. In: *PLoS computational biology* 16.11 (2020), e1006308

10.7 Problems

Problem 10.1 Think of a real world problem which would require same-different judgements that was not covered in this chapter. Formulate it in Bayesian terms.

Problem 10.2 Imagine you are a psychophysicist and want to do an experiment. Come up with an example that was not mentioned in this chapter and sketch how you would set up an experiment to ask if human subjects do same-different judgments in this case.

Problem 10.3 We introduced above equations for common problems. Let us convince ourselves these are correct.

a) Prove Eq. (10.18).
b) Prove Eq. (10.19).
c) Prove Eq. (10.35).

Problem 10.4 . Refer to Section 10.1. Consider the case where the stimuli are the same in half of the cases: $p_{\text{same}} = 0.5$.

a) * Prove that then $d > 0$ reduces to $x_1 x_2 > 0$. Hint: it might be helpful to use the definition and properties of the hyperbolic cosine function.
b) In this case, derive an expression for $p(\hat{C} = 1 | s_1, s_2)$ in terms of cumulative standard normal distribution functions, Φ (see Chapter 8).
c) Based on your answer to (b), derive an expression for proportion correct. Simplify the expression until it has only a single Φ in it.

Problem 10.5 Refer back to Section 10.1.2. Unless $p_{\text{same}} = 0.5$, the decision rule $d > 1$ does not simplify analytically. However, we can still visualize the decision rule in (x_1, x_2) space: in which regions of this space does the Bayesian observer respond "same", and in which regions "different". We want to get an intuition of how σ and p_{same} interact to affect optimal behavior.

a) Choose $\mu = 1$. Consider three values of σ (0.5, 1, and 2), and three values of p_{same} (0.4, 0.5, and 0.6). For each combination of σ and p_{same}, create a plot that shows the boundary between the "respond same" and "respond different" regions in (x_1, x_2) space, where either measurement can take values between -2 and 2. Display the nine plots in a 3-by-3 grid for easy comparison.
b) Describe and interpret the effects of σ and p_{same} on the decision boundary.

Problem 10.6 Refer back to Section 10.1.3. Choose $p_{\text{same}} = 0.4$, $\mu = 1$, and $\sigma = 1$. We will use two numerical techniques for calculating the probability that the observer will report "same" in a given stimulus condition, $p(\hat{C} = 1 | s_1, s_2)$. Note that we need to calculate four numbers, since there are four possible combinations of s_1 and s_2.

a) Use Riemann integration. For both x_1 and x_2, use a grid from -5 to 5 in steps of 0.01. Calculate the four numbers we are looking for.
b) Use Monte Carlo simulation. Compare. The results should be very similar.

Problem 10.7 Let us now look at same-different with continuous variables.

a) Code up the causal inference model and reproduce **Fig. 10.3A**. Use the parameters given in the caption to **Fig. 10.3**.
b) Reproduce **Fig. 10.3C** using Method 2.
c) Explain why fixing $s_1 = 0$ and varying s_2 produces a different result than fixing $s_2 = 0$ and varying s_1.
d) Reproduce **Fig. 10.3C** using Method 3.

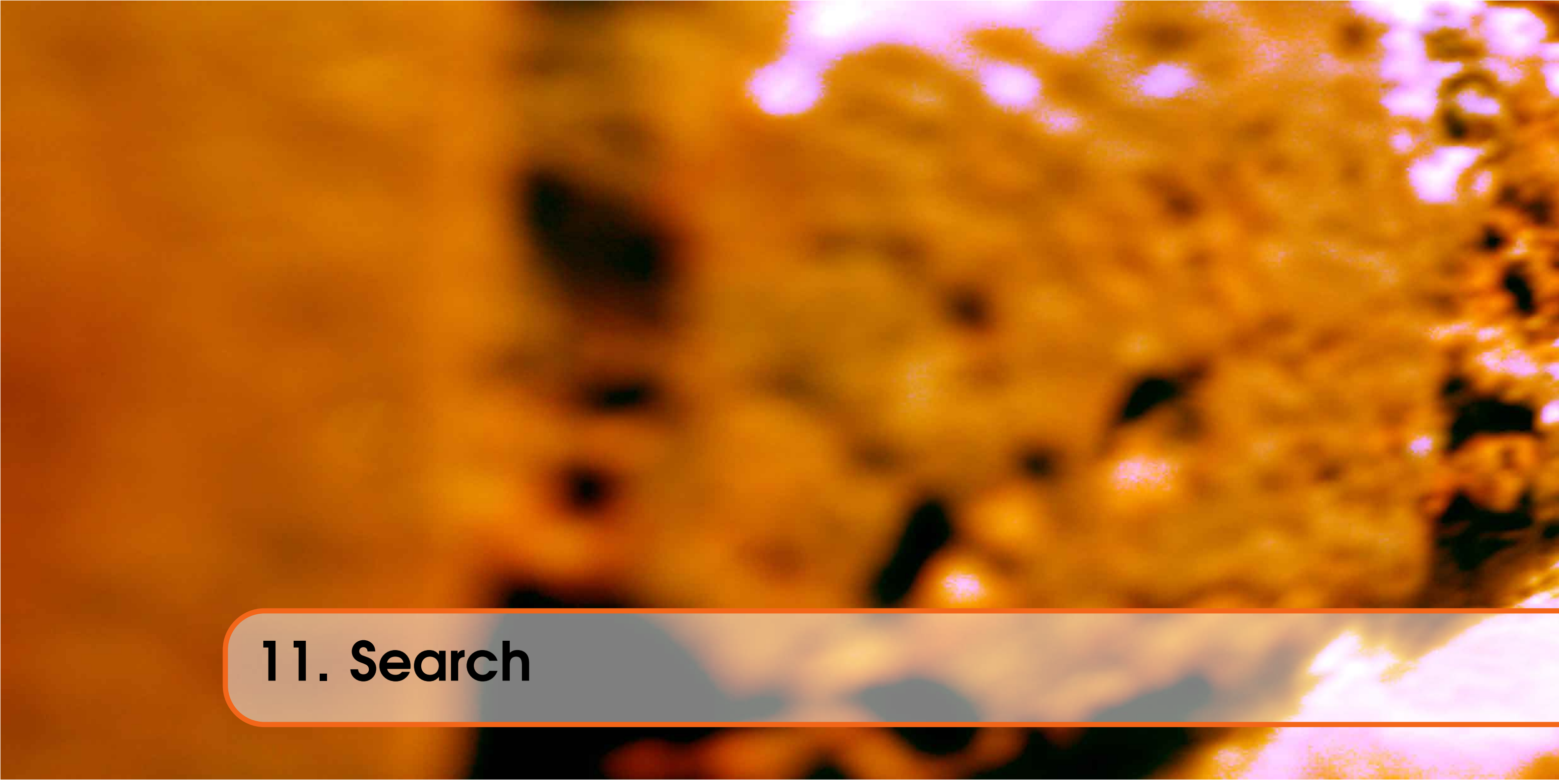
Problem 10.8 Let us now consider the case of bimodal priors. Find x_1, x_2, σ_1, σ_2, and σ_s such that the causal inference model will predict a posterior distribution over s_1 that is visually clearly bimodal.

Problem 10.9 Let us generalize the ideas to multiple stimuli. Suppose there are N stimuli. When $C = 1$, the stimuli are all identical, and their common value is drawn from a normal distribution with mean 0 and standard deviation σ_s. When $C = 2$, all stimuli are drawn independently from the same normal distribution. Assume $p(C = 1) = 0.5$, and independent measurement noise with standard deviation σ.

a) Derive the Bayesian MAP decision rule. The inequality representing the decision rule should have a quadratic function of $x_1, \ldots, x_N$ on the left-hand side and a constant expression on the right-hand side. Hints: 1) Use precision instead of variance notation. 2) Use Section B.7.4. The final decision rule will have on the left-hand side a function of the measurements that can be interpreted relatively easily.
b) Derive the probability of reporting "same" on a "different" trial, when the stimuli are $\mathbf{s} = (s_1, \ldots, s_N)$.

Problem 10.10 A musical tone has a pitch, which defines how high a tone sounds. An experimenter conducts an auditory oddity detection task, as follows. She draws two values of pitch, denoted s (we may think of it as the log of the frequency as that is close to perception), independently from a Gaussian distribution $p(s)$ with standard deviation σ_s (the mean is irrelevant). She then presents to the subject a sequence of three tones, two with the first drawn value of pitch, and one with the second value. The three tones are presented in random order and the subject reports which of the three is the odd one out. Assume that the measured pitch of each tone is independently corrupted by zero-mean Gaussian noise with standard deviation σ.

a) Draw a graphical representation of the generative model, and write down expressions for the probability distributions over the variables.
b) Derive how the Bayesian MAP observer should estimate the temporal location of the oddball (1, 2, or 3) from the measured pitches.
c) Explain why the rule you obtained makes intuitive sense.
d) Assume $\sigma_s = 2$. For each value of σ from 0.05 to 5 in steps of 0.05, simulate 100,000 sets of measurements and use those to estimate the probability correct of the optimal observer. Plot this probability as a function of σ.
e) What is the value of the asymptote as $\sigma \to \infty$?
f) Repeat step (d) for two ad hoc models. In the first ad-hoc model, the observer determines which measurement lies farthest away from the average of the three measurements, and reports the location of that measurement as the oddball location. In the second ad-hoc model, the observer compares the distances between pairs of measurements, finds which of these three distances is smallest, and chooses as the oddball location the location of the measurement not included in that pair. Show that these ad-hoc models lead to lower performance by plotting probability correct of the three models as a function of σ in the same plot.



11. Search

How do we find a target from among many similar objects?

In this chapter, we continue our exploration of inference with multiple stimuli by studying search. Search is the task of looking for a *target* among *distractors*. If the target is known to be present, the search task is a *target localization* task: where is the target? If the target is not known to be present, the search task is a *target detection* task: is the target present?

Search can take place in any sensory modality. For example, you might want to know where in a crowd your friend is standing (visual), determine whether someone is calling your name among many speaking voices (auditory), figure out from which direction an offensive odor is emanating (olfactory), or find a particular coin from among several in your pocket (tactile).

Interestingly, the popular N-alternative forced choice task, while not traditionally presented as such, is a search task: in such a task, the observer localizes a target that is sure to be present among N stimuli, presented either simultaneously or sequentially.

Plan of the chapter

We begin by discussing the ecological relevance of visual search to an animal in the wild and by defining three types of search: target localization, target detection, and target categorization. We consider a specific example of target localization in which a predator attempts to localize a fish, known to be present somewhere in the scene, that hovers above a riverbed of similar appearance. We later consider the same example as a target detection task, in which the predator is unsure whether a fish is present. We consider both target localization and detection in the presence of measurement noise. Finally, we briefly discuss visual search that involves eye movements.

11.1 Ecological relevance of visual search

To an animal in the wild, performing efficient visual search can be a matter of life or death. Frequently, animals must detect whether a predator is present in a visual scene. The predator might be hidden or camouflaged, making it difficult to distinguish from the surrounding visual elements (**Fig. 11.1A**). Predators must similarly attempt to detect and localize well-camouflaged prey. Modern-day humans might want to find a particular piece of paper from among many similar

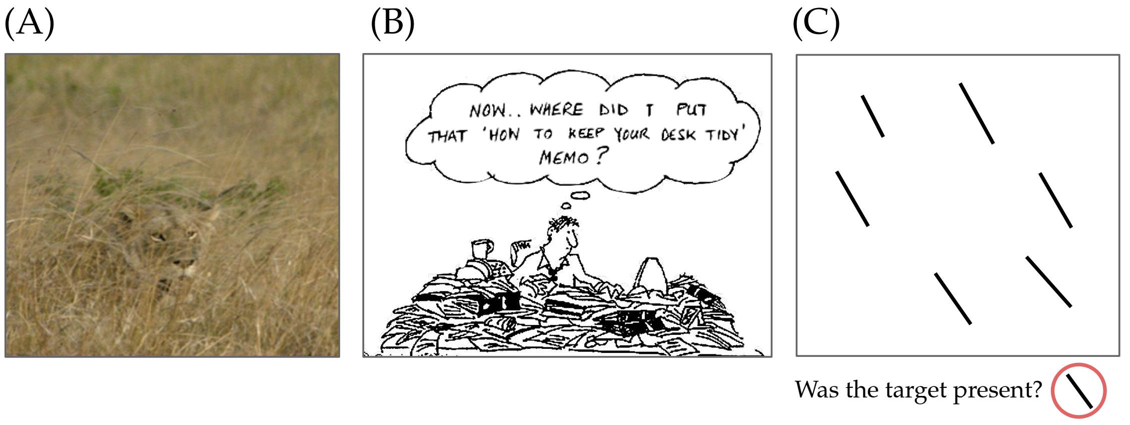

(A)
(B)
(C)
NOW.. WHERE DID I PUT THAT 'HOW TO KEEP YOUR DESK TIDY' MEMO?
Was the target present?

Figure 11.1: Visual search **(A)** In the animal world. **(B)** In the human world. **(C)** In the laboratory; *figure inclusion pending permissions.*

notes on a cluttered desk (**Fig. 11.1B**) or locate a particular product from among many similar ones on the shelves of a grocery store.

In the laboratory, investigators use a variety of simplified tasks to study visual search Doing psychophysics with natural scenes is difficult for several reasons. First, natural scenes are so rich in content that describing them mathematically would require a high-dimensional space. Second, it is not clear what exactly an object is: is the entire tree the object, or is the object an individual branch, or perhaps even a leaf? Third, the noise in the different dimensions of a natural scene typically has a complex and largely unknown correlation structure that goes far beyond the Gaussian noise distribution that we have considered so far. Therefore, in laboratory visual search tasks, as well as in other laboratory psychophysics, it is common to use extremely simplified search scenes that contain a relatively small number of highly distinct objects that differ only along a single stimulus dimension (**Fig. 11.1C**). Obviously, a big gap exists between such simple, artificial scenes and natural tasks, but it is hoped that the majority of the computational principles we can unveil using laboratory tasks matter in natural tasks – in essence, that the laboratory task allows us to study some of the minimal building blocks of the computations that the brain performs in the real world.

There exists a large cognitive psychology literature on visual search, with many descriptive (arguably ad-hoc) models. A line of work dating back to the 1950s has built probabilistic models of visual search, from the earlier pioneers (Peterson, Birdsall, Jaarsma) to later researchers (Palmer, Eckstein, Verghese, Shimozaki, Baldassi, and others). It has been shown recently (Ma, 2011) that visual search, at least for simple stimuli, can, under rather general conditions, be well described using a Bayesian model.

11.2 Types of visual search

Laboratory visual search tasks come in at least three main flavors:

- *Target localization.* It is given that one or more targets are present. The observer has to decide which of the presented stimuli are the targets (discrete), or where the targets are located (continuous).
- *Target detection.* Targets may or may not be present, and the observer has to decide whether any targets are present.
- *Target categorization.* The observer is asked about the category of the target. We will only handle this third flavor in the Problems.

Computationally, these three flavors of search tasks are deeply similar, as we will see in the following pages. In each of these paradigms, there are several factors to consider:

- How many targets can there be? In this chapter, we will for simplicity mostly consider the case of a single target (but see Problem 11.8).
- Are the distractors independent or somehow related to each other? We will throughout the chapter consider mainly the case of independent distractors (but see Problems).
- Is the task hard because of ambiguity, noise, or both? In many visual search tasks, ambiguity is sufficient to make the task hard. It also simplifies our treatment to assume no measurement noise, so in this chapter we will start with that setting.

In each paradigm, the location of the target is a priori unknown to the observer. For this reason, the observer has to consider every possible location. In localization tasks, target location is the world state variable of interest. In detection tasks, target location is a nuisance parameter and the Bayesian observer marginalizes over this variable.

11.3 Target localization: camouflage

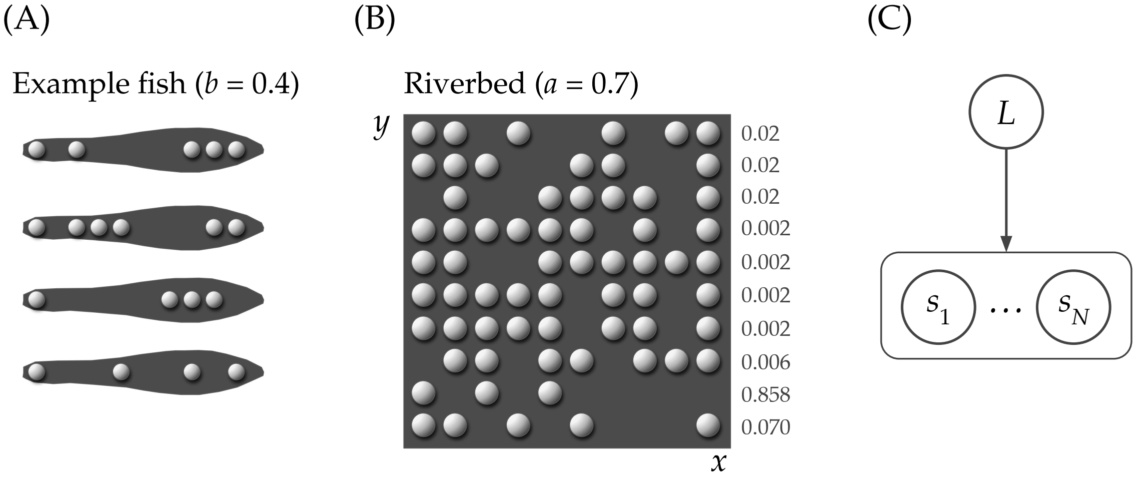


Figure 11.2: The challenge of localizing a camouflaged object. **(A)** Four members of the dotted fish species with dot probability $b = 0.4$. The fish has size 1×10. **(B)** A 10×10 grid of riverbed, with pebble probability $a = 0.7$, in which a fish is present in one row (i.e., one y-position). Next to each row, we list the posterior probability that the fish is present in that row. **(C)** Generative model. Here, the number of rows, N, is 10.

A beautiful naturalistic example of search is seen in animals attempting to perceive camouflaged predators or prey. In Chapter 1, we discussed camouflage as an evolved strategy for producing broad likelihood functions in the observer. Here we consider a toy problem inspired by the image of the flounder (**Fig. 1.9**). Suppose that a dotted fish species inhabits a river, and that a fish sometime hovers just above the pebble-covered riverbed. River fish often point upstream, as this orientation provides easier stability against the current, and faces the fish in the direction of potential food that is being swept downstream. We assume that the dotted fish takes this orientation (**Fig. 11.2A**). We divide the area under consideration into a 10×10 grid and consider a species of dotted fish that has size 1×10. If the fish is present, then the prior probability of it being located in any of the 10 rows is equal. We assume that the fish skin color is identical to that of the riverbed mud, and that each dot on the skin closely resembles a pebble. Suppose that each grid square on the riverbed independently has a probability a of containing a pebble and $1 - a$ of not containing a pebble. For the fish species, the probability of a dot within any grid square on the skin is b, and of no dot is $1 - b$.

Exercise 11.1 Which factors do you think are important for camouflage. Which ones do animals in nature regularly use? Do you know of any that are missing? ■

We will begin our study of search with a localization task. Consider a predator that knows a fish is present in this area; perhaps the predator saw the outline of the fish as it moved in this general direction a moment earlier. The predator's task is to determine, given a visual observation such as in **Fig. 11.2B**, in which row the fish is located. Each row is a "stimulus" s_i. The predator knows that exactly one of N presented stimuli $s_1, \ldots, s_N$ is the target, and has to decide which one. We denote the index of the target by L; thus, s_L is the target stimulus.

Step 1: Generative model (Fig. 11.2C). The generative model contains target location L, which is an integer between 1 and 10, the number of possible y positions for the fish, and the stimuli $(s_1, \ldots, s_N)$, which we will sometimes summarize by $\mathbf{s}$ (boldface notation for a vector). The uniform prior is $p(L) = \frac{1}{N}$. We next need to specify the distribution of the stimulus vector $\mathbf{s}$ given a target location L. In our scenario, the appearances of the target (fish) and each of the distractors (rows of pebbles) are drawn independently from their respective distributions. Thus,

$$p(\mathbf{s}|L) = p(s_1|L)p(s_2|L)\cdots p(s_N|L) \equiv \prod_{i=1}^{N} p(s_i|L). \tag{11.1}$$

We denote the distribution of the target stimulus by $p_{\text{target}}(s)$ and the distribution of the distractor stimulus by $p_{\text{distractor}}(s)$. Then

$$p(s_i|L) = \begin{cases} p_{\text{target}}(s_i) & \text{if } i = L \\ p_{\text{distractor}}(s_i) & \text{if } i \neq L \end{cases} \tag{11.2}$$

We can work out these probabilities using the proportion of dots on a fish and the proportion of pebbles in the riverbed. This calculation parallels the one of the probability of a sequence of dry and rain days in Section 6.1. For example, if the stimulus is a fish, then the probability of a particular dot pattern is obtained by multiplying 10 factors, one for each horizontal position, where a dot contributes a factor b and an empty position contributes a factor $1-b$. As a result, if we denote the number of dots in row i as n_i, we have

$$p_{\text{target}}(s_i) = b^{n_i}(1-b)^{10-n_i} \tag{11.3}$$

$$p_{\text{distractor}}(s_i) = a^{n_i}(1-a)^{10-n_i}. \tag{11.4}$$

This completes the specification of the generative model.

Step 2: Inference. Even in this relatively simple case of visual search, we still have a new form of inference, where the observer needs to combine information across the N items: after all, the target is present in *exactly one* location. That means that evidence in favor of one row containing pebbles should count as evidence in favor of one of the other rows containing the fish.

We now formalize the inference. The observation is a set of stimuli $\mathbf{s}$. The world state variable of interest is target location L, so we want to calculate the posterior over L, $p(L|\mathbf{s})$. By Bayes' rule,

$$p(L|\mathbf{s}) \propto p(\mathbf{s}|L)p(L) \tag{11.5}$$

Given the uniform prior and Eq. (11.1), this becomes

$$p(L|\mathbf{s}) \propto \prod_{i=1}^{N} p(s_i|L) \tag{11.6}$$

Substituting Eq. (11.12),

$$p(L|\mathbf{s}) \propto p_{\text{target}}(s_L) \prod_{i \neq L} p_{\text{distractor}}(s_i) \tag{11.7}$$

Since we are in Step 2, L is the *hypothesized* target location and s_L is the observed stimulus value at that location. The product is over all distractors, which means all stimuli besides the target. Therefore, we can also write

$$p(L|\mathbf{s}) \propto \frac{p_{\text{target}}(s_L)}{p_{\text{distractor}}(s_L)} \prod_{i=1}^{N} p_{\text{distractor}}(s_i) \tag{11.8}$$

Now, the product is over all stimuli and therefore does not depend on L. Since our relationship is only a proportionality, we can write

$$p(L|\mathbf{s}) \propto \frac{p_{\text{target}}(s_L)}{p_{\text{distractor}}(s_L)} \tag{11.9}$$

Interestingly, the fraction is the likelihood ratio of target presence if there had been only a single stimulus, or in other words, the "local" likelihood ratio of classifying s_L as a target versus a distractor.

Finally, to obtain a proper posterior, the right-hand side of Eq. (11.9) needs to be normalized by dividing by the sum of that expression over all L. Eq. (11.9) is the fundamental relationship for localization of a single target among independent distractors.

So far, we used the following aspects of the generative model: a) that distractors are drawn independently from a distractor distribution; b) that there is only a single target, drawn independently from a target distribution; c) that if the target is present, it has the same probability of being at each location (although this is easily generalized, by replacing $\frac{1}{N}$ by $p(L)$ everywhere). We did not yet use the specifics of p_{target} and $p_{\text{distractor}}$ in our camouflage scenario. We now do so by substituting the expressions from Eqs. (11.3) and (11.4) into Eq. (11.9) for the posterior over target location:

$$p(L|\mathbf{s}) \propto \left(\frac{b}{a}\right)^{n_i} \left(\frac{1-b}{1-a}\right)^{10-n_i}. \tag{11.10}$$

Exercise 11.2 Show that this derivation is correct. ■

We calculated these posterior probabilities in the example of **Fig. 11.2B**.

If the predator maximizes accuracy, they would perform maximum-a-posteriori (MAP) estimation of L. This means picking the value of L for which the right-hand side of Eq. (11.9) is highest. We can write this as

$$\hat{L} = \underset{L}{\operatorname{argmax}} \frac{p_{\text{target}}(s_L)}{p_{\text{distractor}}(s_L)}. \tag{11.11}$$

An interesting aspect of this rule is that each candidate target stimulus, s_L, should be considered individually. The observer computes for each individual stimulus the likelihood ratio as if they were *classifying* (categorizing) that stimulus as a target or a distractor. Thus, the stimulus with the highest categorization likelihood ratio is also the one with the highest posterior probability of being the target in this search task. This is not a general law; it crucially depends on the independence assumption we made here.

Since there is no measurement noise, Step 3 is unnecessary: for the same set of stimuli $\mathbf{s}$, the Bayesian observer will always make the same response[1]. However, it is not trivial to calculate performance metrics (such as proportion correct) for given a and b, as this would require marginalizing over the $\mathbf{s}$ associated with these parameters. We will do this in Problem 11.12.

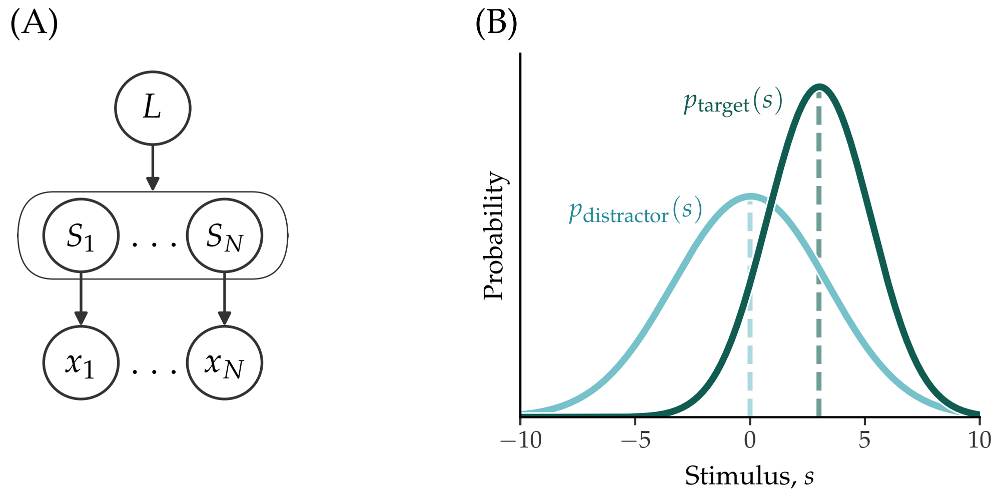


Figure 11.3: Visual search with a single target, independent distractors, and measurement noise. **(A)** Generative model when there are N stimuli. L is the index of the target; it is an integer between 1 and N. **(B)** Example target and distractor distributions. Here, the target distribution is normal with mean 3 and standard deviation 2, and the distractor distribution is normal with mean 0 and standard deviation 3. The two distributions overlap, so distractors can be confused with the target.

11.4 Target localization with measurement noise

In the previous example, stimuli were discrete. We will now consider a continuous example, and we will also introduce our familiar concept of measurement noise into a search problem. In this example, the uncertainty to the observer does not come only from the overlap between target and distractor *stimulus distributions*, but also from the overlap between their *measurement distributions*.

Step 1: Generative model (Fig. 11.3A). We still consider tasks, just like the fish example, in which all stimuli are independent conditioned on L.

Given target location L, we also again have.

$$p(s_i|L) = \begin{cases} p_{\text{target}}(s_i) & \text{if } i = L \\ p_{\text{distractor}}(s_i) & \text{if } i \neq L \end{cases} \tag{11.12}$$

Now, $p_{\text{target}}(s)$ and $p_{\text{distractor}}(s)$ are continuous distributions; examples are shown in **Fig. 11.3B**.

Next, we will make our usual assumption that the measurements are independent conditioned on the stimuli. These two forms of conditional independence go well together. Specifically, we denote the vector of measurements by $\mathbf{x}$, and its conditional distribution by $p(\mathbf{x}|\mathbf{s})$. We now also allow a general prior distribution $p(L)$.

Step 2: Inference. Bayes' rule now takes the form

$$p(L|\mathbf{x}) \propto p(L) \int p(\mathbf{x}|\mathbf{s}) p(\mathbf{s}|L) d\mathbf{s}. \tag{11.13}$$

This equation indicates that the observer marginalizes over the stimuli, which in this problem are nuisance parameters. The marginalization is multi-dimensional. Thanks to conditional independence

[1] This could be a curse rather than a blessing, since actual humans would not necessarily always make the same response. To account for such variability, decision noise can be added to the model.

of measurements, Eq. (11.13) can be written as

$$p(L|\mathbf{x}) \propto p(L) \int \cdots \int \left(\prod_{i=1}^{N} p(x_i|s_i) \right) \left(\prod_{i=1}^{N} p(s_i|L) \right) ds_1 \cdots ds_N. \tag{11.14}$$

The two products combine, and each s_i appears in only one factor in the resulting product. Then, the multi-dimensional integral reduces to a product of one-dimensional integrals (see Box 10.1):

$$p(L|\mathbf{x}) \propto p(L) \prod_{i=1}^{N} \int p(x_i|s_i) p(s_i|L) ds_i. \tag{11.15}$$

Thus, comparing to Eq. (11.6), we have replaced the likelihood $p(x_i|L)$ with the likelihood $\int p(x_i|s_i)p(s_i|L)ds_i$. The rest of the derivation proceeds analogous to the previous section, and we find for the posterior

$$p(L|\mathbf{s}) \propto p(L) \frac{\int p(x_i|s_i) p_{\text{target}}(s_i) ds_i}{\int p(x_i|s_i) p_{\text{distractor}}(s_i) ds_i} \tag{11.16}$$

This again parallels a binary classification task, in particular one we have seen before, in Chapter 8: the ratio is the "local" likelihood ratio of stimulus s_L being a target versus a distractor.

In sum, when target and distractors are all independent conditioned on target location, the likelihood function over *target location* only requires the local likelihood ratios for target/distractor *classification*.

Special case: single target value, single distractor value

We now consider the special case where the target can take on only one particular value, s_T, and the distractor can take on only one particular other value, s_D. Formally, this is written as a set of delta functions,

$$p_{\text{target}}(s) = \delta(s - s_T) \tag{11.17}$$

$$p_{\text{distractor}}(s) = \delta(s - s_D). \tag{11.18}$$

Because we now have measurement noise, localizing the target can be a hard problem: the target is confusable with the distractors because of this noise. We assume that the measurement x_i follows a Gaussian distribution with mean s_i and variance σ_i^2. Then in Eq. 11.16, the delta functions and the integrals "cancel" each other out and we are left with

$$p(L|\mathbf{s}) \propto p(L) e^{\frac{s_T - s_D}{\sigma_L^2}\left(x_L - \frac{s_T + s_D}{2}\right)} \tag{11.19}$$

This parallels the discrimination task of Chapter 7: the ratio is the likelihood ratio of stimulus s_L being s_T versus s_D.

11.5 Target detection: camouflage

We explained in Section 11.2 that localization is only one of various forms of search. Another important form of search is target detection, the problem of determining whether a target object (or multiple target objects) are present or absent in a scene. In this chapter, we assume that there is no more than a single target in the scene. We now continue with the camouflage example from Section 11.3 but treat it as a detection problem. Suppose that the predator is newly arrived on the scene and can no longer assume that a fish is present. The predator's task, given a visual observation of the riverbed, is to determine whether a fish is present ($C = 1$) or not ($C = 0$).

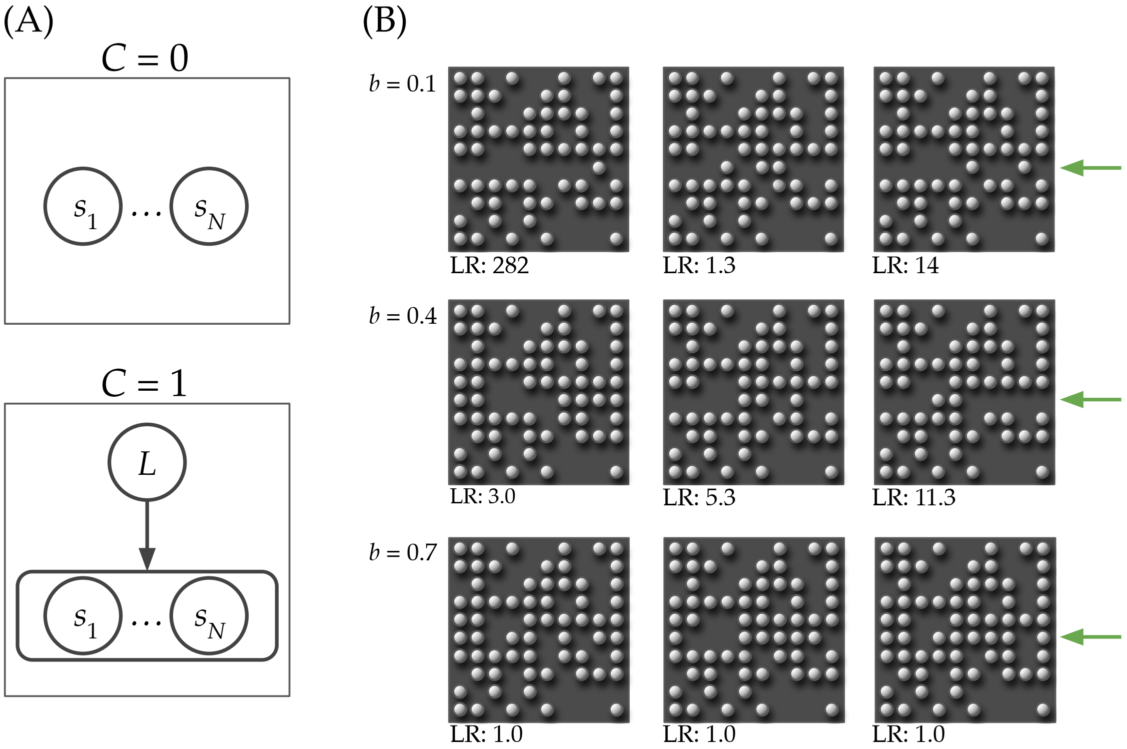


Figure 11.4: The challenge of detecting a camouflaged object. (A) Generative model diagram. For the current example, $N = 10$. (B) A riverbed for which $a = 0.7$. Three fish were randomly selected from species characterized by either $b = 0.1$, 0.3, or 0.7 (nine fish in total). The arrows indicate the true location of the fish. The likelihood ratio for fish presence is shown below each image.

Step 1: Generative model. The generative model (**Fig. 11.4A**) is similar to the one in Section 11.3 but now contains an extra variable, C indicating the absence of presence of a target. The binary variable C comes with a prior $p(C)$. It also causes the generative model to have two components, just as in Chapter 10. If the target is absent ($C = 0$), all stimuli s_i are distractors, i.e. riverbed with a pebble probability of a for each square. If the target is present ($C = 1$), the generative model is the same as in **Fig. 11.2C**, with L denoting target location and $\mathbf{s} = (s_1, \ldots, s_N)$ the set of dot patterns. When $C = 1$, only a single target is present, and each location is equally probable, $p(L) = \frac{1}{N}$.

Step 2: Inference. The observer (predator) is now interested in inferring whether a fish is present, i.e. the binary variable C, from an observation of the riverbed, $\mathbf{s}$. **Fig. 11.4** shows the observations for 9 randomly selected fish, with $b = 0.1$, 0.4, or 0.7 (three fish each), hovering above the same riverbed characterized by $a = 0.7$.

Following our usual procedure when we encounter a binary variable, we express the posterior over C using the log posterior ratio:

$$d = \log \frac{p(C=1|\mathbf{s})}{p(C=0|\mathbf{s})} \tag{11.20}$$

$$= \log \frac{\mathscr{L}(C=1;\mathbf{x})}{\mathscr{L}(C=0;\mathbf{x})} + \log \frac{p(C=1)}{p(C=0)} \tag{11.21}$$

We examine the likelihoods separately, starting with $C = 0$. Using the independence of the observations, we immediately find for the likelihood of $C = 0$:

$$\mathscr{L}(C=0;\mathbf{x}) \equiv p(\mathbf{s}|C=0) \tag{11.22}$$

$$= \prod_{i=1}^{N} p(s_i|C=0). \tag{11.23}$$

The likelihood of $C = 1$ is a bit more complicated because the target location, L, serves as a nuisance variable and therefore we need to marginalize over it:

$$\mathscr{L}(C=1;\mathbf{x}) = p(\mathbf{s}|C=1) \tag{11.24}$$

$$= \sum_{L=1}^{N} p(L)p(\mathbf{s}|L,C=1) \tag{11.25}$$

$$= \frac{1}{N}\sum_{L=1}^{N} p(\mathbf{s}|L,C=1) \tag{11.26}$$

$$= \frac{1}{N}\sum_{L=1}^{N}\left(\prod_{i=1}^{N} p(s_i|L,C=1)\right), \tag{11.27}$$

where we have used conditional independence of the stimuli for the last equality. We recognize the product inside the sum, $\prod_{i=1}^{N} p(s_i|L,C=1)$, from Eq. (11.6). We can borrow from the derivation following that equation to further evaluate Eq. (11.27). This yields

$$\mathscr{L}(C=1;\mathbf{x}) = \frac{1}{N}\left(\prod_{i=1}^{N} p_{\text{distractor}}(s_i)\right)\sum_{L=1}^{N}\frac{p_{\text{target}}(s_L)}{p_{\text{distractor}}(s_L)}. \tag{11.28}$$

Combining Eqs. (11.23) and (11.28), we can now calculate the likelihood ratio of the target being present versus absent,

$$\text{LR} \equiv \frac{\mathscr{L}(C=1;\mathbf{x})}{\mathscr{L}(C=0;\mathbf{x})} \tag{11.29}$$

$$= \frac{1}{N}\sum_{L=1}^{N}\frac{p_{\text{target}}(s_L)}{p_{\text{distractor}}(s_L)}. \tag{11.30}$$

In parallel to Eq. (11.9), the fraction inside the sum is the "local" likelihood ratio of classifying s_L as a target versus a distractor. We could denote the corresponding likelihood ratio by LR_L,

$$\text{LR}_L \equiv \frac{p_{\text{target}}(s_L)}{p_{\text{distractor}}(s_L)}, \tag{11.31}$$

so that Eq. (11.30) becomes

$$\text{LR} = \frac{1}{N}\sum_{L=1}^{N}\text{LR}_L. \tag{11.32}$$

In other words, under the assumptions that we made, the "global" likelihood ratio LR is equal to an average of the local likelihood ratios[2]. The *log* likelihood ratio and the log posterior ratio are

$$\text{LLR} = \log\left(\frac{1}{N}\sum_{L=1}^{N} e^{\text{LLR}_L}\right), \tag{11.33}$$

$$d = \text{LLR} + \log\frac{p(C=1)}{p(C=0)}, \tag{11.34}$$

[2] If we had used a general prior over location, $p(L)$, this average would have been a weighted average, with the weights given by $p(L)$.

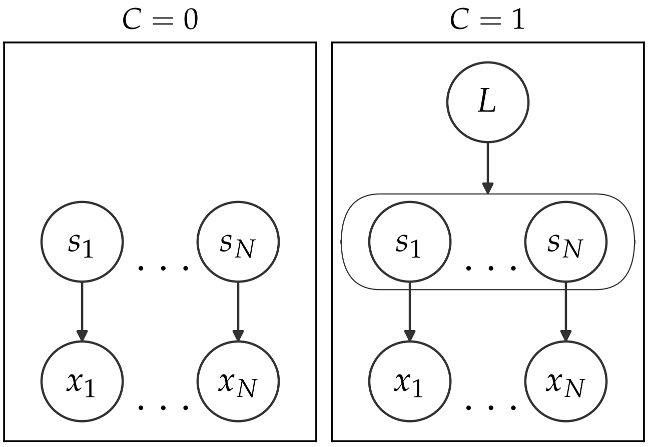


Figure 11.5: Graphical depiction of the generative model for detection of a single target. $C = 0$: target absent. $C = 1$: target present.

where $\mathrm{LLR}_L = \log \mathrm{LR}_L$. Eq. (11.33) shows that the "global" log likelihood ratio of target presence is a nonlinear function of the "local" log likelihood ratios. One can think of this as a spatial integration rule for evidence. Keep in mind that in this derivation, we have made essential use of the specific statistical properties of the task modeled here: that the distractors are independent, and that if there is a target, there is only one target. If these properties do not apply, then Eq. (11.33) might not either.

Finally, we can use the specifics of p_{target} and $p_{\text{distractor}}$ in our camouflage scenario. Doing so, Eq. (11.30) for the likelihood ratio of the fish being present versus absent becomes

$$\mathrm{LR} = \frac{1}{10}\sum_{L=1}^{10}\left(\frac{b}{a}\right)^{n_L}\left(\frac{1-b}{1-a}\right)^{10-n_L}, \tag{11.35}$$

where n_L is the number of dots within the L^{th} row. We explore the consequences of this expression in Problem 11.12.

11.6 Target detection with measurement noise

We now consider target detection with measurement noise. This section can be seen as a combination of Section 11.4 on target localization with measurement noise, and Section 11.5 on target detection without measurement noise.

Step 1: Generative model (Fig. 11.5). The generative model is the same as in Section 11.5 except a) we use a general prior, $p(L)$, instead of a uniform one; b) there is now a layer of noisy measurements generated from the stimuli. We assume that the measurements are independent conditioned on the stimuli.

Step 2: Inference. Inference proceeds as in Section 11.5 except that now the measurements are given, not the stimuli; the latter are unknown to the observer. Thus, the log posterior ratio is

$$d = \log\frac{p(C=1|\mathbf{x})}{p(C=0|\mathbf{x})} \tag{11.36}$$

$$= \log\frac{p(\mathbf{x}|C=1)}{p(\mathbf{x}|C=0)} + \log\frac{p(C=1)}{p(C=0)}. \tag{11.37}$$

The likelihoods are obtained by marginalizing over the stimuli, $\mathbf{s}$. The likelihood of $C = 0$ becomes

$$
\begin{aligned}
\mathscr{L}(C=0;\mathbf{x}) &\equiv p(\mathbf{x}|C=0) && (11.38)\\
&= \int p(\mathbf{x}|\mathbf{s})p(\mathbf{s}|C=0)d\mathbf{s} && (11.39)\\
&= \prod_{i=1}^{N} \int p(x_i|s_i)p(s_i|C=0)ds_i && (11.40)\\
&= \prod_{i=1}^{N} \int p(x_i|s_i)p_{\text{distractor}}(s_i)ds_i && (11.41)
\end{aligned}
$$

and similarly for $C = 1$. Other than that, we follow Section 11.5 and find for the likelihood ratio of target presence,

$$
\begin{aligned}
\text{LR} &\equiv \frac{\mathscr{L}(C=1;\mathbf{x})}{\mathscr{L}(C=0;\mathbf{x})} && (11.42)\\
&= \sum_{L=1}^{N} p(L)\frac{\int p(x_L|s_L)p_{\text{target}}(s_L)ds_L}{\int p(x_L|s_L)p_{\text{distractor}}(s_L)ds_L} && (11.43)
\end{aligned}
$$

This means that Eq. (11.32) for relating the "global" likelihood LR to the local ones LR_L still holds.

Relationship to embedded class task from Section 8.6. We now consider the case that $N = 1$, the target value follows a Gaussian distribution with mean 0 (or any arbitrary value) and variance σ_1^2, and the distractor value follows a Gaussian distribution with the same mean and variance σ_2^2. Then, the target detection task reduces to the embedded class task from Section 8.6 (see in particular Fig. 8.5). An important special case is that the target has a single value ($\sigma_T = 0$). For general N, with otherwise the same assumptions, the embedded class task is still the building block of the target detection task.

Special case: single target value, single distractor value. As in Section 11.4, we consider the case of a single target value, s_T, and a single distractor value, s_D. The local log likelihood ratio of target presence is then equal to

$$
\text{LLR}_L = \frac{s_T - s_D}{\sigma_L^2}\left(x_L - \frac{s_T + s_D}{2}\right) \qquad (11.44)
$$

11.7 Remarks

Some special cases described here have been extensively tested in laboratory experiments, for example detection of a vertical target among distractors whose orientations are independently drawn from a uniform distribution. To create the conditions in which the models would apply, these are experiments in which the subject is typically asked to fixate at the center of the screen, items are presented at a fixed distance from that center, items have a single relevant feature, and the display is presented for a very short period of time (e.g. 100 ms) [94].

At least when distractors are heterogeneous, the Bayesian model of search describes human data well as long as it is additionally assumed that the amount (standard deviation) of the measurement noise increases as the number of stimuli increases [83]. This can be thought of as a "resource limitation".

It is worth noting that Bayesian models of simple visual search have a long history, dating back at least to [**peterson1954**]. However, for many decades, starting with [**nolte1967**], Bayesian models were overshadowed by "signal detection theory models", in particular the maximum-of-outputs or max model. The max model is reasonable when distractors are homogeneous (identical to each other), a case not considered in this chapter. However, it completely fails when distractors are heterogeneous. It can be generalized, but those generalizations turn out to be less convincing than the Bayesian model.

In natural visual search, eye movements are crucial. In laboratory experiments, the role of eye movements is sometimes deliberately minimized by requiring the participant to fixate. This has both experimental and modeling advantages [94]. Nevertheless, Bayesian models can also be applied to visual search tasks in which eye movements are allowed, such as free-viewing a scene while localizing a target. In such tasks, the goal is often to predict the next fixation location. Posterior distributions can be used to evaluate the expected benefits of different candidate fixation locations [89, 124]. These scenes are still not natural, but have carefully controlled statistics. There is a lot of work on modeling visual search in natural scenes. This work is much more useful for applications than anything we discussed in this chapter, but it is beyond the scope of this book because it is not Bayesian.

11.8 Summary and remarks

In this chapter, we introduced the generative models for search. We have learned:

- Search is an ecologically important task.
- Examples of search are target localization and target detection, but there is also a connection to two-interval forced choice designs.
- Search leads to a statistical dependency structure between all potential targets.
- This produces a more complex inference problem which we can still treat using the same Bayesian methods.
- When stimuli and measurements are *conditionally* independent, the resulting integrals can be solved because they factorize.
- Target detection introduces a statistical model where the structure is unknown on top of which element is the target.

11.9 Suggested readings

- Miguel P Eckstein. “Visual search: A retrospective”. In: *Journal of vision* 11.5 (2011), pages 14–14
- Wei Ji Ma et al. “Behavior and neural basis of near-optimal visual search”. In: *Nature neuroscience* 14.6 (2011), pages 783–790
- Helga Mazyar, Ronald Van den Berg, and Wei Ji Ma. “Does precision decrease with set size?” In: *Journal of vision* 12.6 (2012), pages 10–10
- Jiri Najemnik and Wilson S Geisler. “Optimal eye movement strategies in visual search”. In: *Nature* 434.7031 (2005), pages 387–391
- John Palmer, Preeti Verghese, and Misha Pavel. “The psychophysics of visual search”. In: *Vision research* 40.10-12 (2000), pages 1227–1268
- Ruth Rosenholtz. “Visual search for orientation among heterogeneous distractors: Experimental results and implications for signal-detection theory models of search.” In: *Journal of Experimental Psychology: Human Perception and Performance* 27.4 (2001), page 985
- Scott Cheng-Hsin Yang, Mate Lengyel, and Daniel M Wolpert. “Active sensing in the categorization of visual patterns”. In: *Elife* 5 (2016), e12215

11.10 Problems

Problem 11.1 Come up with an example where a search problem happens in everyday life that has not been used as an example in this chapter.

Problem 11.2 In a way this chapter deals with cases where integration happens over many places and their relevant visual content. Make a list of local visual content of coffee cups and their spatial relations. Discuss how real-world recognition of coffee cups may differ from the kinds of modeling we do here.

Problem 11.3 We mentioned that in experiments testing the models of this chapter, the subject is typically asked to fixate at the center of the screen, items are presented at a fixed distance from that center, items have a single relevant feature, and the display is presented for a very short period of time (e.g. 100 ms). Explain how deviations from each of these design elements could create the need for more complex models.

Problem 11.4 Consider the case of Section 11.4 but without measurement noise. Assume that the target distribution is normal with mean μ_T and variance σ_T^2, and that the distractor distribution is normal with mean μ_D and variance σ_D^2, as in **Fig. 11.3B**.

a) Show that the posterior over target location L is

$$p(L|s) \propto p(L)e^{-\frac{1}{2}(J_T-J_D)\left(s_L-\frac{J_T\mu_T-J_D\mu_D}{J_T-J_D}\right)^2}. \tag{11.45}$$

b) Assume $N=3$, $p(L)=(0.3,0.3,0.4)$, and the parameters of **Fig. 11.3B**. Suppose the observations are $\mathbf{s}=(0.9,6.1,-0.2)$. Calculate the posterior over L.
c) Assume that $p(L)$ is uniform. Assume moreover that $\sigma_T<\sigma_D$, as is reasonable in a search task (targets are usually more narrowly defined than distractors). Show that MAP estimation then amounts to choosing the location L for which s_L is closest to $\frac{J_T\mu_T-J_D\mu_D}{J_T-J_D}$.
d) Assume that $p(L)$ is uniform and use the parameters of **Fig. 11.3B**. Vary N from 1 to 8. For each value of N, calculate proportion correct. Plot proportion correct as a function of N.
e) Explain intuitively why in this model, proportion correct decreases as a function of N.

Problem 11.5 Derive Eq. (11.16) starting from Eq. (11.13).

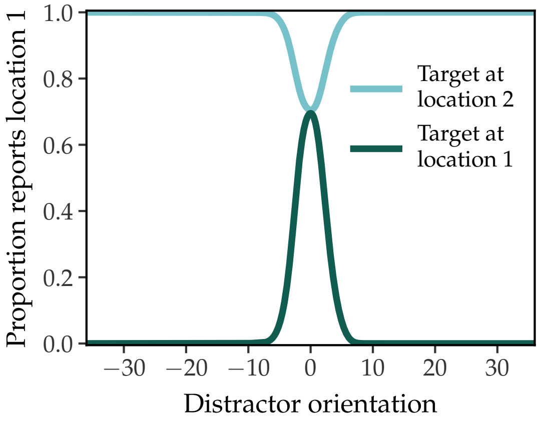
Problem 11.6 In this chapter, we considered tasks in which the distractors are independent. In reality, they might not be. For example, if oriented line segments are part of a textured background, they will tend to point in the same direction. This matters for the decision rule and observer performance. Here, we consider an extreme form of non-independence, namely that all distractors are identical to each other. However, their common stimulus value, which we denote by s_D, still varies from trial to trial, following a distribution $p_{\text{distractor}}(s_D)$. Further assume a distribution over target location $p(L)$, a target distribution $p_{\text{target}}(s)$, and measurement distributions $p(x_i|s_i)$ Derive an expression for the posterior over target location.

Problem 11.7 Assume $N=2$, $p(L=1)=p_1$, independent distractors, a target that always has value 0, a distractor that has a normal distribution with mean 0 and variance σ_D^2, and measurement noise with variance σ^2. The observer has to localize the target.

a) Show that the Bayesian observer reports location 1 when

$$x_1^2-x_2^2<\frac{2\log\frac{p_1}{1-p_1}}{\frac{1}{\sigma^2}-\frac{1}{\sigma_d^2+\sigma^2}}. \tag{11.46}$$

b) Assume $p_1=0.6$, $\sigma_D=10$, $\sigma=2$. By using a grid over x_1 and x_2, numerically calculate and plot the following psychometric curves, which show the proportion of "location 1" reports as a function of the stimulus value of the distractor.



c) Repeat using sampling of x_1 and x_2 instead of a grid.

Problem 11.8 How would Eq. (11.30) change if in the $C = 1$ condition, instead of there always being exactly one target, each stimulus individually is a target with probability ε?

Problem 11.9 In this problem, we examine the distributions of the log posterior ratio. An observer has to detect whether a target orientation is present or absent among N oriented line segments. The target has orientation 0° and each distractor has orientation 10°. On each trial, the experimenter chooses whether the target will be present or absent with equal probability. When the target is present, each stimulus is equally likely to be the target. Assume that the measurement at each location is corrupted by Gaussian noise with standard deviation 5°.

a) Suppose $N = 2$. Simulate the measurement on 5,000 target-present trials and 5,000 target-absent trials. For convenience and without loss of generality, you can assume that when the target is present, it is at the first location.

b) On each simulated trial, compute the log posterior ratio.

c) Plot the resulting histograms of the log posterior ratio, one for target present and one for target absent, in the same plot. Is either histogram symmetric?

d) Assume the observer performs MAP estimation. Calculate proportion correct.

e) Repeat your simulation for set sizes from 1 to 20. Plot proportion correct as a function of set size.

Problem 11.10 In this problem, we study search with a circular world state variable Read Section B.7.6 for background on the Von Mises distribution. An observer detects whether a target, defined by orientation, is present among N line segments. The target always has orientation s_T. Each distractor orientation is drawn independently from a uniform distribution on $[0, \pi)$. The observation at the i^{th} location, x_i, is drawn from a Von Mises distribution with circular mean s_i (the true orientation) and concentration parameter κ,

$$p(x_i|s) = \frac{1}{\pi I_0(\kappa)} e^{\kappa \cos 2(x_i - s_i)}, \tag{11.47}$$

where I_0 is the modified Bessel function of the first kind of order 0. The prior probability that the target is present is 0.5.

a) Show that the Bayesian MAP observer responds "target present" if

$$\sum_{i=1}^{N} e^{\kappa \cos 2(x_i - s_T)} > N I_0(\kappa). \tag{11.48}$$

b) Simulate 10^5 trials with $N = 2$, $s_T = 0$, and $\kappa = 10$. On each trial, draw observations from the generative model. We want to point out that for drawing from Von Mises distributions there are likely to be implementations already in your favorite programming language. Then compute the log posterior ratio of target presence (LPR).

c) Plot the empirical distributions of this LPR when the target is present and when it is absent (using a smooth curve might be better for presentation than using histograms).
d) Plot the receiver operating characteristic (ROC).
e) Repeat parts (c) and (d) for two different set sizes, $N = 4$ and $N = 8$. Plot LPR distributions and ROCs in a way that you can easily compare across N. Interpret the effects of N.

Problem 11.11 Derive the log likelihood ratio at set size N in each of the following visual search scenarios. Assume independent Gaussian noise.

a) The target is always vertical. Distractors are homogeneous but their value is drawn on each trial from a uniform distribution over orientation. The observer reports whether the target is present.
b) The target is always vertical. Distractors are drawn from a Gaussian distribution around vertical with variance σ_D^2. The observer reports whether the target is present. (This was done by Benjamin Vincent.)
c) The target is always vertical. Each distractor is independently chosen to be tilted an amount Δ to the left or to the right of vertical. The observer reports whether the target is present.
d) Distractors are always vertical. The target is drawn on each trial from a Gaussian distribution around vertical with variance σ_T^2. The observer reports whether the target is present.
e) The target is drawn from a symmetric distribution around 0. Distractors are all vertical. The observer reports whether the target is tilted to the right or left of vertical. (This task has been studied extensively by Stefano Baldassi and colleagues.)
f) Distractors are homogeneous but their value is drawn on each trial from a uniform distribution over orientation. The target, if present, has a value such that the target-distractor difference is Δ on each trial. The observer reports whether the target is present.
g) The target is always vertical and always present. Distractors are drawn from a uniform distribution. The observer reports which location contained the target.

Problem 11.12 This problem builds on the detection version of the camouflage scenario in Section 11.5, and in particular on Eq. (11.35) for the likelihood ratio LR of the fish being present versus absent.

a) Calculate the numerical value of LR for each panel and verify your answers against the numbers in **Fig. 11.4**.
b) Vary b from 0.1 to 0.7 in steps of 0.1. Instead of assuming the specific observations (combinations of fish plus riverbed) in **Fig. 11.4**, we now simulate, for every value of b, 10,000 observations randomly generated with that value of b (and still $a = 0.7$). Represent each observation by a vector of ten n_L values. Calculate the mean and standard deviation of $\log \mathrm{LR}$ as a function of b, and plot these values with error bars whose sizes represent standard deviations.
c) Modify the derivation leading up to Eq. (11.35) to treat the case of a species of smaller fish of 3 units long. Hint: you will need to marginalize over both the horizontal and vertical coordinates.
d) Repeat (b) for this new expression and plot in the same plot with a different color.
e) Based on the plot in (d), argue that the visibility of the smaller fish is less affected by suboptimal camouflage.

Problem 11.13 In target classification, a target is present, but its location is unknown. The observer has to classify the target. Download Shen and Ma (2016), *A detailed comparison of optimality and simplicity in perceptual decision making*. Read the first 4 pages, up to "Simple heuristic models", and the sections "Model predictions" and "Model fitting" on pages 5 and 6. In this task, subjects judge whether a target stimulus is tilted leftward (counterclockwise) or rightward (clockwise) with respect to vertical. The only way the target differs from the distractors is that the distractors are identical to each other on a given trial, whereas the target value is drawn independently. We consider

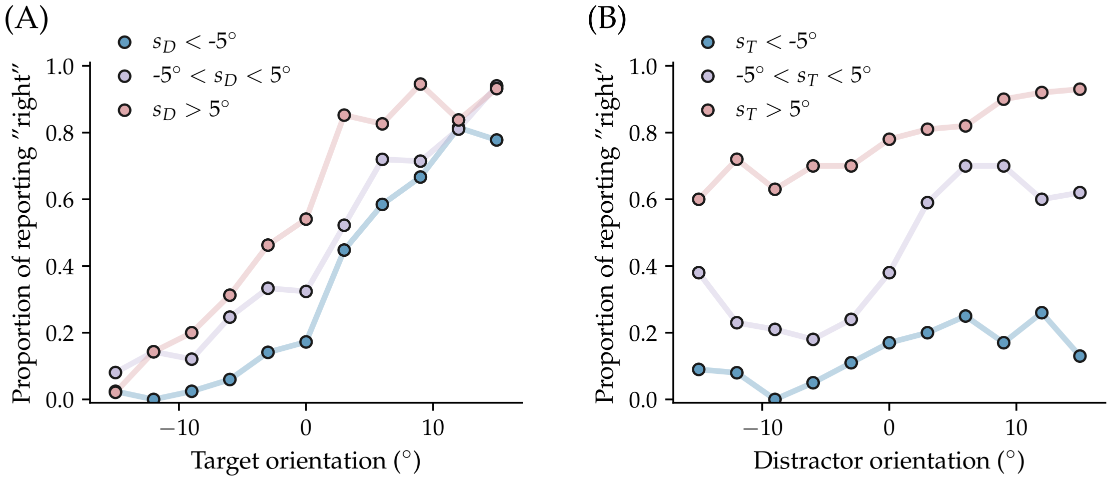


Figure 11.6: Visual search data from one subject from Shen and Ma (2016). (A) Proportion of reporting "target tilted right" as a function of (binned) target orientation. (B) Same as a function of (binned) distractor orientation. In the problem, we fit the entirety of these data simultaneously.

the optimal Bayesian model described in the paper. This model has two parameters, sensory noise level σ and the lapse rate λ.

a) For $\sigma = 1$ and $\lambda = 0$, calculate, separately for each trial, the probability that the observer will report "target tilted right" through Monte Carlo simulation. For each trial, draw 1,000 measurement vectors $\mathbf{x}$.
b) Define a fine grid for σ and λ. Repeat part (a) for every parameter combination on this grid. Save the results in a 3d matrix, with one dimension corresponding to σ, one to λ, and one to trials.

Download ch11_visual_search.csv from https://osf.io/84kpb/. These are data from one subject in the Shen and Ma experiment. The matrix in the file has 2,000 rows (trials) and 3 columns. The first column is the target orientation, the second the distractor orientation, the third the subject's response (-1 for leftward, 1 for rightward).

c) Use the result of (b) to calculate the log likelihood of the subject responses.
d) Find the maximum-likelihood estimates of σ and λ on the grid.
e) Using these parameter estimates, reproduce the two plots in **Fig. 11.6**. The model curves could differ slightly from the figure, depending on the grid you choose and due to sampling noise.

12. Inference in a changing world

How do we estimate the current state of a changing world?

In all previous chapters, we considered world states that did not change. The real world, however, is often changing, and the observer typically wants to estimate the world state at the current time.

Plan of the chapter

Here, we work the Bayesian model out in detail for two common tasks with a changing world state: one in which the world state is continuously changing in a lawful manner (tracking or Kalman filtering), the other one in which changes occur at discrete moments (change point detection).

12.1 Tracking a continuously changing world state

Observers often have to do inference while the world is continuously changing. For example, when you are playing a ball sport, you want to know where the moving ball is now (or arguably a little bit into the future). When you're trying to find the light switch in a dark room, your own hand moves and you want to know where it is now (relative to the light switch). When you're trying to understand a spoken sentence, the meaning evolves while you gather auditory information. In these examples, the world changes in a continuous and lawful manner. Here, we will examine how such changes affect the inference done by a Bayesian observer.

12.1.1 Step 1: Generative model

To make the math simpler, we will discretize time: time moves in unit steps, e.g. seconds. We will denote time by t; thus, t is an integer, and we will make it start at $t = 1$. The generative model is shown in **Fig. 12.1**. The underlying world state s evolves while there is an observation x at each point of time

The top row shows the world state (stimulus), which now changes across time (columns). We denote by s_t the world state at time t. The bottom row contains the measurements at the different time points. We denote by x_t the measurement at time t. This generative model is different from

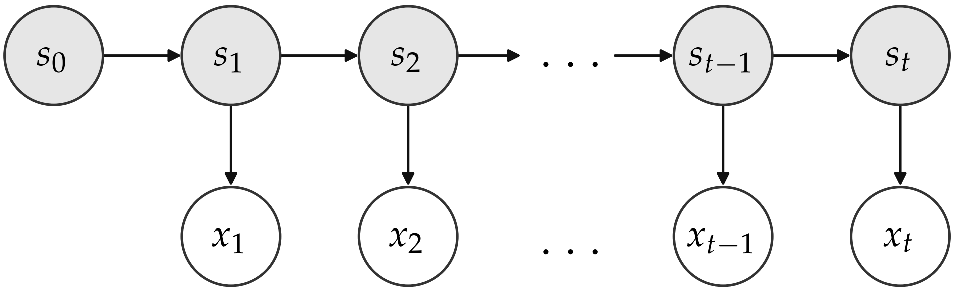


Figure 12.1: Generative model for a world state that changes continuously in time. The subscript refers to time.

Section 5.5, where the observer was making a sequence of measurements, but the world state did not change.

The generative model contains two important assumptions. First, we assume that measurements are conditionally independent given the respective world states. This is a common assumption that we already encountered in Chapter 10. Formally, we can write about the probability of all the measurements(**x**) over time and all the states (**s**) over time.

$$p(\mathbf{x}|\mathbf{s}) = \prod_{i=1}^{t} p(x_i|s_i). \tag{12.1}$$

As we typically do for examples in this book, we assume that the measurement follows a normal distribution:

$$p(x_i|s_i) = \mathcal{N}(x_i; s_i, \sigma^2). \tag{12.2}$$

The observation variance, σ^2, is the same for each measurement and thus does not depend on t. Second, in the top row, we only drew arrows from the world state at one time step to the world state at the next time step. There are no arrows that bridge states more than one time step apart. Formally, we can therefore write

$$p(\mathbf{s}) = p(s_0)p(s_1|s_0)p(s_2|s_1)\cdots p(s_t|s_{t-1}) \tag{12.3}$$

$$= p(s_0)\prod_{i=1}^{t} p(s_i|s_{i-1}) \tag{12.4}$$

The conditional probability $p(s_i|s_{i-1})$ is called the *state transition model*. It describes the dynamics of the world: how much we expect the state at time i to have a value s_i given that the state at time $i-1$ had value s_{i-1}. The assumption that the distribution of s_i only depends on s_{i-1}, and not on s_{i-2}, s_{i-3}, etc. is called the *Markov property*, and the corresponding generative model for $s_1, s_2, \ldots s_T$ (top row of **Fig. 12.1**) is called a Markov model or, more precisely, a first order Markov process. In a sense, a Markov process has *no memory*: the expectation for what happens next only depends on the current state, not on how the current state was arrived at.

Definition 12.1.1 — Markov process. A Markov process $s_1, s_2, \ldots, s_T$ is defined by the property $p(s_t|s_1, \ldots, s_{t-1}) = p(s_t|s_{t-1})$.

Physical systems that obey Newton's laws of motion are Markov processes, since those laws can be stated as *differential equations*. For example, to predict the trajectory of a ball, we only need to know the current position and velocity, not its past positions or velocities.

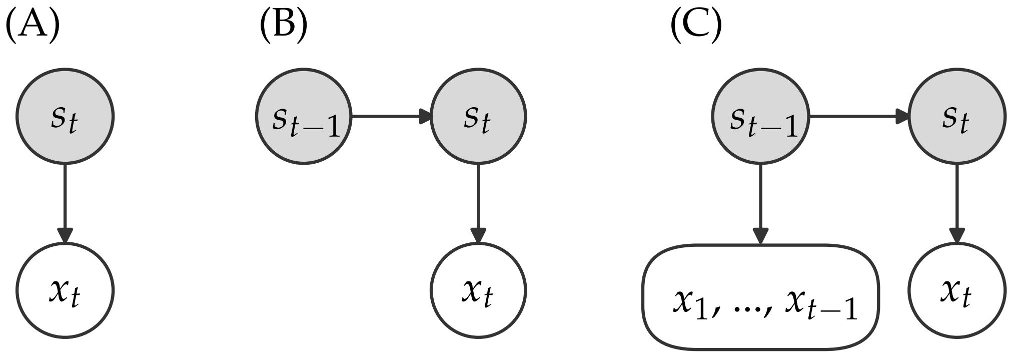


Figure 12.2: Simplified generative models in the problem of tracking a continuous change.

We now have to make specific distributional assumptions. Throughout this chapter, we will only consider the simplest possible state transition model: that the state increases by a fixed value Δ from time step to time step: $s_i = s_{i-1} + \Delta$. However, we do allow for noise. We assume that this noise is Gaussian and has standard deviation σ_s, so that

$$p(s_i|s_{i-1}) = \mathcal{N}(s_i; s_{i-1} + \Delta, \sigma_s^2). \tag{12.5}$$

This could, for example, describe linear motion of your hand, at constant velocity but with some variability. Finally, we will assume a form for the prior at time 0, $p(s_0)$ in Eq. (12.4). We will assume a Gaussian with mean μ_0 and standard deviation σ_0:

$$p(s_0) = \mathcal{N}(s_0; \mu_0, \sigma_0^2). \tag{12.6}$$

This concludes Step 1. Since the states $s_1, s_2, \ldots$ are not directly observable, the generative model we defined here is called a *hidden Markov model.*

12.1.2 Step 2: Inference

The observer is interested in inferring the current state, s_t, based on the entire time series of measurements $x_1, x_2, \ldots, x_t$. The Markov property makes this simpler. In particular, we will be able to write down a recursive relation for the posterior distribution: the posterior over s_t given $x_1, x_2, \ldots, x_t$ can be expressed in terms of the posterior over s_{t-1} given $x_1, x_2, \ldots, x_{t-1}$. In the following subsections, we gradually build up the logic of the inference process.

Suppose first that the only world state is the current one, s_t, and a corresponding measurement x_t (**Fig. 12.2A**). The world state s_t obeys the distribution $p(s_t)$, which the observer uses as a prior. Then we are in the same situation as in Chapter 3. The posterior over s_t is

$$p(s_t|x_t) \propto p(x_t|s_t)p(s_t). \tag{12.7}$$

If there was only one state we would just need to use Bayes' rule to calculate the posterior distribution. Now we add the previous world state, s_{t-1} to the generative model (**Fig. 12.2B**). Its distribution is $p(s_{t-1})$. Then the prior over s_t is not readily available to the observer, but has to be obtained through marginalization over s_{t-1}:

$$p(s_t) = \int p(s_t|s_{t_1})p(s_{t-1})ds_{t-1}. \tag{12.8}$$

This expresses that the observer's belief about the current state, s_t, is obtained by "extrapolation" or "prediction" from the previous state, s_{t-1}. This extrapolation or prediction is captured by the conditional probability distribution $p(s_t|s_{t-1})$. If from one time step to the next, noise is added to the state, the marginalization has the effect of widening the distribution.

Specifically, if $p(s_{t-1})$ is a normal distribution with mean $\mu_{\text{post},t-1}$ and variance $\sigma^2_{\text{post},t-1}$, and we assume Eq. (12.5) for the state transitions, then $p(s_t)$ will also follow a normal distribution with mean $\mu_{\text{post},t-1} + \Delta$ and variance $\sigma^2_{\text{post},t-1} + \sigma_s^2$: the distribution is shifted over by Δ and is wider.

Substituting Eq. (12.8) into Eq. (12.7), we find

$$p(s_t|x_t) \propto p(x_t|s_t) \int p(s_t|s_{t-1}) p(s_{t-1}) ds_{t-1}. \tag{12.9}$$

In the special case of a static world, in which $s_t = s_{t-1}$, we have $p(s_t|s_{t-1}) = \delta(s_t - s_{t-1})$, and Eq. (12.9) reduces to Eq. (12.7).

In Eq. (12.9), we assumed that the observer uses a prior distribution $p(s_{t-1})$. In reality, though, instead of a *prior* distribution at the previous time, we have a *posterior* distribution, $p(s_{t-1}|\text{ past})$, where “past" is the sequence of all previous measurements, $x_1, x_2, \ldots, x_{t-1}$. We can explicitly add those in the generative model (**Fig. 12.2C**). Eq. (12.9) changes by conditioning on the past measurements on both sides. This gives

$$p(s_t|x_t) \propto p(x_t|s_t) \int p(s_t|s_{t_1}) p(s_{t-1}|x_1, \ldots, x_{t-1}) ds_{t-1}. \tag{12.10}$$

This is a *recursive* relationship: the posterior over the state at time t given the measurements through time t is expressed as a function of the posterior over the state at time $t-1$ given the measurements through time $t-1$.

In the special case of a static world, in which $s_t = s_{t-1}$, we have $p(s_t|s_{t-1}) = \delta(s_t - s_{t-1})$, Then Eq. (12.10) becomes

$$p(s_t|x_t) \propto p(x_t|s_t) p(s_t|x_1, \ldots, x_{t-1}), \tag{12.11}$$

which is equivalent to evidence accumulation as discussed in Section 5.5, specifically Eq. (5.25).

If we again postulate that the posterior at time $t-1$ is normal with mean $\mu_{\text{post},t-1}$ and variance $\sigma^2_{\text{post },t-1}$, then the “prior" at time t is normal with mean

$$\mu_{\text{prior},t} = \mu_{\text{post},t-1} + \Delta \tag{12.12}$$

and variance

$$\sigma^2_{\text{prior},t} = \sigma^2_{\text{post},t-1} + \sigma_s^2. \tag{12.13}$$

Moreover, the posterior at time t, $p(s_t|x_1, \ldots, x_t)$, is also normal, with mean

$$\mu_{\text{post},t} = \frac{\frac{x_t}{\sigma^2} + \frac{\mu_{\text{prior},t}}{\sigma^2_{\text{prior},t}}}{\frac{1}{\sigma^2} + \frac{1}{\sigma^2_{\text{prior},t}}} \tag{12.14}$$

and variance

$$\sigma^2_{\text{post},t} = \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma^2_{\text{prior},t}}} \tag{12.15}$$

Exercise 12.1 Derive these equations. You will need Eqs. (12.2) and (12.5). ■

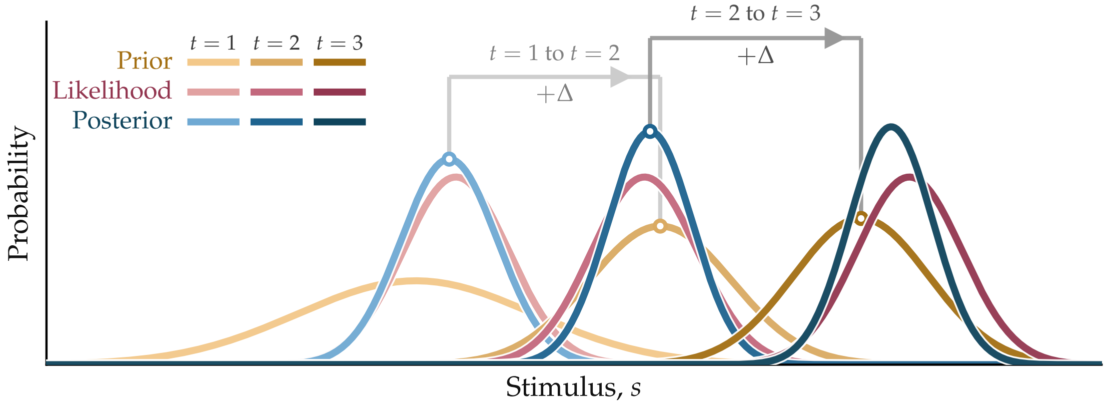


Figure 12.3: Evolution of the posterior over a stimulus that changes at a constant rate. Shown are the prior (yellow), likelihood (red), and posterior (blue) at three subsequent time steps, $t = 1, 2, 3$. At each time point, the posterior is calculated just like in Chapter 3. The posterior becomes the prior at the next time step through a shift (by Δ and a widening, per Eqs. (12.12) and (12.13)).

These equations show an interesting combination of combining a measurement with a prior (Chapter 3) and that prior widening, as discussed above. These "forces" are counteracting: the former will make the posterior narrower, the latter wider. We see this in Fig. 12.3

Eventually, the variance of the posterior (and thus uncertainty) will asymptote at a fixed value (see Problem 12.6).

Eqs. (12.14) and (12.15) define the posterior at time t in terms of the posterior of time $t-1$. However, we have to specially consider what happens at $t = 1$, because at $t - 1$ we don't have a posterior, only a prior, Eq. (12.6). However, that prior serves as the posterior: we can simply write

$$\mu_{\text{post},0} = \mu_0 \tag{12.16}$$

$$\sigma^2_{\text{post},0} = \sigma^2_0 \tag{12.17}$$

These equations define what can be called the "initial condition".

To conclude Step 2, we simply recall that the estimate of a continuous world state that minimizes expected squared error is the posterior mean, which in our case is given by Eq. (12.14).

In this section, we have considered an observer who infers the present state, s_t. However, the formalism can be extended to prediction of a future state. This is simply done by removing the current measurement, x_t, from the generative model and accordingly from Eqs. (12.9) and (12.10). Then, those equations describe the probabilistic prediction for the future state s_t from past measurements $x_1, \ldots, x_{t-1}$.

12.2 Change point detection

So far, we have examined continuous change of a stimulus. Sometimes, however, there is a discontinuous change or "jump" in the world state. For example, the ownership of a restaurant might change hands, causing the quality of the food to suddenly improve. A neurologist may want to detect a seizure on an EEG in a comatose patient (who does not exhibit any outwardly visible symptoms of a seizure). Or a friend might have experienced a life event that drastically changes the nature of their interactions with you.

These situations require a different treatment. We will examine two cases: a) the observer knows that a single change occurred over the entire observable time period, but has to determine

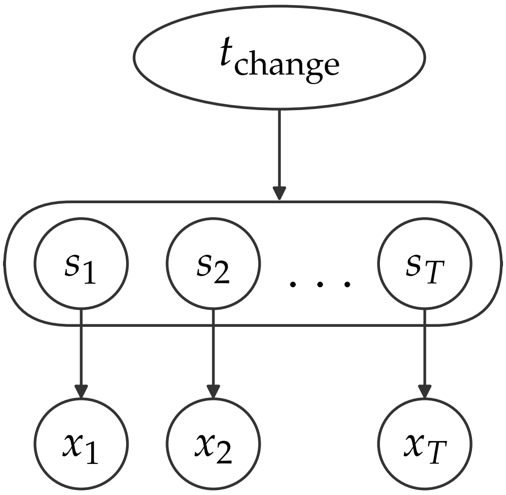


Figure 12.4: Generative model of change point detection with a single change point.

when (this section); b) a change could independently have occurred at any time point during the observable time period (next section).

12.2.1 Single change point

Intuitively, change point detection can be hard because the noise in the measurements can sometimes create large "apparent changes" that are not due to a change in the underlying state. The Bayesian observer solves this problem in an optimal manner.

Step 1: Generative model. The generative model is shown in **Fig. 12.4**. We assume that every time point between $t = 1$ and $t = T$ has the same probability of being the change point:

$$p(t_{\text{change}}) = \frac{1}{T}. \tag{12.18}$$

Next, we define what it means to have a change point at t_{change}:

$$s_t = \begin{cases} s_{\text{pre}} & \text{for } t < t_{\text{change}} \\ s_{\text{post}} & \text{for } t \geq t_{\text{change}} \end{cases}, \tag{12.19}$$

where s_{pre} and s_{post} are fixed and assumed known to the observer. Finally, we make the usual assumption about independent and normally distributed measurements. We additionally assume that the noise level is the same for all measurements:

$$p(\mathbf{x}|\mathbf{s}) = \prod_{t=1}^{T} p(x_t|s_t) \tag{12.20}$$

$$= \prod_{t=1}^{T} \mathcal{N}(x_t; s_t, \sigma^2). \tag{12.21}$$

Step 2: Inference. The world state of interest is change point time t_{change} t_{change}. Therefore, the Bayesian observer computes the posterior over t_{change} given a sequence of measurements, $\mathbf{x}$. We first apply Bayes' rule with a uniform prior:

$$p(t_{\text{change}}|\mathbf{x}) \propto p(\mathbf{x}|t_{\text{change}})p(t_{\text{change}}) \tag{12.22}$$

$$\propto p(\mathbf{x}|t_{\text{change}}). \tag{12.23}$$

Then, we marginalize over $\mathbf{s}$:

$$p(t_{\text{change}}|\mathbf{x}) \propto \sum_{\mathbf{s}} p(\mathbf{x}|\mathbf{s})p(\mathbf{s}|t_{\text{change}}). \tag{12.24}$$

Here, the sum is over all sequences $\mathbf{s}$. There are a total of 2^T such sequences, but given t_{change}, only one is allowed, as specified by Eq. (12.19) . Then, the sum reduces to a single term, which becomes

$$p(t_{\text{change}}|\mathbf{x}) \propto \left(\prod_{t=1}^{t_{\text{change}}-1} p(x_t|s_t = s_{\text{pre}})\right)\left(\prod_{t_{\text{change}}}^{T} p(x_t|s_t = s_{\text{post}})\right). \tag{12.25}$$

This can be written as

$$p(t_{\text{change}}|\mathbf{x}) \prod_{t_{\text{change}}}^{T} \frac{p(x_t|s_t = s_{\text{post}})}{p(x_t|s_t = s_{\text{pre}}} \tag{12.26}$$

Exercise 12.2 Show this.

So far, we have not used the form of the measurement distribution. Using that, the posterior over t_{change} becomes

$$p(t_{\text{change}}|\mathbf{x}) \propto \prod_{t_{\text{change}}}^{T} e^{\frac{\Delta s}{2\sigma^2}\sum_{t_{\text{change}}}^{T}(x_t-\bar{s})}, \tag{12.27}$$

where $\Delta s = s_{\text{post}} - s_{\text{pre}}$ and $\bar{s} = \frac{s_{\text{pre}}+s_{\text{post}}}{2}$, by analogy with the discrimination tasks in Chapter 7. Eq. (12.27) has an intuitive explanation: the evidence for a change point at t_{change} increases the more measurements following t_{change} are on the same side of the mean as s_{post}. However, the evidence decreases if one goes so far back in time that one includes measurements coming from the original s_{pre}. To obtain the actual posterior probabilities, compute the value of the right-hand side of Eq. (12.27) for all values of t_{change}, and divide them by their total (normalization).

The inference concludes with a read-out stage. Since t_{change} is discrete, MAP estimation (picking the mode of the posterior, in order to maximize accuracy) is the most obvious strategy. MAP estimation amounts to maximizing $\sum_{t_{\text{change}}}^{T}(x_t - \bar{s})$.

Exercise 12.3 Why is this the case?

Step 3: Response distribution. The response distribution has to be simulated and we leave this to a Problem.

12.2.2 Random change points

We next consider the problem of change points randomly occurring and the observer inferring their times. The challenge of determining when an apparent change is a true change of the underlying stimulus is then harder.

Step 1: Generative model. The generative model is shown in **Fig. 12.5**. We assume that each C_t takes values 0 (no change) and 1 (change), that all C_t's are independent, and that $p(C_t = 1) = \varepsilon$. Thus, there can be anywhere between 0 and T change points. We next assume that each s_t takes values -1 and 1. If $C_t = 0$, s_t is equal to s_{t-1} while if $C_t = 1$, s_t has the opposite sign. We assume $s_0 = 1$. Then, to each vector $\mathbf{C}$ corresponds a specific vector of stimuli, which we will denote by $\mathbf{s}_{\mathbf{C}}$. We make the standard assumption about the measurements, Eq. (12.21).

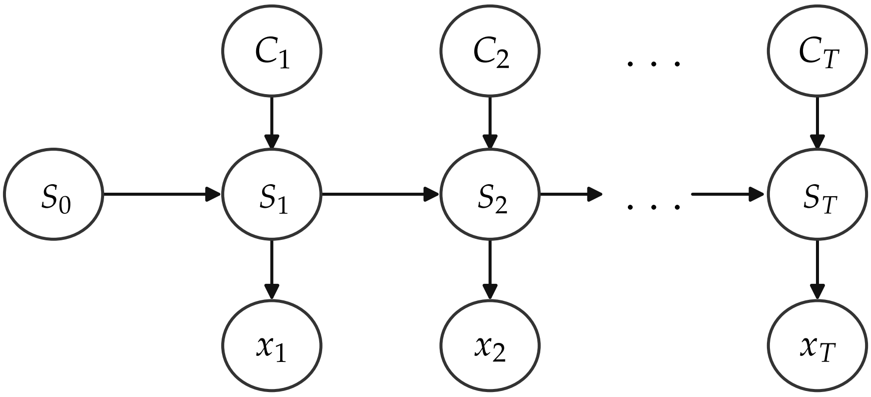


Figure 12.5: Generative model for random change point detection.

Step 2: Inference. Inference is now over the entire time series $\mathbf{C} = (C_1, \ldots, C_T)$. Since each C_T is binary and all are independent, the hypothesis space consists of 2^T binary vectors. The posterior is

$$p(\mathbf{C}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{C})p(\mathbf{C}). \tag{12.28}$$

Here, the prior over $\mathbf{C}$ cannot be assumed uniform, since the construction of the generative model makes some $\mathbf{C}$ more probable than others. To be precise, we have

$$p(\mathbf{C}) = \varepsilon^{||\mathbf{C}||}(1-\varepsilon)^{T-||\mathbf{C}||}, \tag{12.29}$$

where $||\mathbf{C}||$ is the total number of 1's in $\mathbf{C}$ [1].

Exercise 12.4 Why? ■

The likelihood function over $\mathbf{C}$ is

$$\begin{aligned} \mathscr{L}(\mathbf{C};\mathbf{x}) &= p(\mathbf{x}|\mathbf{C}) && (12.30)\\ &= \sum_{\mathbf{s}} p(\mathbf{x}|\mathbf{s})p(\mathbf{s}|\mathbf{C}) && (12.31)\\ &= p(\mathbf{x}|\mathbf{s}=\mathbf{s}_C), && (12.32) \end{aligned}$$

where we recall that $\mathbf{s}_C$ is the unique stimulus vector s that corresponds to the change point vector $\mathbf{C}$. We leave the rest of this model to a Problem.

12.2.3 More realistic change point detection

We made several simplifying assumptions in this section. We assumed that change points are either unique (one per trial) or random (independent across time points). Neither is realistic. In many real-world change detection problems, change points occur “every now and then", which means that there is some prior over the interval between two change points. Moreover, change point detection can usually not be done with the luxury of hindsight, i.e. based on a complete time series of observations; instead, a decision has to made instantly about whether a change occurred. Finally, the change is often not between two specific stimulus values but between two categories. None of

[1] The notation $||\cdot||$ is called the *norm* of $\mathbf{C}$. In this case, we are using the L_1-*norm*, which is simply the sum of the elements.

these aspects makes inference conceptually different from the examples discussed in this section; however, each can introduce substantial technical complications.

In view of such complications, and in particular considering how bad they get as T is large, change point detection is a domain in which exact Bayesian models quickly become implausible as models of human behavior. We will discuss this further in Chapter 15.

12.3 Summary and remarks

In this chapter, we have introduced methods to enable inference in a world that is constantly changing. We have learned:

- Inference often takes place in a changing world. To handle this, a Bayesian decision-maker needs an explicit model of how the world is changing.
- If the dynamics of the world are characterized by a Markov process, then the resulting update equations are tractable.
- Inference boils down to alternating between updating our belief given the temporal changes of the world and updating it given the observations.
- Interestingly, it is not possible to explicitly "solve for" the mean and variance of the posterior at each time. The recursive equations, combined with the initial condition, are usually the best we can do.
- We derived the recipe for optimal inference for a linear state transition model and Gaussian measurement noise. This recipe is also called a *Kalman filter*.
- We also derived recipes for inferring change points from time series of noise observations, in two cases: when there is only one change point in the entire time series, or when change points occur at random.

12.4 Suggested readings

- Ryan Prescott Adams and David JC MacKay. "Bayesian online changepoint detection". In: *arXiv preprint arXiv:0710.3742* (2007)
- J Yu Angela. "Adaptive behavior: Humans act as Bayesian learners". In: *Current Biology* 17.22 (2007), R977–R980
- Kathryn Bonnen et al. "Continuous psychophysics: Target-tracking to measure visual sensitivity". In: *Journal of Vision* 15.3 (2015), pages 14–14
- Daniel Goldreich and Jonathan Tong. "Prediction, postdiction, and perceptual length contraction: a Bayesian low-speed prior captures the cutaneous rabbit and related illusions". In: *Frontiers in psychology* 4 (2013), page 221
- Konrad P Kording, Joshua B Tenenbaum, and Reza Shadmehr. "The dynamics of memory as a consequence of optimal adaptation to a changing body". In: *Nature neuroscience* 10.6 (2007), pages 779–786
- Elyse H Norton et al. "Human online adaptation to changes in prior probability". In: *PLoS computational biology* 15.7 (2019), e1006681
- Kunlin Wei and Konrad Körding. "Uncertainty of feedback and state estimation determines the speed of motor adaptation". In: *Frontiers in computational neuroscience* 4 (2010), page 11
- Robert C Wilson, Matthew R Nassar, and Joshua I Gold. "Bayesian online learning of the hazard rate in change-point problems". In: *Neural computation* 22.9 (2010), pages 2452–2476
- Daniel M Wolpert. "Computational approaches to motor control". In: *Trends in cognitive sciences* 1.6 (1997), pages 209–216

12.5 Problems

Problem 12.1 The muscles in our bodies change over time. Sometimes they get stronger, e.g. after we work out, sometimes they get weaker. How would you formalize changing muscles, and inference about their strength. Do you think this is a real problem?

Problem 12.2 Change points where an aspect of the world undergoes a sudden change happen in many domains. Give an example of a case that we did not use as an example of where that would be important.

Problem 12.3 Formally derive the posterior $p(s_t|x_1,\dots,x_t)$ using the structure of the generative model. "Follow the arrows." No need to substitute any specific (e.g. Gaussian) distributions.

Problem 12.4 Consider the HMM in this chapter. Suppose the observer wants to make a prediction for the future state s_{t+1}, given the measurements $x_1,\dots,x_t$ (i.e., the measurement at time $t+1$ has not been made yet). Derive the mean and variance of the posterior. You can assume the mean and variance of $p(s_t|x_1,\dots,x_{t-1})$ to be known.

Problem 12.5 Assume again the HMM from above. Take $\mu_0 = 0$, $\sigma_0 = 1$, $\sigma_s = 1$, $\sigma = 2$, $\Delta = 1$, and t running from 1 to 30.

a) Draw a sequence of world states $s_0, s_1, \dots, s_t$ according to the distributions in the top row of the generative model.

b) Draw a sequence of corresponding measurements $x_1,\dots,x_t$ (no measurement at time 0).

We will now simulate an observer who uses the sequence of measurements you just drew to infer s_t at every time step.

c) Create a movie consisting of 30 frames in which the t^{th} frame shows:
   a. the posterior distribution over s_t
   b. a vertical dashed black line at the true s_t (from part (a))
   c. a vertical dashed blue line at the measurement x_t (from part (b)).

   Make sure that the scales of the axes remain fixed. Also label the axes.

d) Plot in a single plot, separately from the movie:
   a. the true world state as a function of time (black line)
   b. the measurement as a function of time (blue line)
   c. the posterior mean as a function of time (red line)

e) Plot in a separate plot the posterior standard deviation as a function of time.

Problem 12.6 A time series $y_1,\dots,y_t$ is said to *asymptote* if it monotonically increases or decreases, but as t grows very large, approaches a finite value.

a) Prove that in our treatment of the HMM, the variance of the posterior asymptotes at

$$\sigma^2_{\text{posterior}} = \frac{\sigma_s^2}{2}\left(\sqrt{1+\frac{4\sigma^2}{\sigma_s^2}}-1\right). \tag{12.33}$$

b) Explain intuitively *why* the standard deviation asymptotes.

Problem 12.7 Look up an external source (such as Wikipedia) on the dynamical systems model underlying the Kalman Filter. This is a description of a generative model that includes the generative model of this chapter as a special case. Here, we explore how.

a) Identify the variables in the dynamical systems models with variables or constants in our chapter.

b) Making use of the correspondences from part (a), simplify the prediction and update equations until you arrive at our two recursive equations Eq. (12.14) and Eq. (12.15). Keep in mind that our state estimate is the posterior mean.

Problem 12.8 An observer observes the time series (-0.46, 0.83, -3.26, -0.14, -0.68, -2.31, 0.57, 1.34, 4.58, 3.77). The observer knows that one change occurred and performs inference under the model in Section 12.2.1, with $s_{\text{pre}} = -1$, $s_{\text{post}} = 1$, and $\sigma = 1$. Calculate the observer's posterior distribution over change point time.

Problem 12.9 In the model of Section 12.2.1, assume $s_{\text{pre}} = -\mu$, $s_{\text{post}} = \mu$, and $\sigma = 1$. We expect that as μ grows larger, values of s before and after the change point become easier to distinguish. In this problem, we simulate this process.

a) Set $\mu = 1$. Randomly draw a change point. The change point determines the sequence of stimuli. Now draw a corresponding sequence of measurements. Then apply the MAP decision rule to this sequence, and record whether the MAP observer was correct or not.
b) In part (a), you did a single Monte-Carlo simulation. Now, do 10,000 Monte Carlo simulations for the same μ. Calculate proportion correct across all simulations.
c) Repeat parts (a-b) for all values of μ between 0 and 2 in steps of 0.1. Plot proportion correct as a function of μ.

Problem 12.10 In the model of Section 12.2.1, we assumed that s_{pre} and s_{post} were known. Derive the decision rule in the case when one of these takes the value s_-, the other the value s_+, but the observer does not know which is which. In other words, the change could be from s_- to s_+ or vice versa.

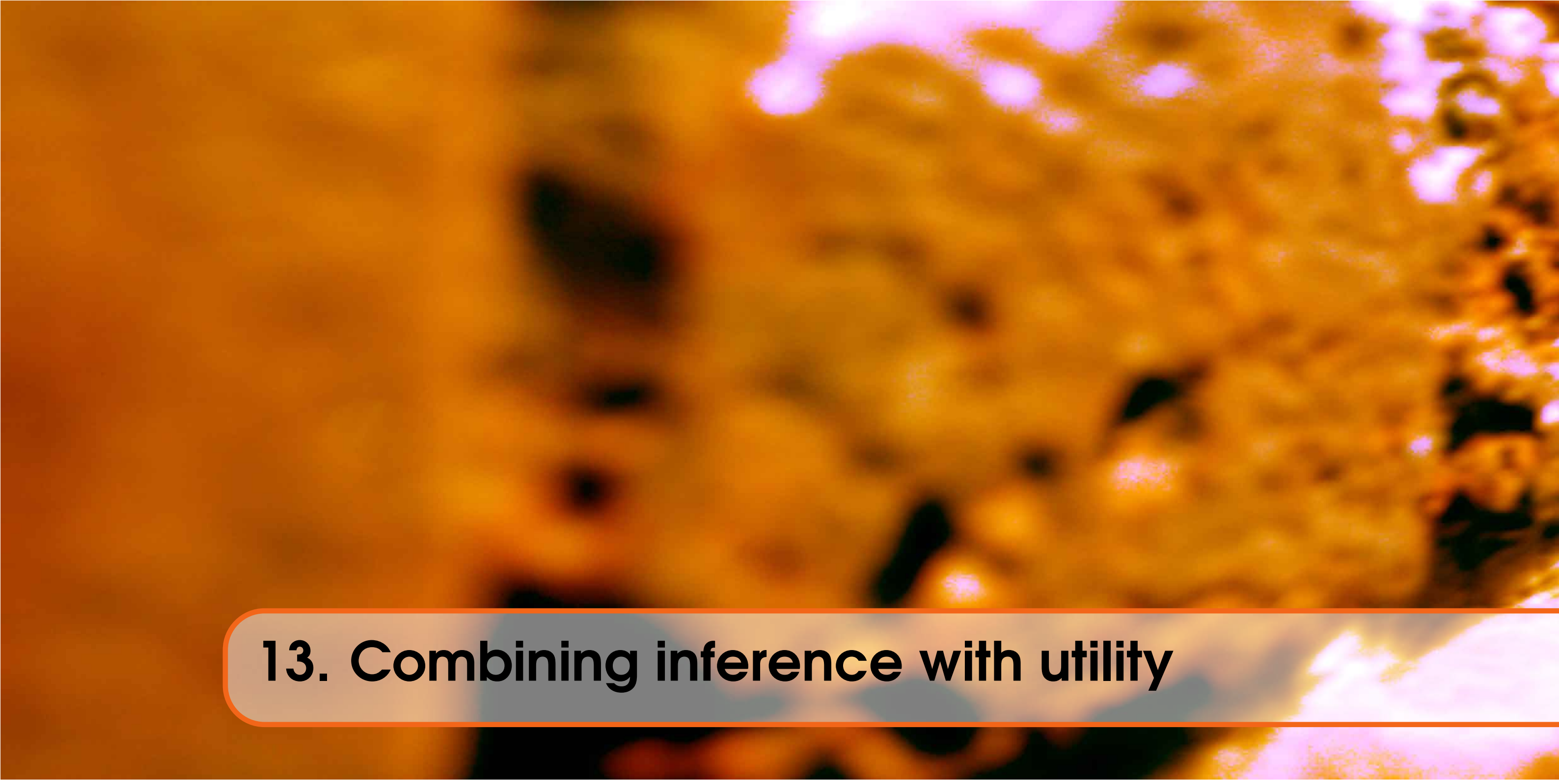
Problem 12.11 In the model of Section 12.2.2, we assumed that s_t could take values -1 and 1 and changed sign when $C_t = 1$. Derive the decision rule in the case when s is drawn from a normal distribution with standard deviation σ_s and a mean μ_t that can take values -1 and 1 and that changes sign when $C_t = 1$.

Problem 12.12 You observe the time series $\mathbf{x}$ = (-0.25, -1.66, -0.34, -0.41, -0.55, -1.88, -2.63, -0.79, 1.54, 0.85, 2.12, 1.22, -0.85, -0.61, -1.14).

a) Using the model in Section 12.2.2, with $s_0 = 1$, $\varepsilon = 0.3$, and $\sigma = 1$, calculate the posterior distribution over the change point vector $\mathbf{C}$ (save as a .mat file but don't bother to 2^15 values).
b) How many $\mathbf{C}$ vectors have a posterior probability greater than 1%?
c) What is the MAP estimate of $\mathbf{C}$?
d) Plot in a single plot $\mathbf{x}$ as a function of time as dots connected by lines, and the MAP estimate of $\mathbf{C}$ as a set of dashed vertical lines at the change point times). Does your inference look plausible?

Problem 12.13 Simulate 1000 synthetic trials from the model in Section 12.2.2, with $T = 10$, $s_0 = 1$, $\varepsilon = 0.3$, and $\sigma = 1$.

a) With what frequency does the MAP estimate of $\mathbf{C}$ have the same number of changepoints as the true $\mathbf{C}$?
b) Repeat part (a) for values of σ on a grid from 0.1 to 2 in steps of 0.1. Plot the frequency as a function of σ.



13. Combining inference with utility

How can we make optimal decisions when potential rewards and costs are involved?

So far, we have only talked about perception. We have modeled it as a process in which the observer generates a posterior distribution over the world state variable of interest, then produces an estimate in a way that is as close as possible to the truth (for instance to maximize expected accuracy, which leads to MAP estimation, or to minimize expected squared error, which could lead to posterior mean estimation). In real life, however, we often do not simply estimate world states; instead, we perform actions that have consequences, which in turn lead to rewards or costs. A few examples:

- Deciding when to cross a road: you are not only computing a posterior distribution over the speeds of oncoming cars, but you are also combining this information with the utility that you derive from saving time, and the negative utility associated with ending up in the hospital or causing an accident.
- Deciding whether to greet a person approaching you: you are not only computing a posterior distribution over the familiarity or identity of the person, but you are also weighing the awkwardness of ignoring someone you know and the other awkwardness (in some locales) of greeting a stranger.
- Deciding whether to buy travel insurance: you are not only computing the posterior probability that something will go wrong on your trip, but also the cost of the insurance and the cost of having to pay out of pocket when not insured.
- Deciding whether to drink old milk: you are not only computing the posterior probability that the milk has gone bad, but also the time, effort, and financial costs of buy new milk, as well as the cost of falling sick after drinking milk gone bad.
- A physician often has to decide whether to order a bothersome or expensive medical test to rule out a low-probability, but very serious diagnosis.
- Suppose you are hiking in foggy weather on a mountain trail with a cliff on your right side. From your limited visual information, you can estimate your distance from the cliff with some uncertainty. You then need to decide where along the trail to walk, choosing from among a continuum of possible headings. The center of the path may be most comfortable

on your feet. Making an error by veering off towards the left is relatively harmless; perhaps the path becomes rockier. An error to the right, in contrast, could be fatal. Because the costs are asymmetric, the optimal decision will be to bias your position towards the side of the path that is farther from the cliff.

These examples illustrate that the computation of a posterior probability distribution is a first step in a decision process that also takes into account costs and rewards. The theory of taking costs and rewards in Step 2 of the Bayesian modeling process is also called *Bayesian decision theory.*

The scientific fields of optimal control, economics and decision theory each make use of cost functions in one form or another. In economics and decision theory, researchers often specify a *utility function*. This may be related to economics' traditional focus on goods versus effort: the higher the utility, the better for the actor. In optimal control, researchers usually specify a *cost* or *loss function*. This may be related to the field's focus on minimizing the energetic cost of producing movements or the fuel costs of rockets. However, in each of the fields, there are positive and negative factors contributing to the function. Ultimately the two formulations are equivalent, as cost can be thought of as negative utility.

Plan of the chapter

We begin the chapter by showing how the optimal decision depends on our probability distribution over world states and on our preferences over outcomes. We first consider how to make an optimally binary decision (e.g., should I take my umbrella or not). We next consider how to choose among several actions (e.g., where should I search for my lost keys?) or even a continuum of actions. Finally, we use Bayesian decision theory to revisit the deceptively simple process of deciding which value to read-out from a posterior distribution. We will learn that taking the MAP estimate (mode of the posterior) is not generally the optimal decision. Depending on our cost function – i.e., the penalty or reward associated with different outcomes – it may be optimal to read out the posterior mean, median, mode, or another value.

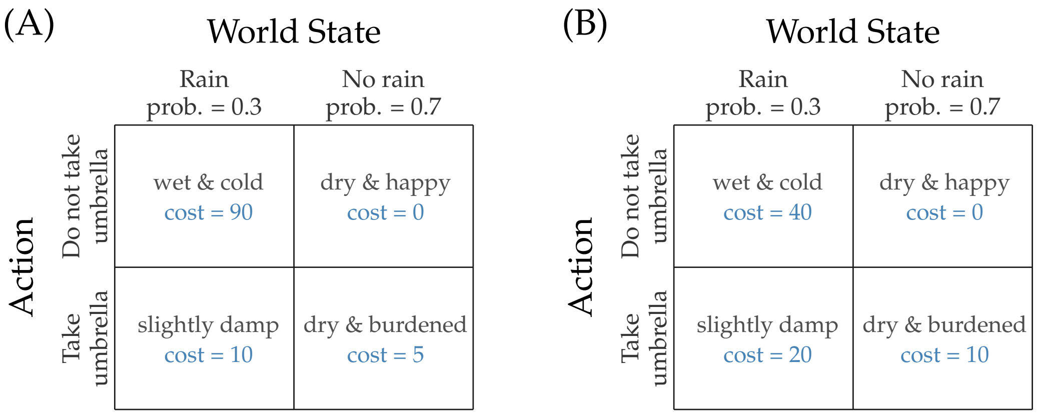
13.1 Examples

13.1.1 Deciding between two actions

Suppose that, as you prepare to leave home, you wonder whether it will rain. Combining a quick assessment of the cloudy sky (visual data) with knowledge of weather patterns in your area, you estimate the posterior probability of rain at 30%. Your posterior distribution, $p(\text{it will rain} \mid \text{visual data, background knowledge}) = 0.3$ and $p(\text{it will not rain} \mid \text{visual data, background knowledge}) = 0.7$, represents your belief about the world state of interest, but it does not dictate how you should *behave*. You need to make a decision: should you carry an umbrella or not?

It might at first seem that, since you believe that an upcoming rainfall is less likely, you should simply leave your umbrella at home. However, it should become clear upon reflection that your decision whether to carry an umbrella will be based not only on your estimate of the chance of rain but also on the value you attach to different possible *outcomes* that could result from your choice of action. If you decide not to carry the umbrella, and it rains, then you will suffer the undesirable consequence of becoming wet. On the other hand, if you decide to carry the umbrella, and it does not rain, then you may feel inconvenienced by holding the unnecessary umbrella. An outcome may be undesirable, in which case we associate it with a *cost*, or desirable, in which case we associate it with a *utility*. **Fig. 13.1A** illustrates the costs, specified by one individual, of the four possible outcomes in the umbrella problem.

Under the framework of Bayesian decision theory, the optimal behavior is the decision that minimizes expected cost, or, equivalently, maximizes expected utility. The expected cost is the cost associated with each possible outcome multiplied with the probability of that outcome. Referring



(A) World State

Action	Rain prob. = 0.3	No rain prob. = 0.7
Do not take umbrella	wet & cold cost = 90	dry & happy cost = 0
Take umbrella	slightly damp cost = 10	dry & burdened cost = 5

(B) World State

Action	Rain prob. = 0.3	No rain prob. = 0.7
Do not take umbrella	wet & cold cost = 40	dry & happy cost = 0
Take umbrella	slightly damp cost = 20	dry & burdened cost = 10

Figure 13.1: On a 100-point scale, two people rank the unpleasantness (cost) of four possible outcomes. **(A)** Walking in the rain without an umbrella results in a wet-and-cold outcome of cost 90; walking in the rain with an umbrella leaves one only slightly damp (cost 10); and so on. The optimal decision for this individual is to carry his umbrella. **(B)** Another individual assigns different costs. The optimal action for this individual is to leave her umbrella at home.

to **Fig. 13.1A**, we see that the expected cost of carrying the umbrella is:

$$\text{EC(umbrella)} = 0.3 \cdot 10 + 0.7 \cdot 5 = 6.5 \tag{13.1}$$

That is, if we choose to carry the umbrella, we have a 30% chance of incurring a cost of 10, and a 70% chance of incurring a cost of 5. The expected cost of 6.5 can be thought of as the average cost that would result, if we chose to take the umbrella each day, over many days with weather identical to the current day. By contrast, the expected cost of not carrying the umbrella is:

$$\text{EC(no umbrella)} = 0.3 \cdot 90 + 0.7 \cdot 0 = 27 \tag{13.2}$$

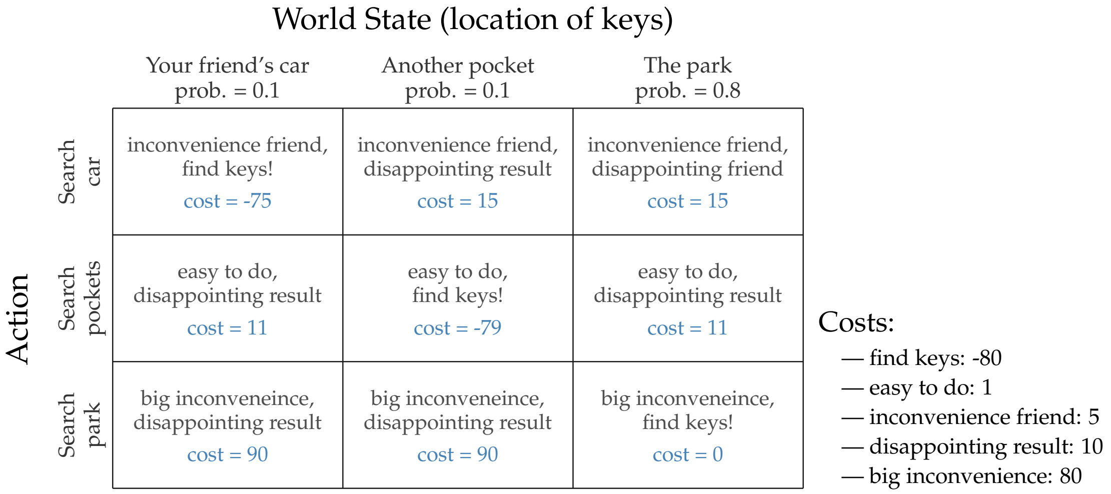
If we choose not to carry the umbrella, we have a 30% chance of incurring a cost of 90, and a 70% chance of incurring a cost of 0. The expected cost of 27 can be thought of as the average cost that would result, if we chose not to take the umbrella each day, over many days with weather identical to the current day. Because $6.5 < 27$, the action that minimizes expected cost is to carry the umbrella; that is the optimal action for this individual.

Importantly, the cost or utility placed on an outcome reflects a personal preference; it is inherently subjective. Indeed, two people with the identical posterior distribution over the world state may choose opposite courses of action, because of the distinct values that they place on particular outcomes. To illustrate this point, consider a second individual who agrees that the chance of rain is 30%, but for whom the costs of the outcomes are different (**Fig. 13.1B**). Note that the two people agree in their *ranking* of the outcomes from highest to lowest in cost, but they assign different numerical costs to the outcomes. For this second person, the optimal action is to leave the umbrella at home.

Exercise 13.1 Why is this the case? ■

13.1.2 Deciding among several actions

The procedure we have illustrated for selecting one of two actions applies equally to situations that require the selection of one of several actions. As an illustration, suppose that, upon getting a ride home from a picnic in the park with a group of friends, you realize that your house keys are not in



World State (location of keys)

Action	Your friend's car prob. = 0.1	Another pocket prob. = 0.1	The park prob. = 0.8
Search car	inconvenience friend, find keys! cost = -75	inconvenience friend, disappointing result cost = 15	inconvenience friend, disappointing friend cost = 15
Search pockets	easy to do, disappointing result cost = 11	easy to do, find keys! cost = -79	easy to do, disappointing result cost = 11
Search park	big inconveneince, disappointing result cost = 90	big inconveneince, disappointing result cost = 90	big inconveneince, find keys! cost = 0

Costs:

— find keys: -80
— easy to do: 1
— inconvenience friend: 5
— disappointing result: 10
— big inconvenience: 80

Figure 13.2: On a -100 to +100 point scale, the costs of different outcomes in the key search problem. Costs for each outcome were calculated by addition of the costs associated with each feature of the outcome (inset). If you decide to the search the car, you will need to call your friend and ask him to do that for you, which imposes an inconvenience on your friend and you do not like that (cost 5); if you decide to search the grass at the park where you picnicked, you will need to take a long trip back to the park, and probably spend lot of time searching there, a big inconvenience (cost 80). Thoroughly searching your other pants' pockets and the pockets in your coat is easy to do (cost 1). Searching and not finding your keys would be disappointing (cost 10). Finally, finding your keys would be very rewarding (cost -80).

your right pants' pocket, where you customarily keep them. Based on knowledge of your recent activities, you quickly generate a probability distribution over the location of your lost keys. They might have fallen out in your friend's car; you might have placed them in a different pocket; or you might have lost them at the park. Depending on where your keys actually are, and where you decide to search for them, nine outcomes are possible, and each of these has a particular cost to you (**Fig. 13.2**). The decision you need to make is: where should you search first?

The optimal decision will be the one associated with minimal expected cost. Computing the expected cost of each action, we have:

$$\text{EC(search in car)} = 0.1 \cdot -75 + 0.1 \cdot 15 + 0.8 \cdot 15 = 6 \tag{13.3}$$

$$\text{EC(search in pockets)} = 0.1 \cdot 11 + 0.1 \cdot -79 + 0.8 \cdot 11 = 2 \tag{13.4}$$

$$\text{EC(search in park)} = 0.1 \cdot 90 + 0.1 \cdot 90 + 0.8 \cdot 0 = 18 \tag{13.5}$$

Thus, despite the fact that you consider it most probable that the keys are in the park, it is optimal to search first in your pockets.

Box 13.1 — The drunkard and the lamppost. The tale of the drunkard who loses his keys walking from the bar to his car, but decides to search for them under a lamppost, presents an amusing case of suboptimal decision making. From a Bayesian decision theoretic perspective, the drunkard has correctly assigned low cost (high reward value) to the outcome (search under lamppost, find keys), because if his keys are there he is likely to encounter them easily and quickly given the light. However, he has failed to take into consideration that the probability is zero that his keys are in that location, since he never was near the lamppost to begin with. Being biased towards searching where it is easiest is also called the *streetlight effect*.

Figure inclusion pending permissions. ■

Exercise 13.2 Suppose that, upon searching your pockets, you fail to find the key. Where would you search next? To find out, use Bayes' rule to update your probability distribution over location, given the new data that your keys are not in your pockets, then once again find the action that minimizes expected cost. ■

Box 13.2 — Bayesian search. The example we have been considering is one of Bayesian search. This is similar conceptually to the visual search examples that we discussed in Chapter 11, but in those examples we did not consider a cost function. Bayesian search has been used successfully to discover valuable lost objects; it is commonly used to search for ships lost at sea. A famous example was the loss of the U.S. Navy's nuclear submarine, USS Scorpion, which disappeared during a voyage in the Atlantic in May, 1968, with 99 crew aboard. Interviews with Navy experts were used to assign probabilities to different scenarios that might have caused the sinking of the sub, and computer simulations were then run to define a prior probability distribution for the sub's location on the ocean floor. A search grid was constructed, with a prior probability assigned to each square on the map. In search operations, each grid square can be associated with a particular cost, based in part on the difficulty of finding the sub if it were at that location; the seafloor differs in depth, and in some areas has narrow canyons that would increase the difficulty of the search. If a grid square is searched and the sub not found, the probability distribution over the map can be updated, and the optimal search location recalculated. The USS Scorpion was found in October, 1968, in about 10,000 feet of water, approximately 400 miles southwest of the Azores islands. Bayesian search was similarly helpful in searching for Air France flight 447, which perished over the Atlantic Ocean in 2009.

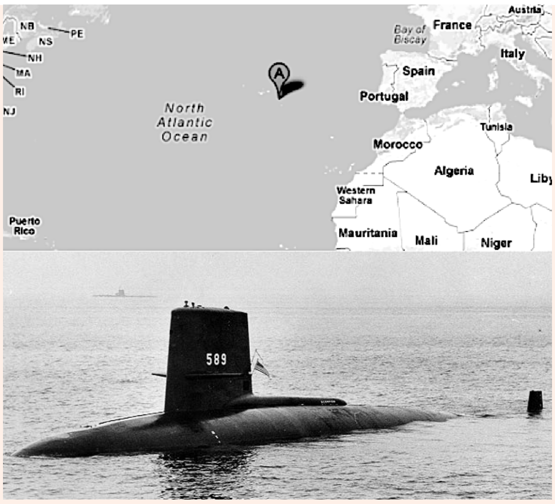


Figure inclusion pending permissions.

13.2 Mathematical formulation: expected utility

Notation for this chapter	
s	World state
a	Action
x	All observations/measurements
$U(s,a)$	Utility of action a when the world state is s

We will consider an *agent* who considers actions a while the world state s is not known. Unlike in earlier chapters, s is not necessarily the stimulus; it could for instance also be a category, which we often denoted by C. In previous chapters, the action is an estimate of the world state s (and we would refer to the agent as the observer), but in this chapter, it does not have to be; for example, if s is whether the milk is still good, than a could be "drink" or "not drink".

We define a *utility function* $U(s,a)$, which is a function of both state s and action a. Reward corresponds to positive utility, cost to negative utility. U is typically real-valued. In previous chapters, when $a = \hat{s}$, then U is typically a function of the difference between s and $\hat{s}$.

Suppose now that the observer has a posterior $p(s|x)$ over s. Here, x stands for the collective set of measurements or observations made by the observer; it does not have to be a single scalar measurement, as in previous chapters. Under this posterior, the expected value of the utility of action a, also more concisely called the *expected utility* of a, is (see Appendix A)

$$\text{If } s \text{ is discrete: } \mathrm{EU}(a) = \sum_s U(s,a)p(s|x) \tag{13.6}$$

$$\text{If } s \text{ is continuous: } \mathrm{EU}(a) = \int U(s,a)p(s|x)ds \tag{13.7}$$

Note that the sum and the integral are over s, so the result does not depend on s. The simplest version of Bayesian decision theory postulates that the agent maximizes expected utility: they choose the action that – averaged over all possible world states weighted by their respective posterior probability – yields the highest possible utility:

$$a_{\text{optimal}} = \underset{a}{\operatorname{argmax}}\, \mathrm{EU}(a). \tag{13.8}$$

Box 13.3 — The multiple faces of "decision-making". In psychology and neuroscience, different subcommunities have different notions of what constitutes the field "decision-making". Here are three:

- Decision-making must involve a **non-trivial cost/reward structure**. People holding this notion typically work in behavioral economics, neuroeconomics, or in the field called "judgment and decision-making", which is concerned with cognitive biases. Decisions rarely involve perceptual uncertainty; for example, subjects choose between two gambles/lotteries presented on the screen. A fundamental question in this field is what, if anything, people optimize. With this notion in mind, the current chapter is the first one that touches upon true decision-making, but Eq. would be considered an oversimplification.
- Decision-making must involve **accumulation of evidence**. People holding this notion typically measure subjects' reaction times in perceptual or cognitive tasks: the subject has freedom to decide when to respond. A prominent model in this field is the *drift-diffusion model*. Notions of optimality are quite rare in this field. A fundamental question is how observers terminate a decision process; this is typically modeled as a decision variable hitting a "bound". The work in this book is only tangentially related to this field, since we do not consider reaction time paradigms. The point of contact is the accumulation of evidence discussed in Chapter 5.
- In this book, we use an inclusive notion: decision-making is any process that maps stimuli to response. This includes purely perceptual decisions, purely utility-based decisions, and anything in between.

■

13.2.1 Binary classification

To start, we will work out this framework in a simple case: Consider the generative model of Chapter 7 or 8, where a class C (0 or 1) has to be inferred from a measurement x. The observer reports class. Thus, $a = \hat{C}$ like in those chapters, and the utility function is a function $U(C,\hat{C})$. But here is the twist: a correct response of $C = 0$ yields utility $U(0,0) = U_0$, while a correct response of $C = 1$ yields utility $U(1,1) = U_1$. An incorrect response yields no utility: $U(0,1) = U(1,0) = 0$. If U_0 were equal to U_1, maximizing expected utility would reduce to maximizing accuracy, for which we already know MAP estimation to be the solution. Here, however, we allow U_0 and U_1 to be different from each other. Intuitively, we would expect that if, for example, $U_1 > U_0$, then we would expect the observer to be more inclined to report class 1. The expected utility of the action "report class 1" is then

$$\begin{aligned} \mathrm{EU}(\hat{C}=0) &= \sum_{C=0}^{1} U(C,\hat{C}=0)p(C|x) && (13.9)\\ &= U(C=0,\hat{C}=0)p(C=0|x) + U(C=1,\hat{C}=0)p(C=1|x) && (13.10)\\ &= U_0 p(C=0|x) + 0 \cdot p(C=1|x) && (13.11)\\ &= U_0 p(C=0|x). && (13.12) \end{aligned}$$

This is simply the probability of being correct multiplied by the reward following a correct response. Similarly, $\mathrm{EU}(\hat{C}=1) = U_1 p(C=1|x)$. If the agent maximizes expected utility, then they will choose $\hat{C} = 1$ if the expected utility of this choice is greater than of the alternative, $\hat{C} = 0$. Substituting the expressions for EU, we find that this is the case when

$$\frac{U_1 p(C=1|x)}{U_0 p(C=0|x)} > 1. \qquad (13.13)$$

We now write out the posterior ratio as the product of a prior ratio and a likelihood ratio:

$$\frac{U_1}{U_0}\frac{p(C=1)}{p(C=0)}\frac{p(x|C=1)}{p(x|C=0)} > 1. \tag{13.14}$$

We learn from this that utilities in a way acts like priors: we can consider the product of the utility ratio and the prior ratio, $\frac{U_1}{U_0}\frac{p(C=1)}{p(C=0)}$ as a "new" prior ratio. In other words, the agent's bias towards class 1 changes in the same way when the prior ratio changes by a factor, as when the utility ratio changes by that factor.

13.2.2 Continuous estimation

In the examples discussed above, both the world state variable and the action choice could take on only a limited number of discrete values. In contrast, many daily examples involve continuous world states and/or choices among a continuum of possible actions. We can no longer apply costs individually to every outcome in such cases; rather, we need to use a utility function over the continuum of possible world states or actions.

For instance, returning to our umbrella example, we could consider a more nuanced scenario in which the world state can take a continuum of values reflecting the rate of the rainfall, from 0 (no rain) to 10 (extremely heavy rainfall). This results in a continuum of possible outcomes, each with a specific cost, even when we consider only two actions (to take or not to take the umbrella). If we tried to create an outcome table to represent this situation, we would have two rows (as in **Fig. 13.1**) but an infinite number of columns. Clearly, another approach is needed here.

In such cases, we need to construct a function that reflects cost over the continuous space of outcomes for each action. Perhaps the cost to us of being in the rain grows linearly with the amount of water that lands on us. The utility function could then be expressed as $U(sa,a) = -(A+Ba)s$, where s represents the rainfall rate, a represents our action (0 for carrying, and 1 for not carrying the umbrella), and A and B are constants. Thus, whether we have our umbrella or not, we would be increasingly displeased by greater rainfall, but the rise in our displeasure, as a function of rainfall rate, is greater when we lack the umbrella. We would then replace the sum by an integral:

$$a_{\text{optimal}} = \underset{a}{\text{argmax}}\ \text{EU}(a) \tag{13.15}$$

$$= \underset{a}{\text{argmax}} \int U(s,a)p(s|x)ds. \tag{13.16}$$

To make the problem even more realistic, we could consider not just a continuum of rainfall rates, but also a continuum of possible actions reflecting not just our decision to take or leave the umbrella, but also our walking speed. We might walk slowly or attempt to minimize our exposure time to a possible rainfall by sprinting to our location (or go at any speed in-between). If we tried to create an outcome table to represent this situation (e.g., **Fig. 13.1**), we would have an infinite number of rows and columns. Again, to deal with this situation we would need to specify a cost function over the space of outcomes.

The specification of a cost function for real-life decision problems is a difficult task, although there are multiple obvious ingredients of the cost function. We want to satisfy our immediate needs and desires as well as progress towards more distal goals. This implies that obtaining food, drink, shelter, mating, and other factors will have utility. At the same time, we do not want to overly exert ourselves, we want to minimize our energy consumption, the risk of damage to our body and other factors that are negative to our well-being.

Exercise 13.3 Sketch a cost function that could describe much of your behavior today. ■

13.3 Relation to previous chapters

13.3.1 Discrete tasks

In previous chapters, in discrete tasks, the observer cared about maximizing accuracy. If the discrete world state is s, then this corresponds to a utility function that is 1 (or any other positive number) when $\hat{s} = s$ and 0 otherwise. We can write such a “correctness” utility function as

$$U(s, \hat{s}) = \begin{cases} 1 & \text{if } \hat{s} = s \\ 0 & \text{otherwise} \end{cases} \tag{13.17}$$

13.3.2 Continuous tasks

In continuous tasks, this utility function is not very sensible. For example, in a location estimation task, it is impossible to get the position *exactly* right, $\hat{s} = s$. Moreover, you will be more pleased when you make a small error than when you make a big error. We need to capture this in the utility function.

One reasonable try is to postulate that the cost of reporting the estimate $\hat{s}$ when the true stimulus is s is the squared difference between the two: a quadratic cost function. For the utility function, that means.

$$U(s, \hat{s}) = -(s - \hat{s})^2. \tag{13.18}$$

When the posterior is $p(s|x)$, expected utility is then

$$\begin{aligned} \mathrm{EU}(\hat{s}) &= \int U(s, \hat{s}) p(s|x) ds & (13.19) \\ &= -\int (s - \hat{s})^2 p(s|x) ds. & (13.20) \end{aligned}$$

To find the value of the estimate that maximizes EU, we compute the derivative of EU with respect to $\hat{s}$:

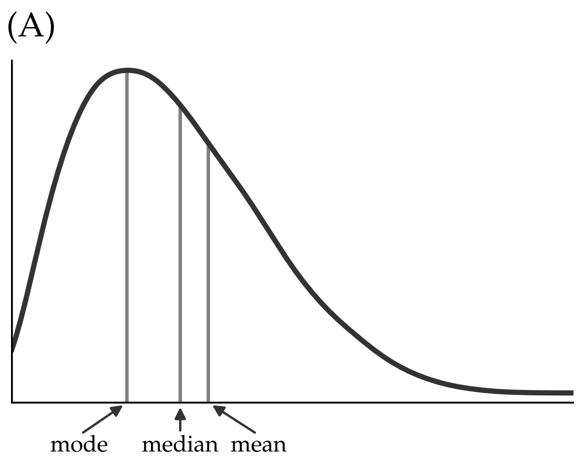
$$\begin{aligned} \frac{\partial \mathscr{L}(\mathrm{EU})}{\partial \hat{s}} &= \int 2(s - \hat{s}) p(s|x) ds & (13.21) \\ &= \int 2sp(s|x) ds - \int 2\hat{s} p(s|x) ds & (13.22) \\ &= 2 \int sp(s|x) ds - 2\hat{s} \int p(s|x) ds & (13.23) \\ &= 2\mathbb{E}[s|x] - 2\hat{s} \cdot 1, & (13.24) \end{aligned}$$

where $\mathbb{E}[s|x]$ is the posterior mean. The derivative equals 0 when $\hat{s} = \mathbb{E}[s|x]$, i.e. when the estimate is equal to the posterior mean. Thus, the posterior mean is the optimal readout given the quadratic cost function. (We mentioned this in Chapter 4 but without proof.)

For every utility functions, the procedure of maximizing expected utility will lead to a different decision rule. The quadratic cost function is often used in practice as the mean is easy to calculate, and the utility function often approximates what subjects care about (be close to the target). However, other cost functions are also plausible. Perhaps most notable is the *absolute error* cost function,

$$U(s, \hat{s}) = -|s - \hat{s}|. \tag{13.25}$$

Maximizing expected utility now leads the decision-maker to report the *median* rather than the mean of the posterior (see Problem 13.2).



(B)

$$\hat{s}_{\text{mean}} = \langle s \rangle = \int_{-\infty}^{\infty} s p(s|x) ds$$

$$\int_{-\infty}^{\hat{s}_{\text{median}}} p(s|x) ds = \int_{\hat{s}_{\text{median}}}^{\infty} p(s|x) ds$$

$$\hat{s}_{\text{mode}} = \text{argmax } p(s|x)$$

Figure 13.3: **(A)** An asymmetrical probability density; this might represent an observer's posterior probability distribution over sound source location on a particular trial. The three lines, from left to right, represent the mean, median, and mode of the distribution. **(B)** Definitions of mean, median, and mode. The mean is the center-of-mass of the distribution; the median divides the distribution into two halves of equal area; the mode is the point at which the distribution is highest.

In the context of perceptual estimation problems, the choice of utility functions is particularly important when the posterior distribution is asymmetric. For symmetric unimodal distributions such as the Gaussian distribution, the mean, median and mode are identical. For asymmetric distributions, however, they are distinct (**Fig. 13.3**). Furthermore, some distributions, while symmetric, are bimodal. In Chapter 10, when discussing causal inference, we encountered bimodal (two-peaked) posterior distributions (see **Fig. 10.4**. For such distributions, the mean and mode are generally distinct and research has suggested that human subjects report closer to the mean of such a bimodal posterior (Kording et al., 2007).

13.3.3 What it means to decide optimally

Optimal decision-making means minimizing expected utility. In the previous section, we argued that given the posterior, the utility function dictates which readout of the posterior is optimal. However, why would the posterior itself be needed for optimality? Why could the observer not use some other probability distribution over the stimulus computed from the observations, say $q(s|x)$? A situation where this might happen is that the observer does not know the correct structure of the generative model. For example, in the sound localization task, the observer might use an incorrect prior over the position of the sound source. This would give rise to a different posterior, $q(s|x)$. Such a wrongly estimated posterior distribution could be used for decision making.

Using a q that differs from p will in general not maximize expected utility. To understand this, suppose the observer is free to follow ANY estimation strategy to go from x to $\hat{s}$, let us say some function F. The utility obtained on a single trial is then $U(s,F(x))$, and the overall expected utility incurred across all trials in an experiment is

$$\text{OEU} = \sum_{s,x} U(s,F(x)) p(s,x). \tag{13.26}$$

(We use OEU to distinguish from EU, which is computed on a single trial.) Now consider the specific F that we have considered throughout this chapter, namely the one that maximizes expected utility under the true posterior, $p(s|x)$. We denote this mapping by $F*$. For $F*$, we thus know by definition that for any x, the quantity $\sum_s U(s,F^*(x))p(s|x)$ is as high as it can possibly be: there is no F for which $\sum_s U(s,F(x))p(s|x)$ is higher. Since this statement is true for any x, it must also be true when we average over x. Thus, $F*$ is such that $\sum_x p(x)\sum_s U(s,F^*(x))p(s|x)$ is as high as it can possibly be. But this quantity is exactly equal to the "overall expected utility" OEU above. In particular, that means that constructing F based on some substitute posterior distribution $q(s|x)$ cannot yield a higher overall expected utility than using the true posterior distribution. In this sense, the computation of the Bayesian posterior is the essence of optimal decision-making.

13.3.4 Percept and report

In previous chapters, we postulated that the outcome of Step 2, the estimate of the world state, would automatically be what the observer *perceived*. This relied on the observer's "natural" utility functions used in perceptual tasks, which might be correctness in a discrete task and negative squared error in a continuous task. If the experimenter (or the environment) introduces a different utility function, as in Section 13.2.1, then there will be a disconnect between the *percept* and the *report* of the world state.

13.4 Complications

The framework laid out in previous sections is a simplification. Here we discuss a few common complications.

13.4.1 Cost functions for uncertain outcomes

In the examples we have considered so far, each combination of world state and action has mapped deterministically onto a single outcome, to which we assigned a cost. In reality, however, it is often the case that many outcomes could result – with different probabilities – from a particular combination of action and world state. For example, in **Fig. 13.1**, we assumed that, if we were to leave the house without our umbrella (action) and it were to rain (world state), then we would get wet (outcome). In reality, however, a range of possible outcomes might result from this combination of action and world state, depending probabilistically on the availability of places along our path under which we could take cover from the rain. In such situations, given a particular combination of action and world state, one can define a probability distribution across outcomes: $p(o|a,s)$ (**Fig. 13.4**). Now, each *outcome* (o) is associated with a utility, and the optimal decision-maker would choose the action a that maximizes expected utility. Thus, the formula for the optimal action (Eq. (13.4)) generalizes to:

$$a_{\text{optimal}} = \underset{a}{\text{argmax}}\ \text{EU}(a) \tag{13.27}$$

$$= \underset{a}{\text{argmax}} \int \left(\int U(o)p(o|s,a)do \right) p(s|x)ds. \tag{13.28}$$

In other words, you average the utility function both over possible outcomes given your hypothesized action and over possible states of the world.

13.4.2 Nonlinearity between reward and utility

If an experimenter gives double the money as a reward to a human subject, or double the juice to a monkey subject, then it does not mean that their utility is doubled. In particular, utility might be subject to the *law of diminishing returns*: the same increment in objective reward might yield a smaller increase in subjective utility when added to a larger base amount. This can be modeled by relating utility U to reward R through a power law:

$$U = R^{\alpha}. \tag{13.29}$$

This provides a resolution to the famous St. Petersburg Paradox (see Box).

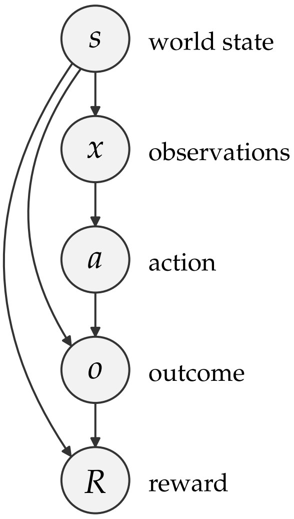


Figure 13.4: General framework for Bayesian decision-making.

Box 13.4 — History of utility theory. Daniel Bernoulli and before him Gabriel Cramer worked on the St. Petersburg paradox.

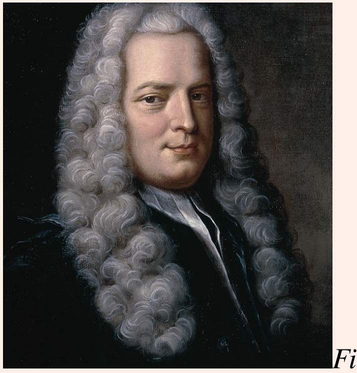


Figure inclusion pending permissions.

This paradox relates to a simple game of chance: on every round, a fair coin is thrown, and the game ends as soon as the coin lands tails (T). The subject then gets 2^h dollars, where h is the number of heads (H) observed. For example, if the sequence of landings observed was T, the player would win \$1; HT, \$2; HHT, \$4; HHHT, \$8; and so on. It turns out that the expected payoff of this game is infinite. The expected utility is the sum over all possible outcomes of the payoff times the probability of the outcome. Representing each possible outcome by its number of heads, we have:

$$\mathrm{EU} = \sum_{h=0}^{\infty} p(h) 2^h = \sum_{h=0}^{\infty} \left(\frac{1}{2}\right)^{h+1} 2^h = \sum_{h=0}^{\infty} \frac{1}{2} = \infty. \tag{13.30}$$

Figure inclusion pending permissions.

The paradox is that, although the expected utility is infinite, people would not pay more than a few dollars to enter the game. A way to resolve this paradox is through the introduction of a utility function that reflects the decreasing marginal utility of money: 1,001,000 dollars is only slightly more valuable than 1,000,000 dollars, whereas 1,000 dollars is much more valuable than nothing.

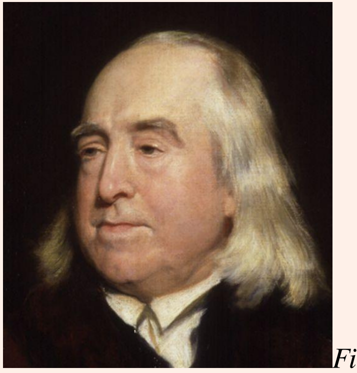


Figure inclusion pending permissions.

Jeremy Bentham, another inventor of the idea of utility, applied it more directly to the pleasures and pains of humans. He used these ideas to derive how society should be organized – namely by maximizing the utilities of all the citizens, the philosophy of utilitarianism. Bentham saw all moral and legal norms as derivable from this simple principle using methods from logic and experimentation. ■

13.4.3 Distortions of probability

In *prospect theory*, which mostly deals with economic decisions such as whether to take \$10 or a 50% chance of winning \$25, it is sometimes alleged that in the computation of expected utility, the probabilities of the outcomes are not linearly taken into account. However, the probabilities under consideration are then usually explicitly presented one like the 50% in the example. Whether posterior probabilities are nonlinearly weighted is unknown.

13.4.4 Decision noise

In economic decisions, subjects do not always make the same decision in the same situation. In perception, such behavioral variability can often easily be explained as a consequence of measurement noise, like we have done throughout the book. In economic decisions without a perceptual component, that is not an option. Therefore, often a form of *decision noise* is introduced to account for human behavior: given a set of actions a with corresponding expected utilities $\mathrm{EU}(a)$, the observer does not always chooses the same a. The most common way to implement decision noise is by postulating that the *probability* of choosing a is proportional to an exponential function of EU(a):

$$p(a) = \frac{e^{\beta \, \mathrm{EU}(a)}}{\sum_a e^{\beta \, \mathrm{EU}(a)}}. \tag{13.31}$$

When β grows very large, the action with the highest expected utility gets a huge boost, so that its probability approaches 1: this is the original case of the EU-maximizing agent. When $\beta = 0$, the observer randomly chooses an action with equal probability, regardless of EU. For any other β, the agent does something in between randomly choosing and maximizing. For this reason, the decision rule is also called a *softmax* rule and the parameter β is also called the *inverse temperature* (by analogy with thermodynamics): the lower β, the higher the "temperature", and the noisier the system.

13.5 Experimental approaches

13.5.1 Imposing utility functions

Often, when we are examining the behaviors of a human subject or animal, we want to compare them against what is optimal. We can compare predictions from Bayesian Decision theory with actual behavior. Yet, there is a fundamental hurdle in this comparison. For this type of analysis to be constructive, we must know beforehand what it is the experimental subject is trying to achieve, i.e., what their cost function is. With this in mind, researchers have designed experiments where the cost of the task is relatively explicit. In a set of reaching studies, it was observed that people are remarkably close to the optimal choices prescribed by decision theory when final positions translated into known monetary gains and losses (Trommershäuser et al., 2003, Trommershauser et al., 2005, Maloney et al., 2006). Similar experiments have studied visual tasks (see Whiteley and Sahani, 2008) and force-producing tasks (Kording et al., 2004 Kording and Wolpert, 2004). This further demonstrates people's abilities to integrate statistical information in a Bayes-optimal manner, not just for estimation, but also for action selection.

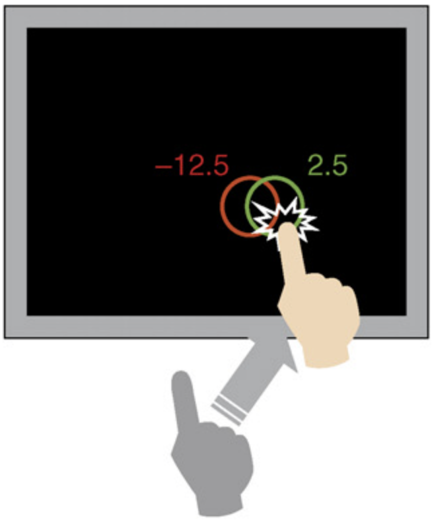


Figure inclusion pending permissions.

Julia Trommershauser and colleagues used the following experiment to study movement under uncertainty. Imagine you have to make a pointing movement to inside a small green circle on a screen, but avoid the inside of a small red circle. If you hit inside the green circle but not inside the red, you earn 2.5 points. If you hit inside the red circle but not inside the green, you lose 13.5 points. If you hit in the intersection of both circles, you lose 10 points. If you hit outside both circles, you earn 0 points. If you take too long, you earn 0 points as well, so you are forced to move fast. Assume the radius of the circles is 1 and the distance between the centers is D. Also assume that the place where you hit the screen is not exactly where you aim, because it is corrupted by movement noise. This noise has a two-dimensional Gaussian distribution with standard deviation σ. There is a point where you should aim to maximize the expected number of points you earn. This point will depend on the costs and rewards, as well as on the separation of the circles and your own motor noise (see Problems). Humans learn to optimally take into account these quantities to decide on their optimal reaching point.

13.5.2 Inferring utility functions

Utility functions are hypothesized functions that human subjects are supposed to optimize. There are some known properties of utility functions that are important. Of primary importance is the fact that they cannot be uniquely measured. Suppose we have a cost function $C(o)$ that a given subject is trying to optimize through a choice of action. Then $F(C(o))$, where F is any monotonic function, leaves the preference of one action over another action invariant. Hence, a person using $C(o)$ and another using $F(C(o))$ will have the same set of preferences over actions. As the cost function is only observable indirectly through preferences over actions, it is impossible to distinguish cost functions from their monotonic transformations. Nevertheless, while it is not possible to measure cost functions directly, it is possible to measure indifference curves. If we have a multidimensional cost function we can produce games where we ask about which pairs of outcomes subjects are indifferent to. Through many such measurements, cost functions can be reasonably well characterized. Using inverse decision theory, studies have examined implicit cost functions subjects use when performing motor tasks and when penalizing target errors. These inferred cost functions have highly nonlinear and nontrivial forms (Todorov, 2004). These findings highlight a crucial problem in decision theory: good fits to behavior may be obtained with incorrect cost functions. Inverse decision theory can thus be used as a means of searching for violations in the assumptions we make using the Bayesian approach to decision making.

13.6 Summary and remarks

In this chapter, we have formulated the problem of optimal decision making in the context of utility maximization. We have learned:

- Utility and cost are both universal ways of phrasing optimal decision making. They are thus near universally used in philosophy and optimization.
- Without uncertainty optimal decision making boils down to choosing the action with highest utility.
- In the context of outcome uncertainty we instead need to optimize expected utility.
- Both binary and continuous decision making can be phrased in this context.
- Inverse reinforcement learning can be used to estimate the utility function used by an agent that optimizes said utility function.

13.7 Suggested readings

- Konrad P Körding et al. “A neuroeconomics approach to inferring utility functions in sensorimotor control”. In: *PLoS biology* 2.10 (2004), e330
- Konrad Paul Körding and Daniel M Wolpert. “The loss function of sensorimotor learning”. In: *Proceedings of the National Academy of Sciences* 101.26 (2004), pages 9839–9842
- Laurence T Maloney, Julia Trommershäuser, and Michael S Landy. “Questions without words: A comparison between decision making under risk and movement planning under risk.” In: *Integrated models of cognitive systems* 29 (2007), pages 297–313

- Chris R Sims. “The cost of misremembering: Inferring the loss function in visual working memory”. In: *Journal of vision* 15.3 (2015), pages 2–2
- Lawrence D Stone et al. “Search analysis for the underwater wreckage of Air France Flight 447”. In: *14th International Conference on Information Fusion*. IEEE. 2011, pages 1–8
- Emanuel Todorov. “Optimality principles in sensorimotor control”. In: *Nature neuroscience* 7.9 (2004), pages 907–915
- Julia Trommershäuser et al. “Optimal compensation for changes in task-relevant movement variability”. In: *Journal of Neuroscience* 25.31 (2005), pages 7169–7178
- Julia Trommershäuser, Laurence T Maloney, and Michael S Landy. “Statistical decision theory and the selection of rapid, goal-directed movements”. In: *JOSA A* 20.7 (2003), pages 1419–1433
- Louise Whiteley and Maneesh Sahani. “Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes”. In: *Journal of vision* 8.3 (2008), pages 2–2

13.8 Problems

Problem 13.1 Suppose that you are walking in the dark in an area that is occasionally inhabited by lions. You hear a suspicious noise that may indicate the presence of a lion. In deciding whether to run away or not, you apply the following cost structure to each of four possible outcomes:

		World State	
		Lion	No lion
Action	Run	physical effort no injury cost = 2	physical effort no injury cost = 2
	Stay	no effort severe injury cost = 100	no effort no injury cost = 0

a) Given the cost structure shown, if you believe that there is only a 30% chance that the lion is present, should you run or stay? Give a complete calculation.

b) How low would your belief in the lion’s presence have to be before you decided to stay rather than to run? Give a complete calculation.

c) The outcomes listed in the table above represent a simplistic view of the problem. To make the problem more realistic, consider the following: In reality, if the lion is present, it may catch you even if you run; if you stay, the lion might (with some probability) decide not to attack you; if you run, you have some probability of injuring yourself by falling or colliding with objects (trees, boulders). Try to modify your solution to the problem, taking into account these realistic considerations. In your calculations, you may use what you believe to be realistic probabilities for each of these contingencies.

Problem 13.2 Suppose we use absolute error instead of squared error as the utility function: $U(s,\hat{s}) = -|s-\hat{s}|$. Prove that then the posterior *median* rather than the posterior mean maximizes expected utility.

Problem 13.3 This problem refers to the Trommershauser outcome uncertainty task in Section 13.5.1. For simplicity, we regard this as a one-dimensional problem: we only consider aim points that lie on the (infinite) line through the centers of the two circles. Define the point in the middle of the two centers as the origin. Thus, the centers of the two circles are at $-D/2$ and $D/2$.

a. Derive an expression for the expected utility of aiming at point a on this line.

b. Choose a reasonable range of values for D and a reasonable range of values for σ. Numerically compute the optimal aim point as a function of both D and σ. Plot this optimum as a function of D and σ using a color plot.

c. Interpret your plot.

Problem 13.4 Consider a one-dimensional stimulus s. The subject makes a measurement x of s. The subject then has to set an interval of size I in the space of s and gets rewarded when the true stimulus s falls inside the interval. The catch is that the amount of reward declines exponentially with the size of the interval: reward $= e^{-I}$. Thus, the subject could set a larger interval to be more sure to include s, but would earn a lower reward when successful.

a) Suppose the posterior $p(s|x)$ is Gaussian with mean $\mu_{\text{posterior}}$ and standard deviation $\sigma_{\text{posterior}}$, and assume that the subject sets the interval around the posterior mean, i.e. $[\mu_{\text{posterior}} - I/2$, $\mu_{\text{posterior}} + I/2]$. Derive an expression for the expected utility of setting the interval to have size I.

b) Plot expected utility EU as a function of I for three values of $\sigma_{\text{posterior}}$: 0.5, 1, and 2. Plot all three curves in the same plot, and color code them.

c) Numerically compute for each of these values of $\sigma_{\text{posterior}}$ which interval size would maximize EU.

d) What is the value of this highest possible EU in each of the three cases?

e) Consider a wider and finer range of $\sigma_{\text{posterior}}$: 0.1 to 3 in steps of 0.1. Plot the EU-maximizing interval size as a function of $\sigma_{\text{posterior}}$. This should be a monotonically increasing function.

f) Using the same range of $\sigma_{\text{posterior}}$, plot the maximum value of EU as a function of $\sigma_{\text{posterior}}$.

Problem 13.5 Consider the task of Chapter 7 – discrimination between s_+ and s_- based on a measurement x following a Gaussian distribution with mean equal to the true stimulus and standard deviation σ – but with potentially asymmetric utilities:

		Subject report	
		s_+	s_-
True stimulus	s_+	U_{++}	U_{-+}
	s_-	U_{+-}	U_{--}

Furthermore, assume a flat prior: $p(s_+) = p(s_-) = 0.5$.

a) Derive the optimal decision rule. Use the Φ notation (cumulative standard normal distribution).

b) Derive expressions for the probability correct when the true stimulus is s_+, and when it is s_-.

Problem 13.6 Repeat Problem 13.5 for the task of Section 8.1: binary classification of a continuous stimulus. We denote the classes by $C = -1$ and $C = 1$. Assume the improper CCSDs

$$p(s|C=-1) = \begin{cases} k & \text{if } s < 0 \\ 0 & \text{if } s > 0 \end{cases} \tag{13.32}$$

$$p(s|C=1) = \begin{cases} 0 & \text{if } s < 0 \\ k & \text{if } s > 0 \end{cases} \tag{13.33}$$

where k is a constant.

Problem 13.7 Consider a stimulus variable s that take values on the circle, such as motion direction. The posterior is $p(s|x)$.

a) For estimation of a circular variable, using squared error as a cost function does not make sense. Why not? Explain using a concrete example.

b) A sensible utility function is the cosine of the estimation error, $U(s, \hat{s} = \cos(\hat{s} - s)$. Show that the estimate that maximizes expected utility on a given trial is the circular mean of the posterior, denoted μ_{post}, which is defined by the equations

$$\cos\mu_{\text{post}} = \mathbb{E}[coss|x] \tag{13.34}$$

$$\sin\mu_{\text{post}} = \mathbb{E}[sins|x]. \tag{13.35}$$

Problem 13.8 The woodchuck does not actually chuck wood, it rather lives off grass, herbs and insects. Every day, again and again, it has to allocate its resources between various possible activities. Lets say you have records what a woodchuck does, e.g. where it walks, how fast, and what it eats and when. How would you build a normative model of why the woodchuck behaves the way it does and how would you test that with data?

Problem 13.9 In the realm of cognitive judgments, posterior probability distributions can be interpreted in the context of "knowing what you know", also called metacognition. Research in education has shown that to become an effective learner, students have to have strong metacognitive skills. Students can acquire such skills by reflecting on what they know and how confident they are about their knowledge. In some places, new forms of testing are designed to incorporate such confidence judgments. In a "Variable Weight" testing format, a test could consist of fifteen questions. Students choose ten about which they are confident (indicated by placing their response on the left side of the answer sheet). The questions for which the students indicate they are confident of their answer and in fact answer correctly are weighted more heavily than questions they answer correctly but for which they indicate less confidence. In another innovation,

> "Students are asked a question that requires them to indicate whether they are absolutely sure, fairly sure, or just guessing at their answer. The points for their response are dependent on the correctness of their answer and the confidence they state. For example, if students indicate they are absolutely sure, they earn nine points if they are correct, but no points if they are wrong. However, if they indicate they are unsure or just guessing, they earn three points if they are correct and two points if they are wrong. Students quickly learn that carefully reflecting on their confidence improves their grade." (Randy M. Isaacson, "Building a Metacognitive Curriculum: An Educational Psychology to Teach Metacognition", Tomorrow's Professor)

Explain using the equations in this chapter why this works.

Problem 13.10 MAP estimation (using the correct generative model) is the read-out method that maximizes expected reward if reward is obtained only when the estimate is exactly correct (1, 0 otherwise), that is, when the estimate $\hat{s}$ is equal to the true stimulus s on a given trial. Over many trials, the expected reward collected by an estimator with a MAP payoff is equal to:

$$R = \int_s p(\hat{s} = s|s)p(s)ds \tag{13.36}$$

. Thus, this expected reward (13.36) must have a higher value when the observer performs MAP estimation than when they use any other method of estimation. To verify this, we evaluate Eq. 13.36 for the MAP observer (using the correct generative model). We will assume a prior and stimulus distribution $p(s) = \mathcal{N}(\mu, \sigma_s^2)$ and a measurement distribution $p(x|s) = \mathcal{N}(s, \sigma^2)$.

- Calculate the expected reward for the MAP estimator
- Consider estimators that are weighted averages of the measurement and the mean of the stimulus distribution. Show that among these estimators, the MAP estimate is the one that maximizes expected reward.
- Repeat for linear combinations of the measurement and the mean of the stimulus distribution: $\hat{s} = ax + b\mu$.

14. The neural likelihood function

How can we incorporate neural variability into a generative model?

The brain performs inference based on neural activity. In the earlier chapters, we abstracted this process by introducing the measurement, denoted x, defining the likelihood $\mathscr{L}(s;x)$, and computing the posteriors in a variety of tasks based on such likelihood functions. To understand the neural basis of inference, in this chapter we introduce the formalism of populations of neurons responding to sensory stimuli. We will use this formalism to define the likelihood function over a stimulus represented by the spike counts in a neural population observed on a particular trial.

Any neural representation of variables in the world implies a likelihood function about variables in the world. A conditional probability distribution $p(r|s)$, expresses how stimuli, s, in the world give rise to neural responses, $\mathbf{r}$. The corresponding likelihood function is the same expression but viewed from the perspective of the observer, who has access to $\mathbf{r}$ but not to s:

$$\mathscr{L}(s;\mathbf{r}) = p(\mathbf{r}|s). \tag{14.1}$$

We will call this likelihood function the *neural likelihood function* to distinguish it from the likelihood function $\mathscr{L}(s;x)$ that we encountered in earlier chapters. All information that the neural activity $\mathbf{r}$ provides the observer about the stimulus s is contained in the neural likelihood function over the stimulus. $\mathscr{L}(s;\mathbf{r})$ is the answer to the question "If the stimulus had value s, what would be the probability that it would have produced the observed activity $\mathbf{r}$?"

As with any likelihood function, each likelihood function we will encounter here allows us to define an associated uncertainty. Following Section 3.4.1, we will define uncertainty as the width (standard deviation) of the likelihood function. Here, this uncertainty will reflect the uncertainty about the stimulus that remains after having observed $\mathbf{r}$.

Plan of the chapter

To understand the neural likelihood function, we will assume that neurons are independent and that the spike count of each neuron follows a Poisson distribution. We will consider two types of tuning curves: bell-shaped and monotonic. We will begin by discussing the case of a single neuron and then progress to a population of neurons. We will compute population likelihood functions in cases where they have a compact functional form. We will link the neural likelihood function to the

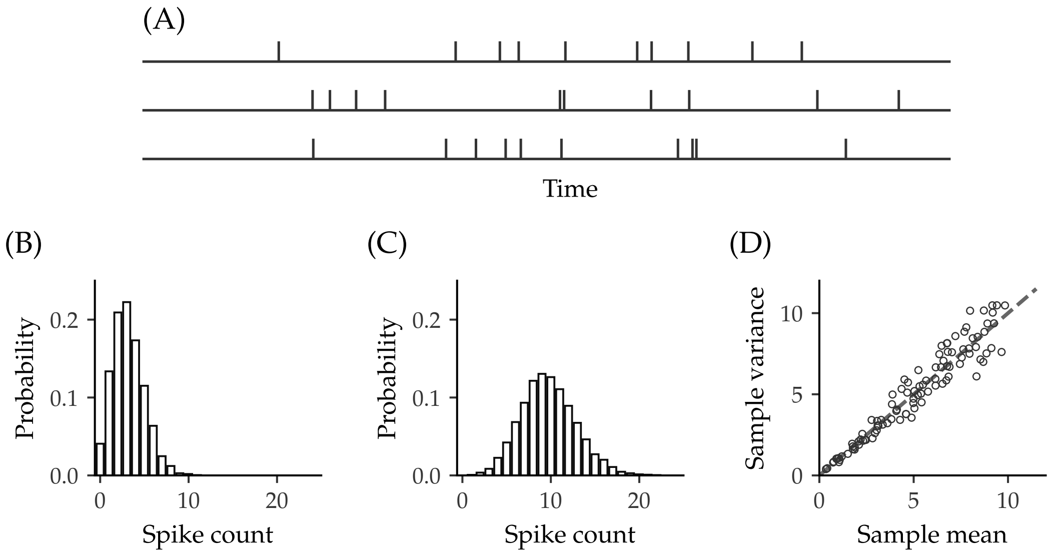


Figure 14.1: Poisson variability in a single neuron. **(A)** Hypothetical spike trains evoked in the same neuron by the same stimulus, repeated four times (trials). Not only do the spike times differ between trials; the spike counts also differ. **(B)** Probability distribution of the spike count of a single Poisson neuron, with rate parameter λ equal to 3.2. Note that the x-axis starts at 0. **(C)** Same but with $\lambda = 9.7$. **(D)** For a Poisson random variable, the variance is equal to the mean. To illustrate this, we drew 100 values of the rate parameter uniformly from $[0, 10]$. For each value, we simulated a sample of 100 spike counts and calculated its sample mean and sample variance.

behavioral likelihood function from earlier chapters. Finally, we will describe an application to brain-machine interfaces.

Throughout this chapter, keep in mind that "the stimulus" is a simple stimulus feature that is encoded at an early stage of sensory processing, such as the contrast of a blob of light, the orientation of a line segment, the location of a sound, or the width of a groove. It does *not* include states of the world that are more complicated or task-dependent, such as "the curvature of a contour", "the presence of a search target", "the category to which a tone belongs", or "the amount of scatter in a cloud of dots". Correspondingly, **r** is the neural activity in that early stage of sensory processing – e.g. retina, LGN, V1, A1, or S1.

14.1 Generative model of the activity of a single neuron

As Step 1, we start with the generative model of the spike count of a single neuron, e.g. the number of spikes elicited by a flash of light that is presented for 10 ms. We describe the spike count of a neuron in a given time interval using a Poisson distribution. Poisson variability (**Fig. 14.1A**) is defined for a non-negative integer; it can be 0. Suppose a stimulus s is presented and the mean spike count of a neuron in response to this stimulus is λ, which does not need to be an integer. λ is also called the rate of the Poisson process. Then the actual spike count will vary from trial to trial, around λ. For every possible count r, we seek its probability. A Poisson process (or in our context, a Poisson spike train) is defined as follows. Imagine a fixed time interval, and divide it into small bins (e.g. 1 millisecond each). We assume that each bin can contain 0 spikes or 1 spike, and that the occurrence of a spike is independent of whether and when spikes occurred earlier (it is sometimes said that a Poisson process "has no memory"). It can be proved in such as case (see Problems) that for a Poisson process with mean λ, the probability of observing a total of r spikes on a single trial is given by the Poisson distribution,

$$p(r|s) = \frac{1}{r!}e^{-\lambda}\lambda^{r}, \tag{14.2}$$

where $r!$ is pronounced “r factorial" and stands for the product $1 \cdot 2 \cdot 3 \cdots r$. The Poisson distribution is shown for $\lambda = 3.2$ and $\lambda = 9.7$ in **Fig. 14.1B-C**. Since the Poisson distribution is discrete, drawing it as a continuous curve would be a mistake. Keep in mind that, while r is an integer, λ can be any positive number. For large enough λ, the distribution is close to symmetrical and looks roughly Gaussian; this is not true for small λ.

An important property of the Poisson distribution is that the variance of a Poisson-distributed variable is equal to its mean: if the mean spike count of a Poisson neuron is λ, then the variance of this neuron’s spike count is also λ. (**Fig. 14.1D**). The ratio of the variance to mean of a neuron’s spike count is called the *Fano factor*; for a Poisson process, the Fano factor is 1.

For our generative model of neural firing, we need to specify the probability of a firing rate, r, as a function of the stimulus, s. To do this, we note that λ is a function of the stimulus: it is the height of the tuning curve (the neuron’s average firing rate) at stimulus level s. Therefore, in terms of the stimulus, Eq. (14.2) can be written as

$$p(r|s) = \frac{1}{r!}e^{-f(s)}f(s)^{r}. \tag{14.3}$$

Box 14.1 — Neural tuning curves. The concept of tuning curves became popular with the pioneering experiments of Hubel and Wiesel in the mid-1960s. They recorded from primary visual cortex (V1) in cat while stimulating with illuminated oriented bars (see **Fig. 14.2A**). They found that the response of a cortical neuron was systematically related to the orientation of the stimulus. There exists one orientation of the stimulus where the neuron fires most rapidly: the neuron’s preferred orientation. For other orientations, the activity decreases with increasing angle relative to the preferred orientation. A plot of the mean firing rate (e.g., spikes per second) as a function of angle describes the neuron’s tuning curve. In the case of many visual neurons, this is a unimodal function (See **Fig. 14.2B**).

Tuning curves can have a wide variety of shapes, depending on the species, the brain area, and stimulus feature. For example, in motor cortex, we find that neural responses influence the direction of movement of the hand of a monkey. Instead of narrow unimodal functions we usually find very broad tuning curves. In auditory cortex, the frequency of the sounds stimulus affects the firing rate of the neuron in a complex tuning curve. And in the hippocampus, a region of the mammalian brain involved in memory acquisition and navigation, there is a two dimensional representation of positions. In experiments with rats, firing rates of hippocampal neurons depend on self-position in two dimensions. Some tuning curves are not bell-shaped, but monotonic (**Fig. 14.2C**). The important thing in all these cases is that reasonably simple tuning curves characterize the mapping from sensory stimuli to the activity of neurons.

This probability distribution, sometimes called the *distribution of neural variability* or *neural noise distribution*, serves as a generative model: it tells us the statistics of the observation r given a world state s.

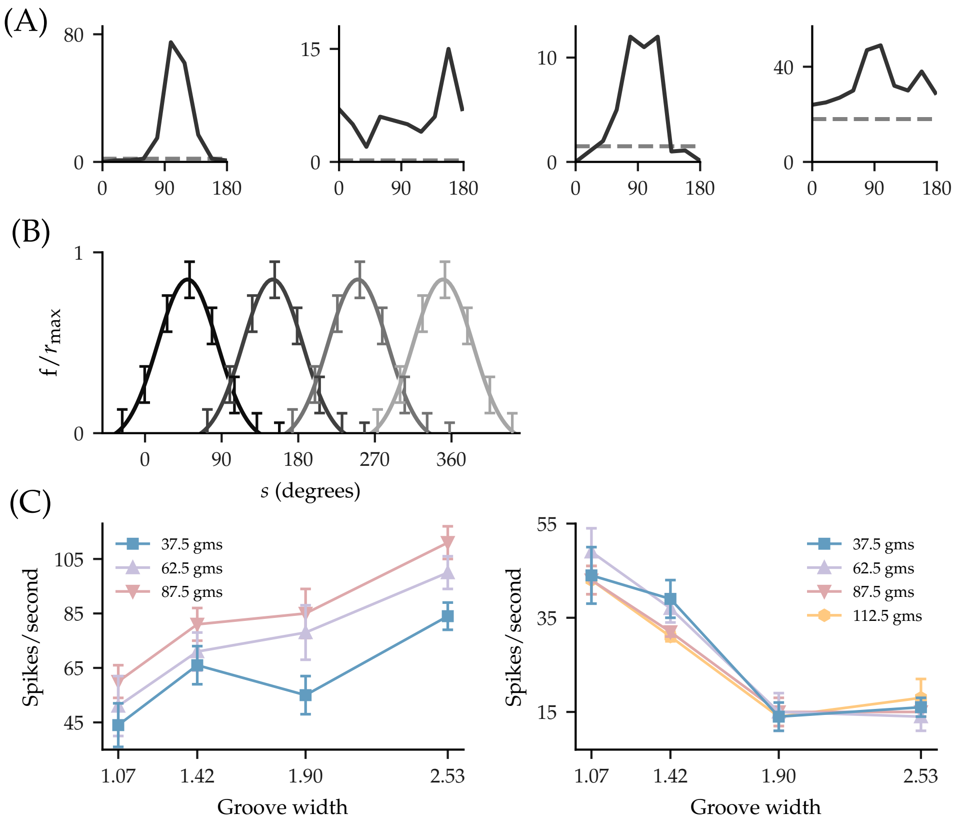


Figure 14.2: Empirical tuning curves. **(A)** Tuning curves for orientation in macaque primary visual cortex (V1). The dashed line represents the spontaneous firing rate. Reproduced from Shapley, Hawken, Ringach (2003). **(B)** Normalized tuning curves for the direction of air current in four interneurons in the cercal system of the cricket. Reproduced from Theunissen and Miller (1991). **(C)** Tuning curves for the width of the groove in a tactile grating in macaque second somatosensory cortex (S2). Different curves are for different magnitudes of the contact force (expressed as mass). Reproduced from Pruett, Sinclair, and Burton (2000).

14.2 Neural likelihood function for a single neuron

Suppose now that r spikes are observed, where r is a specific number such as 0, 2, or 11. Given r, the neural likelihood of a hypothesized stimulus value s is the probability that r spikes were produced by that value of s. In other words, we copy Eq. (14.3) but consider it as a function of s rather than r:

$$\mathscr{L}(s;\mathbf{r}) = \frac{1}{r!}e^{-f(s)}f(s)^r. \tag{14.4}$$

We consider two example cases which we will follow throughout this section:

14.2.1 Case 1: A bell-shaped tuning curve

Suppose that the tuning curve of the neuron has a Gaussian shape with peak location (preferred stimulus) $s_{\text{pref}} = 0$, width $\sigma_{\text{tc}} = 10$, baseline $b = 1$, and gain $g = 5$:

$$f(s) = ge^{-\frac{(s-s_{\text{pref}})^2}{2\sigma_{\text{tc}}^2}} + b. \tag{14.5}$$

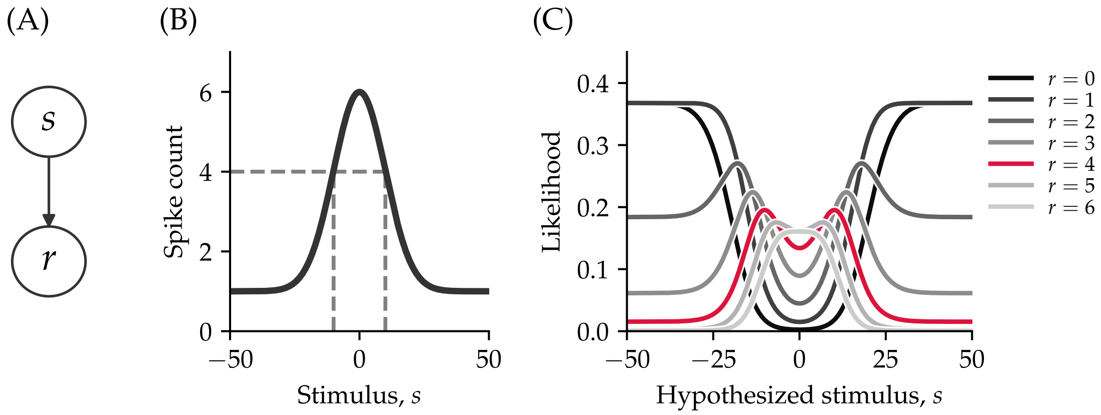


Figure 14.3: Single-neuron inference. **(A)** Generative model (to make: s in top; r in bottom. **(B)** Idealized tuning curve, $f(s)$, of a single neuron with preferred stimulus 0. **(C)** Likelihood function over the stimulus when 4 spikes are observed in a one-neuron brain with Poisson variability. **(D)** Likelihood functions over the stimulus for different observed spike counts in a one-neuron brain.

This is depicted in **Fig. 14.3B**. Note that in spite of the similarity to the Gaussian probability distribution, this is not a probability distribution! In particular, it is not normalized.

A point on the curve is the mean spike count of the neuron in response to a particular stimulus.

Suppose we are told this neuron fired 4 spikes in a given time interval, and asked what we can say about the stimulus. Based on **Fig. 14.3B**, we might say the stimulus was approximately -10 or +10, because then the neuron would produce the expected number of spikes. However, **Fig. 14.3B** only shows us the average spike count over many trials. The trial-to-trial response is noisy, as expressed by Eq. (14.3). Therefore, a total of 4 spikes could also have been produced by a stimulus value of say 3.7 – it just so happened that on this trial, the neuron fired fewer spikes than average. 4 spikes could even indicate that the stimulus was -21, although it would require that the neuron happened to fire many more spike than its average spike count at this stimulus. Clearly, some stimulus values are more likely than others, and we can define the likelihood of a hypothesized stimulus value as the probability of observing 4 spikes in response to that stimulus value.

Equation for the likelihood function. We can formalize this intuition by simply substituting the expression for the tuning curve, Eq. (14.5) into Eq. (14.4) to obtain the neural likelihood function. We have plotted the resulting function for r=4 in **Fig. 14.3C**. This odd-looking function tells us how likely each possible stimulus value is based on the observation (the spike count of 4). The shape confirms our intuition: values of (approximately) +10 and -10 are most likely, 0 is still quite likely, but -30 is very unlikely. To compute the likelihood function, we not only used the tuning curve (Eq. (14.5)) but also the form of neural variability (Eq. (14.3)). This allows us to say more about the stimulus than only that +10 and -10 are most likely.

This likelihood function has several interesting properties. First, unlike the likelihood functions we encountered in earlier, non-neural chapters, this likelihood is not Gaussian. In fact, there is no tuning curve $f(s)$ we could have used to get a likelihood function that is exactly Gaussian for every r.

A second important point is that the likelihood function in **Fig. 14.3C** is not normalized. In fact, the area under the curve is infinite! The “tails” extend to arbitrarily large values. This is simply because the tuning curve has a baseline of 1 spike, so the probability that the observed spike count was 4 is nearly the same value for s=30 as, say, for s=1000. In general, likelihood functions are not normalized. In the present chapter, as we are discussing neural models, the likelihood function over the stimulus will never be normalized.

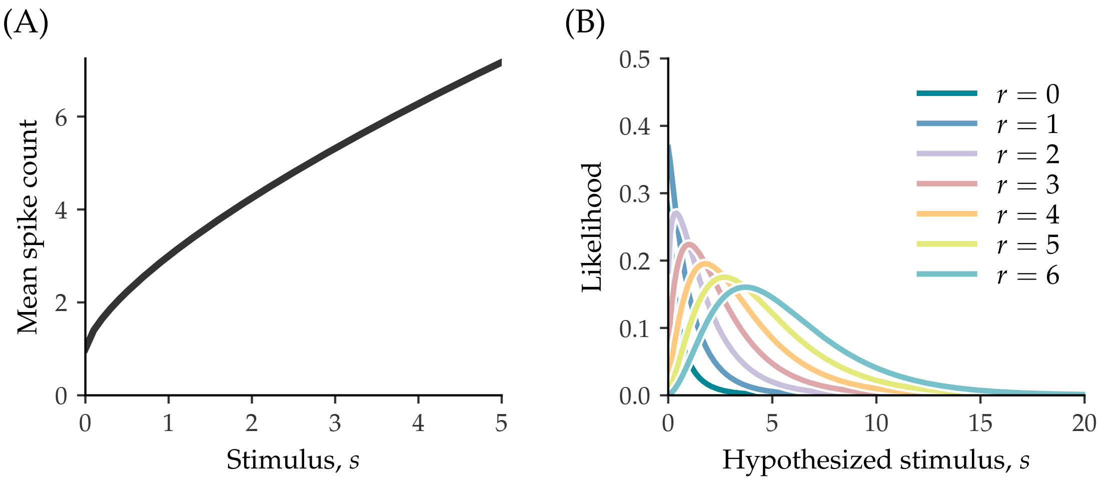


Figure 14.4: **(A)** Example of a monotonic tuning curve. We used $f(s) = as^b + c$ with $a = 2$, $b = 0.7$, and $c = 1$. **(B)** Corresponding likelihood functions, under the assumption of Poisson variability and with observations of $r = 0, \ldots, 6$. Although the tuning curve is monotonic, each of the likelihood functions is bell-shaped.

Fig. 14.3C showed the likelihood function over the stimulus when the observed spike count was 4. We can also calculate the likelihood function over the stimulus for other observed spike counts. This is done in **Fig. 14.3D**. This shows that the likelihood function can have dramatically different shapes for different observations. For example, when 0 spikes are observed, any of the central values of the stimuli are very unlikely, so the likelihood function has an inverted U-shape.

14.2.2 Case 2: A monotonic tuning curve

We have so far considered a real-valued variable with bell-shaped tuning. We will now consider a non-negative variable with monotonic tuning, as we encountered in **Fig. 14.2C**. An example would be a power law tuning curve with baseline:

$$f(s) = as^b + c, \tag{14.6}$$

where we enforce $a > 0$ and $c \geq 0$, so that mean spike counts are guaranteed non-negative regardless of s. Such a tuning curve would make sense for magnitude variables such as length, weight, contrast, and loudness. An example of such a tuning curve is shown in **Fig. 14.4A**.

What do we expect the likelihood to look like in this case? Suppose we observe that this neuron fires 4 spikes. The likelihood reflects how probable this observation is under different hypothesized values of s. From the tuning curve, we know that this neuron fires 4 spikes on average when the stimulus is $s = 6$; therefore, we expect the probability that it will fire exactly 4 spikes to be quite high for a hypothesized stimulus value of 6. On the other hand, if the stimulus were 0, then the neuron's average number of spikes would be 0.5, making it improbable for the neuron to fire 4 spikes. Similarly, if the stimulus were 100, then the neuron's average number of spikes would be 51, which would again make our observation of 4 spikes improbable. Therefore, we would expect the likelihood to be high for $s = 6$ and drop off gradually as s moves away from 6 in either direction. Thus, we expect that the neural likelihood will be bell-shaped. This intuition is not specific to a spike count of 4. In general, any particular spike count will give the highest likelihood to one stimulus value and lower likelihoods to stimuli on either side. We have plotted the likelihoods based on several different observed values of r in **Fig. 14.4B**. In line with our intuition, these likelihood functions are bell-shaped.

14.3 Neural likelihood function based on a population of neurons

The neural likelihood functions that we have encountered so far have been very wide. However, when we recall that these likelihood functions were based on the firing of just a single neuron, and that this neuron was noisy (Poisson), it is in fact remarkable how much we can already say about the stimulus. Moreover, most of us have more than one neuron in our brain, and therefore the information that we have about stimuli in the world is based on the simultaneous firing of a population of neurons.

We will now consider a population consisting of an arbitrary number of neurons; we call the number of neurons n. On a given trial, the neurons in this population will produce a set of spike counts, $r_1,\ldots,r_n$, which we will often denote shorthand by a vector $\mathbf{r}$ and call the "pattern of population activity". Mathematically, $\mathbf{r}$ is a high-dimensional vector. If 1000 neurons were selective to s, then $\mathbf{r}$ would be a 1000-dimensional vector. We assume that the variability in $\mathbf{r}$ across trials is independent across neurons conditioned on s:

$$p(\mathbf{r}|s) = p(r_1,\ldots,r_n|s) \tag{14.7}$$

$$= p(r_1|s)\cdots p(r_n|s) \tag{14.8}$$

$$\equiv \prod_{i=1}^{n} p(r_i|s) \tag{14.9}$$

As a consequence, when we observe a specific pattern of population activity $\mathbf{r}$, the neural likelihood function over s is

$$\mathscr{L}(s;\mathbf{r}) = \prod_{i=1}^{n} p(r_i|s). \tag{14.10}$$

Each factor in this product can be thought of as the likelihood function based on a single neuron's spike count. Thus, the population likelihood function is the product of single-neuron likelihood functions.

The computation of the likelihood function based on a set of neurons with independent noise is conceptually similar to cue combination as discussed in Chapter 5. One neuron's spike count is analogous to one measurement, and the likelihood (Eq. (14.10)) is obtained by multiplying the likelihoods from the individual neurons, just as the likelihoods from individual cues are multiplied together in cue combination (Eq. (5.1)). A difference is that each of the individual neuronal likelihoods is by no means Gaussian.

To make further progress, we assume that every neuron's spike count follows a Poisson distribution, but each neuron has its own mean – for the i^{th} neuron, the mean is given by the i^{th} tuning curve evaluated at s. These tuning curves are described by functions $f_i(s)$, with i ranging from 1 to n. Then,

$$p(r_i|s) = \frac{1}{r_i!}e^{-f_i(s)}f_i(s)^{r_i} \tag{14.11}$$

As a consequence, when we observe a specific pattern of population activity $\mathbf{r}$, the neural likelihood function over s is

$$\mathscr{L}(s;\mathbf{r}) = \prod_{i=1}^{n} \frac{1}{r_i!}e^{-f_i(s)}f_i(s)^{r_i}. \tag{14.12}$$

It is often convenient to take the logarithm of the likelihood function. With the independence assumption only, the population log likelihood function is equal to the sum of the log likelihood functions of the individual neurons:

$$\log \mathscr{L}(s;\mathbf{r}) = \sum_{i=1}^{n} \log p(r_i|s). \tag{14.13}$$

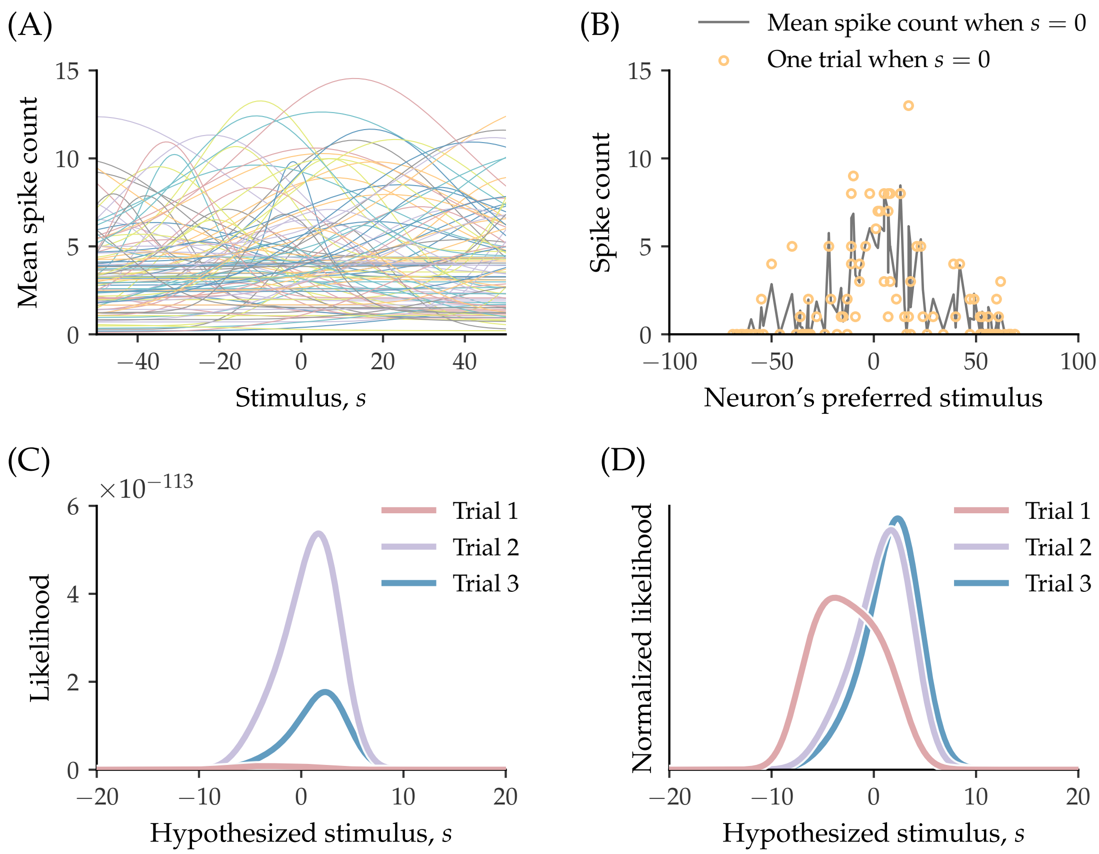


Figure 14.5: Inference based on a heterogeneous population consisting of 100 neurons. **(A)** Tuning curves. **(B)** Example pattern of activity. **(C)** Three example neural likelihood functions from this population, all obtained with $s = 0$. **(D)** Normalized versions of the likelihoods.

With both the independence assumption and the Poisson assumption, we have:

$$\log \mathscr{L}(s;\mathbf{r}) = \sum_{i=1}^{n} \log\left(\frac{1}{r_i!}e^{-f_i(s)}f_i(s)^{r_i}\right). \tag{14.14}$$

While completely equivalent to Eq. (14.12), the logarithmic form is often considered easier to work with, because it has sums instead of products. At any time, one can recover the likelihood function from the log likelihood function by exponentiating it.

14.3.1 Case 1: Bell-shaped tuning curves

We now consider the neural population from Section 14.2.1, in which the single-neuron tuning curve of Eq. (14.5) is replaced by a different tuning curve for each neuron:

$$f_i(s) = g_i e^{-\frac{(s-s_{\text{pref},i})^2}{2\sigma_{\text{tc},i}^2}} + b_i. \tag{14.15}$$

where g_i, $s_{\text{pref},i}$, $\sigma_{\text{tc},i}$, and b_i are the gain, preferred stimulus, width, and baseline of the tuning curve of the i^{th} neuron. An example is shown in **Fig. 14.5A**. Amplitude, width, and baseline are highly variable, as is the case in real recordings. (What is not yet very realistic is that each tuning curve is a member of the same, compact parametric family of functions.

We simulated three patterns of activity in the population of 100 independent Poisson neurons with tuning curves as in **Fig. 14.5A** (preferred stimuli equally spaced between -60 and 60), elicited by the stimulus $s = 0$. These patterns would correspond to neural recordings on three trials on which $s = 0$ was presented. Such a pattern could look like **Fig. 14.5B**.

Per Eq. (14.10) the likelihood over s is the product of the probabilities that neuron 1 fires 2 spikes, neuron 2 fires 3 spikes, etc.:

$$\mathscr{L}(s;\mathbf{r}) = p(r_1 = 2|s)p(r_2|s)\cdots p(r_{31} = 0|s). \tag{14.16}$$

The individual probabilities are obtained from the Poisson equation, Eq. (14.11), with Eq. (14.15) for the tuning curves. Three resulting likelihoods are plotted together in **Fig. 14.5C**.

Several properties of these likelihood functions merit discussion. First, the scale on the y-axis is very small, namely of the order of 10^{-113}. This low magnitude of the likelihood functions is not a mistake: it results from the fact that the probability of multiple events (spikes in multiple neurons) is always less than the probability of any one of those events. Every time a neuron is added, the likelihood gets smaller.

Second, even in a large population, the likelihood can be far from Gaussian (see Trial 1). Some likelihood functions have two maxima, others have a flat top, yet others are skewed. At the same time, almost all likelihood functions have a distinct, dominant peak, and most look more or less Gaussian. This is an empirical generality: the more neurons a population encoding a continuous stimulus contains (and the more they fire), the closer to Gaussian the likelihood functions look. The smooth and structured form of the likelihood function stands in contrast to the messy and apparently structureless population pattern of activity in **Fig. 14.5B**.

Third, the likelihood functions vary enormously in peak height. The one from Trial 1 is so low overall that it is not even visible on this scale. Its peak height is $8.2 \cdot 10^{-115}$. Since the shape of the likelihood function (the single peak and the narrow width) are its most important features, as these will influence the observer's conclusion as to the value of the stimulus. In Fig 14.5D, we have normalized the same three likelihood functions. Because of the normalization, the area under each likelihood function is now equal to 1. This makes the Trial 1 likelihood function clearly visible. The normalized likelihood function is a posterior distribution. In particular, it is the posterior distribution when the prior is uniform.

Exercise 14.1 Why is it the posterior distribution when the prior is uniform? ■

Besides visibility in plots, there is usually no reason to normalize the likelihood function. The shape is the same with or without normalization, and usually the shape is most important. The unnormalized likelihood function is the fundamental entity that is encoded in the neural population pattern $\mathbf{r}$.

A brief summary:

- A likelihood function over the stimulus can always be obtained from single-trial patterns of neural population activity. This likelihood function can be used by a Bayesian observer in subsequent computation.
- The likelihood function is different on every trial, even if the stimulus is kept the same. This is because the likelihood function is determined by the pattern of neural activity on a trial, and this pattern varies stochastically from trial to trial.
- Likelihood functions come in a variety of shapes and sizes (peak heights), but when the number of neurons is large, often look Gaussian.
- The neural likelihood function contains all information that can objectively be obtained from the population activity. No more information can be obtained, and any different information would be incorrect.

Box 14.2 — Representation of uncertainty or uncertainty associated with a representation?. Neural likelihood functions are sometimes described as a representation of uncertainty. However, the term “representation" in neuroscience usually refers to a world state, such as “representation of motion" or “representation of face identity". In this sense, uncertainty is a strange thing to represent, since uncertainty is a property of a belief of an observer, instead of a world state. A more accurate formulation might be that the neural likelihood function captures the uncertainty associated with a (single-trial) representation.

There is a category of models that place uncertainty exclusively in the external world. In those, uncertainty is computed purely based on the sensory input and not an internal representation or neural activity. Subsequently, the researcher then aims to find neural correlates of uncertainty. This approach is a descriptive model, not a process model, because it does not specify step by step the way sensory information gets processed. In such process models, however, it would be more justified to speak about the representation of uncertainty. We do not consider descriptive models in this book, because we take seriously the notion that the brain utilizes a generative model of its internal representations during inference. ■

14.4 Toy models

So far, our population likelihood functions were complicated and did not provide any intuition about how its properties depend on neural activity. To achieve such intuition, we will in this section consider specific examples of the bell-shaped curve families that allow for compact expressions. We emphasize that this will force us to consider biologically rather unrealistic settings; however, the intuition that we will gain will generalize. Borrowing from the language of physics, the models we describe in this section will be “toy models”: they capture the essence without being realistic.

Within the class of tuning curve defined by Eq. (14.5), we can gain intuition by assume that preferred stimuli of the neurons are equally and densely spaced across the entire real line (there are thus infinitely many neurons), their tuning curves are translated versions of each other, and have baseline 0. In other words, we use Eq. (14.15) with $b = 0$, $g_i = g$, and $\sigma_{\text{tc},i} = \sigma_{\text{tc}}$:

$$f_i(s) = ge^{-\frac{(s-s_{\text{pref},i})^2}{2\sigma_{\text{tc}}^2}}. \tag{14.17}$$

This population is depicted in **Fig. 14.6A**.

To calculate the log likelihood, we start from the expression for the log likelihood from Eq. (14.14), which we split up into separate terms:

$$\log \mathscr{L}(s;\mathbf{r}) = -\sum_{i=1}^{n} \log r_i! - \sum_{i=1} f_i(s) - \sum_{i=1} \log f_i(s). \tag{14.18}$$

The first term depends on the spike counts so will change from trial to trial, but on a given trial it is just a constant; it does not depend on s. We now make the common simplifying approximation that the sum of the neural tuning curves over neurons is independent of the stimulus:

$$\sum_{i=1}^{n} f_i(s) \approx k. \tag{14.19}$$

In general, the left-hand side will depend on s, but if tuning curves are sufficiently dense and sufficiently similar to each other, Eq. (14.19) is a good approximation. (Since we only defined tuning curves in a limited region of space, the approximation will only hold in that region; the sum will drop to zero for values of s outside this region.) For the population in **Fig. 14.6**, that is the case. For the population in **Fig. 14.5A**, however, Eq. (14.19) is not a particularly good approximation. Both are shown in **Fig. 14.6B**.

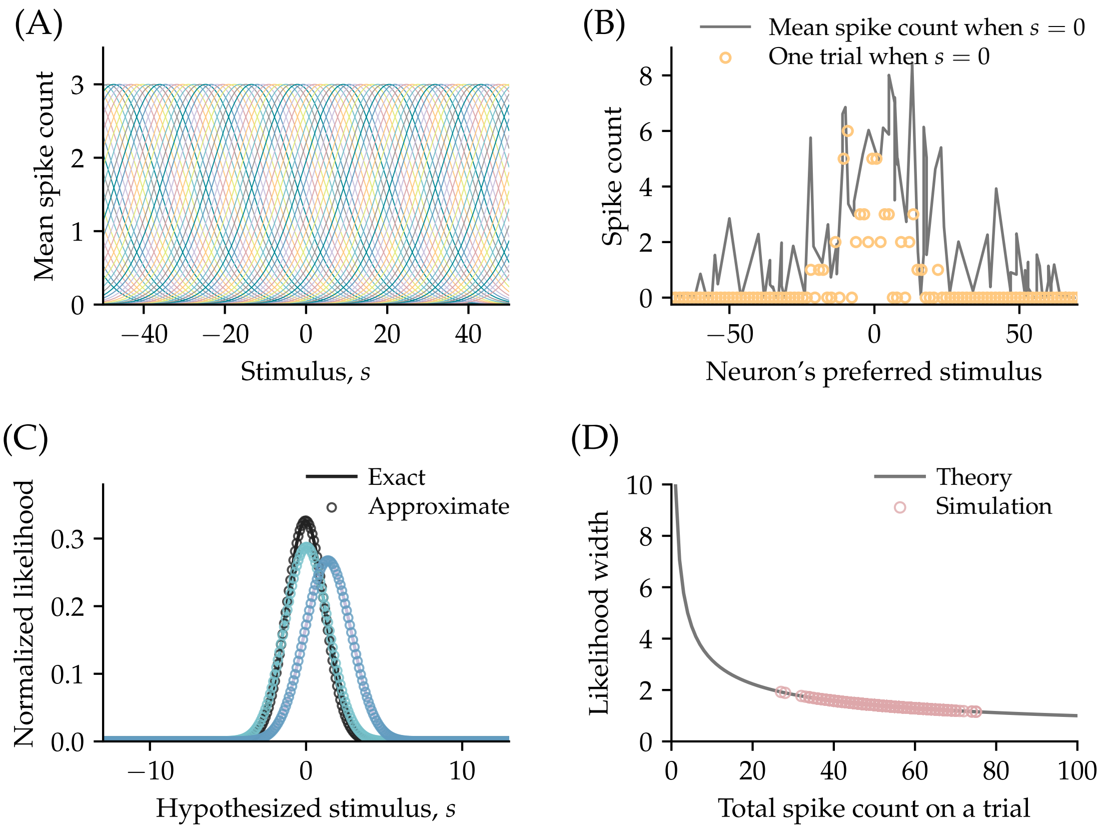


Figure 14.6: Toy model with homogeneous tuning curves. **(A)** Dense, translation-invariant tuning curves with a Gaussian shape (width 10) and baseline 0. **(B)** Mean activity of each neuron (sorted by preferred stimulus) when the true stimulus is $s = 0$ (line), with a single-trial pattern of activity (open circles). **(C)** Three example neural likelihood functions with the constant-sum approximation overlaid. The approximation is very good. **(D)** Sensory uncertainty from a neural population: Width of the neural likelihood function as a function of the total single-trial spike count in the population.

If one is only interested in the *shape* of the likelihood function, as is usually the case (for example to obtain an estimate of the stimulus, or an estimate of uncertainty), then the multiplicative constants don't matter. We can then even write the likelihood as

$$\mathscr{L}(s;\mathbf{r}) \propto \prod_{i=1} f_i(s)^{r_i}, \tag{14.20}$$

where the proportionality sign absorbs all s-independent factors. This is not the same as normalizing, since we do not compute the normalization constant. To summarize what we have found so far, using the constant-sum approximation we obtained a concise expression for the neural likelihood function, which takes the form of products of tuning curves raised to the powers of the corresponding spike counts. The higher the spike count, the higher the power, and the more influence that neuron's tuning curve has on the likelihood function. Thus, if tuning curves are single-peaked (unimodal), as in **Fig. 14.6**, then the likelihood tends to peak near the preferred stimuli of the highest-firing neurons.

In log space, Eq. (14.18) becomes

$$\log \mathscr{L}(s;\mathbf{r}) \approx -\sum_{i=1}^{n} \log f_i(s) + \text{constant}. \tag{14.21}$$

We substitute Eq. (14.17) into Eq. (14.18), and simplify to find

$$\log \mathscr{L}(s;\mathbf{r}) = \frac{1}{2\sigma_{\text{tc}}^2}\sum_{i=1}^{n} r_i(-s_{\text{pref},i})^2 + \text{constant}. \tag{14.22}$$

Exercise 14.2 a) Verify this.
b) Why was the assumption $b = 0$ important?

Undoing the log in Eq. (14.22) and simplifying, we obtain for the neural likelihood function:

$$\mathscr{L}(s;\mathbf{r}) \propto e^{-\frac{(s-\mu_{\text{likelihood}})^2}{2\sigma_{\text{likelihood}}^2}}, \tag{14.23}$$

where

$$\mu_{\text{likelihood}} \equiv \frac{\sum_{i=1}^{n} r_i s_{\text{pref},i}}{\sum_{i=1}^{n} r_i} \tag{14.24}$$

$$\sigma_{\text{likelihood}}^2 \equiv \frac{\sigma_{\text{tc}}^2}{\sum_{i=1}^{n} r_i} \tag{14.25}$$

Exercise 14.3 This is not obvious and deriving it requires several steps. Do the derivation.

There is a good reason to write the likelihood function like this: we recognize the form of an (unnormalized) Gaussian function! In other words, when a population of independent Poisson neurons have Gaussian tuning curves without baselines, and we make the constant-sum approximation, then the neural likelihood function over the stimulus is Gaussian. This property makes this special case valuable as a toy model. We plot normalized likelihood functions obtained from this population using the complete equation, overlaid with the approximate expression, Eq. (14.23) (**Fig. 14.6C**). The approximation is essentially indistinguishable from the complete expression. In addition, we notice that both the mode of the likelihood function and its width vary from trial to trial. In Chapter 3, the likelihood varied in location (mode), but not in width.

Since for a Gaussian distribution, mean and mode are the same number, the maximum-likelihood estimate is $\mu_{\text{likelihood}}$:

$$\hat{s}_{\text{ML}} = \mu_{\text{likelihood}}. \tag{14.26}$$

Thus, under the assumptions made, the ML estimate of the stimulus is a weighted sum of the preferred stimuli of the neurons in the population, with weights given by the neurons' spike counts. This is also called the *population vector decoder*.

The variance of the posterior is given by Eq. (14.25). As usual, the corresponding standard deviation is interpreted as the uncertainty that the observer has about the stimulus. This expression, plotted in **Fig. 14.6D**, makes intuitive sense: the higher the total spike count in the population, the narrower the likelihood function and the lower the sensory uncertainty. Also, the narrower the tuning curve (smaller σ_{tc}), the narrower the likelihood function. Sensory uncertainty is present in a *distributed* manner, since all neurons contribute to the sum in Eq. (14.25).

For a given stimulus, the width of the likelihood function, $\sigma_{\text{likelihood}}$, varies from trial to trial because the total spike count does. In other words, the observer's sensory uncertainty will vary substantially even as the physical stimulus is the same. The width of the likelihood function is on average lower when gain is higher. Moreover, what would have been harder to anticipate, the variation in the width is also lower when gain is higher.

Exercise 14.4 a) When one neuron fires one spike, Eq. (14.25) states that the width of the likelihood is equal to the width of the tuning curve. Explain why this is correct and why this scenario is different from the one-neuron brain discussed in Section 14.2

b) What is the likelihood function when the entire population is silent?

Eq. (14.25) implies that sensory uncertainty itself varies from trial to trial as the total spike count varies; in other words, not all likelihoods are equally wide. This stands in contrast to the first half of the book, in which the likelihood always had the same width for the same stimulus condition. However, the neural formulation is more accurate than the behavioral formulation; in real neural systems, the width of the likelihood function will vary from trial to trial.

The assumptions we made (independent Poisson; Gaussian tuning curves with zero baseline; constant-sum approximation) allowed us to write down explicit equations for the maximum-likelihood estimate and the width of the likelihood function. That we were able to do this depended strongly on these assumptions: if we had removed any one of them, no such equations could have been formulated. More generally, it is rare that closed-form expressions for the maximum-likelihood estimate and the width of the likelihood function can be written down; therefore, one typically needs to compute likelihood functions via numeric simulation.

To summarize, the case of independent Poisson neurons with Gaussian tuning curves and the constant-sum approximation is a useful toy model, because the likelihood function is exactly Gaussian and we can find analytical, intuitive expressions for both the maximum-likelihood estimate (population vector or weighted average) and the likelihood width (inversely related to the total spike count).

14.5 Relation between behavioral and neural concepts

Let us take stock of what we have learned in the previous two sections. We introduced a *generative model for neural activity*, namely independent Poisson variability combined with certain assumptions about tuning curves. Using this generative model, we computed the *neural likelihood function* based on a population pattern of activity **r**. We found specific expressions for the maximum-likelihood estimate and the width of the likelihood function. At this point, it is useful to compare and contrast these results with the generative model we introduced in Chapter 3. We have put corresponding quantities in the table below:

We can see from this table that the behavioral model was simplified in several ways: first, the likelihood function was always Gaussian, while in the neural model, it is not. Second, the maximum-likelihood estimate is identical to the measurement, which is made possible by the fact that the measurement lives in the same space as (has the same domain as) the stimulus; in the neural model, the observation lives in a completely different space (the space of n-dimensional vectors of positive integers) than the stimulus, and therefore also than the maximum-likelihood estimate of the stimulus. Third, in the behavioral model, the likelihood width was the same from trial to trial; in the neural model, since it depends on the observation, it varies from trial to trial.

Looking back at the behavioral models, we can appreciate now that the concept of a measurement was an abstraction. The brain itself does not have scalar measurements with which to do inference; it only has neural action potentials. In fact, we could now *define* the concept of a measurement in terms of the neural model: the measurement is the maximum-likelihood estimate of the stimulus based on the neural observation, namely the population pattern of activity **r**. In this view, the measurement x of a stimulus s is the value of s under which the observed neural activity **r** is most probable. We can think of the measurement x as a “processed form” of the neural activity **r**. This is illustrated in **Fig. 14.7**. In this view, the Gaussian distribution to describe $p(x|s)$

Quantity	**Behavioral model**	**Neural model**
Observation	Scalar measurement x. Possible values: same as of stimulus s	Vector of spike counts $\mathbf{r} = (r_1, \dots, r_n)$. Possible values: positive integers
Noise distribution	$p(x \mid s)$, typically Gaussian with mean s and standard deviation σ	$p(\mathbf{r} \mid s)$, for example independent Poisson with Gaussian tuning curves and a constant-sum approximation
Likelihood over s	$\mathscr{L}(s;x) = p(x \mid s)$, Gaussian if noise distribution is Gaussian	$\mathscr{L}(s;\mathbf{r})$, Gaussian in the example but not in general
Maximum-likelihood estimate	x (variable)	In the example, $\frac{\sum_{i=1^n r_i s_{\text{pref},i}}}{\sum_{i=1^n r_i}}$ (variable)
Width of likelihood function	σ (fixed)	In the example, $\frac{\sigma_{\text{tc}}}{\sqrt{\sum_{i=1^n r_i}}}$ (variable)

Table 14.1: Comparison between behavioral and neural Bayesian quantities in the toy model of Section 14.4.

is an approximation to the neural distribution of the maximum-likelihood estimate, which we will examine in the next section.

The internal representation may directly relate to the readings of one of our primary sensors. For example, the pattern of activation of photoreceptors in the retina or of the mechanoreceptors embedded in our skin may be considered an internal representation. Alternatively, the internal representation might be associated with activity in a downstream brain area. For example, the activity in primary visual cortex is also an internal representation.

The neural activity elicited by a stimulus will vary randomly from trial to trial, even when the physical stimulus itself is identical each time. Thus, we say that the internal representation of a stimulus is noisy. As we discussed in Chapter 1, noise originates from many sources, including thermal noise, photon shot noise, neurotransmitter release, and ion channel opening and closing. We define a *measurement* as the best possible guess about the stimulus based on the internal representation alone. "Best possible" here means that this value has the highest probability of generating the observed internal representation. If the stimulus is an orientation of a line and the internal representation is the pattern of activation of retinal photoreceptors, then the measurement would be the best guess of orientation obtained from this pattern. If the stimulus is the size of an object and the internal representation is the pattern of activation of skin mechanoreceptors when holding the object, then the measurement would be the best guess of size based on this pattern. In the auditory localization example, the measurement could be the best guess of location based on the activation of the hair cells in the inner ears.

Since the internal representation is noisy, the measurement will be noisy as well. The full internal representation typically occupies a high-dimensional space that is very different from the stimulus space; it could for example be a space of neural activity, such as the firing rates of a population of sensory neurons in the cortex. One could think of the mapping from internal representation to measurement as "pre-processing", since the computation we will focus on takes the measurement(s) as input.

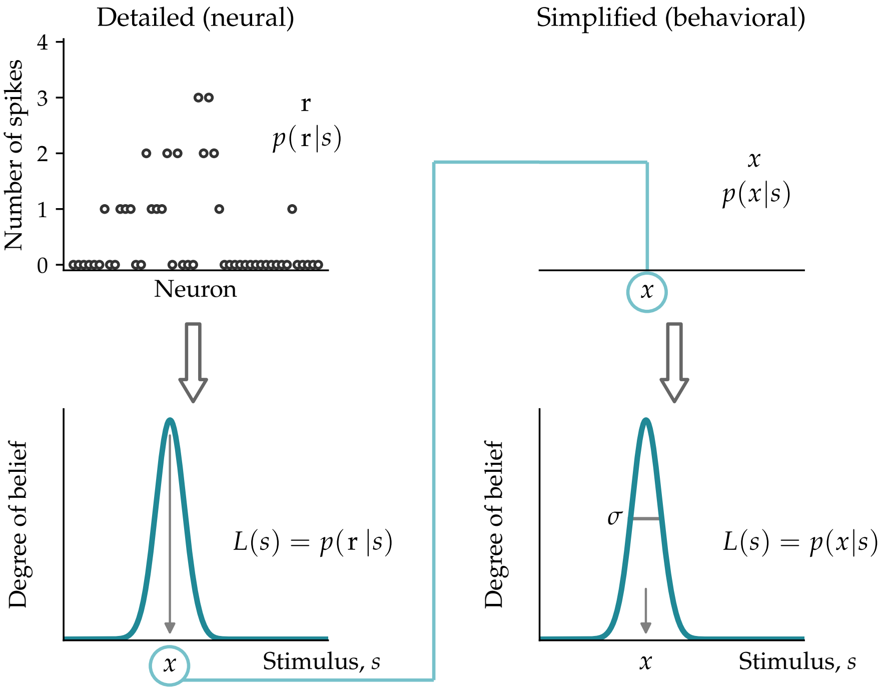


Figure 14.7: Relationship between the neural likelihood function and the behavioral likelihood function.

14.6 Statistics of likelihood mode and width across many trials

Step 3 of the Bayesian modeling framework outlined in Chapter 4 and followed throughout the behavioral chapters was to predict behavior across many trials, or, to be more precise, to calculate the distribution of the estimate conditioned on the stimulus. In the neural context, this distribution almost always has to be computed numerically.

There is a particular inference task that was trivial in the behavioral context but is not in the neural one, which is simply to estimate the stimulus on a continuum under a uniform prior. In that case, the observer's MAP estimate is the maximum-likelihood estimate and the width of the posterior is equal to the width of the likelihood function. In the behavioral model, the maximum-likelihood estimate was x, so the distribution of the maximum-likelihood estimate given the stimulus was the same as the noise distribution, namely $p(x|s)$. To be mathematically precise,

$$p(\hat{s}|s) = p_{x|s}(\hat{s}|s). \tag{14.27}$$

The likelihood function on every trial, the posterior distribution on every trial, and the distribution of the maximum-likelihood estimate all had standard deviation σ in this case.

In the neural toy model of Section 14.4, the maximum-likelihood estimate is $\mu_{\text{likelihood}}$ and the width of the likelihood function is $\sigma_{\text{likelihood}}$. Both quantities are random variables, with distributions that they inherit from the distribution of the spike counts $\mathbf{r}$. We will now examine the distributions of the mode and the width of the likelihood function over many trials, so that $\mathbf{r}$ follows Eq. (14.11). Neither distribution can be calculated analytically, and they are not theoretically Gaussians. Only in the limit that gain is very high (high spike counts on average) can one state that statistically, the distribution of the maximum-likelihood estimate becomes indistinguishable from a Gaussian. This is a property known as asymptotic normality; it is essentially a limit when the amount of information in the population is large, because a sufficient number of neurons has responded with a sufficient number of spikes. If this condition is not satisfied, it is not guaranteed that the Gaussian

Quantity	Behavioral model	Neural model
Distribution of MLE	Gaussian with mean s and standard deviation σ (same as noise distribution)	No analytical form; only Gaussian in the limit of many spikes
Distribution of likelihood width	Delta function at σ (always the same)	Wide distribution (no analytical form)

distribution is a good approximation to the true distribution of the maximum-likelihood estimate. Under similar conditions, it can be shown as well that the mean value of the maximum-likelihood estimate over many trials is equal to the stimulus itself, i.e.

$$\mathbb{E}[\hat{s}_{\mathrm{ML}}|s] = s. \tag{14.28}$$

The technical statement is "the maximum-likelihood estimate is asymptotically unbiased". These properties provide some degree of justification for our assumptions in the behavioral models that the measurement has a Gaussian distribution and is on average equal to s. However, it is important to keep in mind that both are approximations to the underlying neural model.

Besides looking at the distributions of the mode and the width of the likelihood function separately, we can also look at the trial-to-trial correlations between both quantities. Intuitively such a correlation would mean that on trials when you are less certain, you also perform worse. We will examine this in a Problem.

14.6.1 Distinction between $\mathscr{L}(s;\mathbf{r})$ and $\mathscr{L}(s;I)$

Having discussed the neural likelihood function provides us with a basis to formulate a neurally based framework for perception. The general structure of a perception model we described in Chapter 3, consisted of a stimulus s, a measurement x, and a stimulus estimate. This was a simplified conceptualization. In describing the auditory task there, we somewhat loosely referred to sound location as the "stimulus", ignoring its neural component.

The likelihood over s based on $\mathbf{r}$, denoted $\mathscr{L}(s;\mathbf{r})$, was the focus of this chapter so far. We could also have computed a likelihood function over s from the sensory input I rather than the neural activity $\mathbf{r}$:

$$\mathscr{L}(s;I) = p(I|s). \tag{14.29}$$

This function represents all information that can be obtained about s from the sensory input I. It is not necessarily the same as $\mathscr{L}(s;\mathbf{r})$. In general, information will be lost between the sensory input I and $\mathbf{r}$, as mentioned above. As a consequence, $\mathscr{L}(s;\mathbf{r})$ will be wider than $\mathscr{L}(s;I)$, reflecting a greater amount of uncertainty.

Descriptive models that correlate brain activity with uncertainty typically consider uncertainty a state of the world. In other words, they derive their measure of uncertainty from a likelihood $\mathscr{L}(s;I)$.

14.7 Using the neural likelihood function for computation

So far in this chapter, we defined the neural generative model, which is Step 1 of the Bayesian modeling recipe we outlined in Chapter 3. We have described part of Step 2, namely the likelihood function over the stimulus based on the activity in a typical sensory population. We know from previous chapters in which sensory noise played a role that this likelihood function can be used in a multitude of ways: in combination with prior information (Chapter 3), in combination with other likelihood functions (Chapter 5), or to infer a higher-level categorical variable (e.g. Chapters 7, 8, 10, 11). In all these cases, each likelihood function over a stimulus is an elementary building block that is used to build the posterior distribution over the world state of interest.

The situation is exactly the same for the neural likelihood. Wherever before the present we used a stimulus likelihood function $\mathscr{L}(s;x) = p(x|s)$, we can now replace it by a neural likelihood function, $\mathscr{L}(s;\mathbf{r}) = p(\mathbf{r}|s)$, and everything else would go through as before.

As an example, we discuss the combination of a neural likelihood function with a Gaussian prior $p(s)$,

$$p(s) = \mathscr{N}(s;\mu,\sigma_s^2). \tag{14.30}$$

The posterior distribution over s is

$$p(s|\mathbf{r}) \propto p(\mathbf{r}|s)p(s) \tag{14.31}$$

Under the toy model of Section 14.4, but not otherwise the likelihood is proportional to a Gaussian with mean given by Eq. (14.24) and standard deviation by Eq. (14.25). Then we are exactly back to the case of Chapter 3, with the only difference the substitutions from Table 14.1 for x and σ^2. Thus, we can immediately import the equation for the posterior mean estimate from Chapter 3, Eq. (3.23), and make these substitutions, so that

$$\hat{s}_{\text{PM}} = \frac{\frac{1}{\sigma_{\text{tc}}^2}\sum_{i=1}^n r_i s_{\text{pref,i}} + \frac{\mu}{\sigma_s^2}}{\frac{1}{\sigma_{\text{tc}}^2}\sum_{i=1}^n r_i + \frac{1}{\sigma_s^2}}. \tag{14.32}$$

This equation provides the answer to the question: if the sensory input activity is $\mathbf{r}$ and the stimulus is drawn from $p(s)$, how does the brain make the best possible estimate of the stimulus? In other words, it is a neural Bayesian stimulus-response mapping.

The inference problems in other chapters involving a noisy measurement x can be treated using the same substitution. These do not contribute new understanding, so we will not go into them. However, one has to ask how these downstream computations are implemented. It is one thing to write down a formal expression for the posterior distribution of a quantity of interest, another to specify neural operations that can compute this posterior distribution. One could approach this problem roughly in two ways. One is to manually construct such operations, and demonstrate through maths and simulations that these operations suffice. Another is to train an artificial neural network to learn a mapping consisting of simple building blocks (linear operations and point nonlinearities). The former, as illustrated by Ma et al. 2006, could be theoretically more satisfying but the assumptions that are needed to do the math place severe limitations on the task and the form of neural variability. The latter, as illustrated by Orhan and Ma 2017, pretty much always works but might lead to a network that is hard to interpret.

Box 14.3 — Artificial Neural Networks to obtain likelihoods when data depends on the variables of interest in a nonobvious way. In some cases, for example those where neurons with all their properties are involved, it is hard to formulate a good likelihood function. In such cases we may want to use deep learning to obtain a likelihood function, which may be e.g. a Gaussian with data dependent mean and variance. Let us say that the measured data can be characterized by a vector x of variables. For example, x may be the set of activities of each of the neurons. We may want to train a deep learning systems to estimate the mean with one neural network defined function $f_\theta(\mathbf{x})$ and the standard deviation with another neural network defined function $g_\theta(\mathbf{x})$ for a given data point. And then optimize with gradient descent for the following loss derived from the log likelihood of a Gaussian:

$$\log \text{Loss} = \frac{(y - f_\theta(\mathbf{x}))^2}{2g_\theta(\mathbf{x})} + \log g_\theta(\mathbf{x}) \tag{14.33}$$

This leads to the system optimizing its estimates of mean an variance simultaneously in a data driven way. However computed, such a likelihood can readily be combined with a system that is Bayesian otherwise. ■

14.8 Applications

Although the focus of this book is how the brain performs inference based on noisy sensory information, there is a parallel literature on how an experimenter can decode brain state on a trial-to-trial basis. Traditionally, this literature has focused on point estimates; however, in recent years, more attention has been given to decoding entire likelihood functions – thus linking to the rest of this chapter.

Functional magnetic resonance imaging (fMRI) is a method that uses big magnets and microwaves to measures the three dimensional oxygenation of blood in the head due to neural activity. The brain is divided into voxels (like pi*xels*, but *vo*lume elements – small cubes), and one records "percent signal change" in each voxel in response to presented stimuli. Of natural interest is an application of "mind-reading" of decoding what the brain thinks or sees based on fMRI data. The logic here is a bit different from the rest of the book, since our model is not a model of the observer. After all, human subject's observations are the activations of neurons, not voxels. Experimenters, however, use the generative model for voxel activity to decode the stimulus. Thus, the observer is the experimenter, not the human subject. Just like the observer must first learn the parameters of the generative model of its observations to perform inference, the experimenter has to learn the parameters of the generative model of voxel activity. For that purpose, we use training data, in which the stimulus is considered known on every trial. Learning the parameters can be done using maximum-likelihood estimation, but it is actually better to maintain a posterior distribution over parameters, similar to the posterior over σ in Section 6.3. Then, using the learned parameters, one can decode a posterior distribution over the stimulus from the voxel activity on new trials.

The calculation of posterior distributions based on neural activities for perception has a close analog in the technical problem of decoding in the context of brain machine interfaces. Following a spinal cord injury that results in paralysis, a person might be outfitted with prosthetic arms. But how can these prosthetic limbs be controlled? One possibility is to use voice commands, but this procedure is cumbersome. Another possibility is to use eye movements. A third possibility, that at first seems worthy of a science-fiction story, is to control the prosthetic device with ones thoughts. This can be accomplished through the use of a Brain-Machine interface that reads the user's intent from their neural activity. Recorded neural signals from the motor cortex, properly interpreted via Bayesian inference, indicate her intended movements. In such BMI scenarios, we are interested in calculating the posterior distribution over the movement intent, given the recorded neural activity.

14.9 Summary and remarks

In this chapter, we have extended the discussion of generative models to neural activity and introduced the concepts of neural likelihoods and neural uncertainty. We have learned:

- Neural variability can be a source of uncertainty.
- The generative model of the activity $\mathbf{r}$ in a neural population in response to a stimulus s includes tuning curves, neural variability, and a relationship between neurons.
- After assuming conditional independence for the relationship between neurons, we needed to define a probability distribution $p(r_i|s)$ for each neuron i. We assumed Poisson variability.
- The rate parameter of the Poisson distribution is the mean spike count of the neuron in response to the stimulus. It is defined by the tuning curve of the neuron.
- By inverting the generative model so defined, we can obtain neural likelihood functions $\mathscr{L}(s;\mathbf{r})$. These are often not Gaussian.

- Neural likelihood functions can be used for further computation, in the same way as we used non-neural likelihood functions $\mathscr{L}(s;x)$ in previous chapters.

14.10 Suggested readings

- Christopher R Fetsch et al. "Neural correlates of reliability-based cue weighting during multisensory integration". In: *Nature neuroscience* 15.1 (2012), pages 146–154
- Peter Földiák. "The 'ideal homunculus': statistical inference from neural population responses". In: *Computation and neural systems*. Springer, 1993, pages 55–60
- Mehrdad Jazayeri and J Anthony Movshon. "Optimal representation of sensory information by neural populations". In: *Nature neuroscience* 9.5 (2006), pages 690–696
- Wei Ji Ma et al. "Bayesian inference with probabilistic population codes". In: *Nature neuroscience* 9.11 (2006), pages 1432–1438
- A Emin Orhan and Wei Ji Ma. "Efficient probabilistic inference in generic neural networks trained with non-probabilistic feedback". In: *Nature communications* 8.1 (2017), pages 1–14
- Terence David Sanger. "Probability density estimation for the interpretation of neural population codes". In: *Journal of neurophysiology* 76.4 (1996), pages 2790–2793
- Ruben S Van Bergen et al. "Sensory uncertainty decoded from visual cortex predicts behavior". In: *Nature neuroscience* 18.12 (2015), pages 1728–1730
- Edgar Y Walker et al. "A neural basis of probabilistic computation in visual cortex". In: *Nature Neuroscience* 23.1 (2020), pages 122–129
- Richard S Zemel, Peter Dayan, and Alexandre Pouget. "Probabilistic interpretation of population codes". In: *Neural computation* 10.2 (1998), pages 403–430

14.11 Problems

Problem 14.1 Are the following statements true or false? Explain.

a) The closer a stimulus is to the preferred stimulus of a Poisson neuron, the lower is the response variance of this neuron when the stimulus is presented repeatedly.
b) When neurons have similar and equally spaced tuning curves, then the mean population pattern of activity in response to a stimulus has the same width as the tuning curve.
c) When neurons have similar and equally spaced tuning curves, then the neural posterior has the same width as the tuning curve.
d) The variance of a single neuron responding to a stimulus can be determined from the value of its tuning curve at that stimulus value.
e) In any population, the variability of population activity is known if one knows the variability of each single neuron.

Problem 14.2 We assume a population of 9 independent Poisson neurons with Gaussian tuning curves and preferred orientations from -40 to 40 in steps of 10. The tuning curve parameters have values g=10, b=0, and tc=20. A stimulus s=0 is presented to this population. What is the probability that all neurons stay silent?

Problem 14.3 In Section 14.4 (toy model), we assumed zero-baseline Gaussian tuning curves.

a) What changes if the baseline is not zero?
b) For each of the baseline values 0, 0.25, 0.5, and 1, numerically compute and plot 10 likelihood functions (assume the true stimulus is zero), and describe what you observe.

Problem 14.4 In the toy model, we assumed the same tuning width σ_{tc} for all neurons. Derive the equivalent of Equations and for the mode and the width of the likelihood function if every neuron has its own tuning width, say $\sigma_{\text{tc},i}$ for the i^{th} neuron.

Problem 14.5 Assume a population of independent Poisson neurons responding to a non-negative stimulus s. Each neuron has a linear tuning curve, i.e.

$$f_i(s) = a_i(s). \tag{14.34}$$

Show that for any pattern of activity in this population, the neural likelihood function is a gamma distribution, and find expressions for its parameters.

Problem 14.6 Here, we explore whether the width of the likelihood function might be correlated with the error of the maximum-likelihood estimate.

a) Simulate the toy model for 10,000 trials (all at $s = 0$, and gain 1) and create a scatter plot of the squared error of the maximum-likelihood estimate versus the squared width (variance) of the likelihood function. What is the correlation coefficient?

b) Choose several different values of gain. Does the strength of the correlation depend on gain?

c) Now divide the trials into four quartiles for the variance of the likelihood function. Compute the variance of the maximum-likelihood estimate for each of these four groups of trials. Is there a correlation?

Problem 14.7 In Chapter 9, we considered top-level nuisance variables. When such a variable, say c, is unknown, the Bayesian observer would marginalize over c:

$$\mathscr{L}(s;\mathbf{r}) = p(\mathbf{r}|s) = \int p(\mathbf{r}|s,c)p(c)dc. \tag{14.35}$$

Show in the toy model that the likelihood mode and width only depend on $\mathbf{r}$, not on the prior over contrast, $p(c)$.

Problem 14.8 Some stimulus variables, such as motion direction, are periodic (directional). We can think of such variables as taking values on a circle, for example from $-pi$ to π radians. Consider a laboratory experiment in which motion direction is drawn from a Von Mises distribution with circular mean μ_s and concentration parameter κ_s:

$$p(s) \propto e^{\kappa_s \cos(s-\mu_s)} \tag{14.36}$$

a) Assume that motion direction, denoted s, is encoded in a population of n independent Poisson neurons. The tuning curve of the i^{th} neuron has a Von Mises shape with gain g, preferred direction $s_{\text{pref},i}$, and concentration parameter κ_{tc}:

$$f_i(s) = g e^{\kappa_{\text{tc}} \cos(s-s_{\text{pref},i})}. \tag{14.37}$$

Show that the likelihood function over the stimulus based on a population pattern of activity in this population, $\mathbf{r} = (r_1, \ldots, r_n)$, is proportional to a Von Mises distribution,

$$\mathscr{L}(s;\mathbf{r}) \propto e^{\kappa_L \cos(s-\mu_L)}, \tag{14.38}$$

and find expressions for $\cos\mu_L$, $\sin\mu_L$, and κ_L, each in terms of the r_i's.

b) To which well-known population decoder is the maximum-likelihood estimator μ_L obtained in part (a) equal?

c) Draw 2000 motion directions s from the stimulus distribution described in part (c) with $\mu_s = 0$ and $\kappa_s = 4$. Each drawn stimulus represents one experimental trial. For each trial, draw a pattern of activity of the population described in part (a) in response to the motion direction on that trial; assume $g = 0.2$, preferred directions at every multiple of 10°, and $\kappa_{\text{tc}} = 1$.

d) Then, again for each trial, compute μ_L and the mean of the posterior. You may use the expressions obtained in part (a). If you did not solve those parts, you can do the computations numerically. Whichever programming language you use, there is probably an implementation of the inverse tangent function.

e) Create a figure consisting of 2 by 2 subplots. The top left subplot should show a scatterplot of μ_L against s. The top right subplot should show a scatterplot of the mean of the posterior against s. Both subplots should have both an x-range and a y-range from $-\pi$ to π; also draw the diagonal as a dashed black line, for reference. The bottom left subplot should show a histogram of the value of the utility function in part (e) when $\hat{s} = \mu_L$ (maximum-likelihood estimation). Use 20 bins. The bottom right subplot should show the same histogram but with $\hat{s}$ being the mean of the posterior.

f) Does the maximum-likelihood estimate or the posterior mean have a higher expected utility? Which correlates better with the true stimulus? Are these properties expected?

15. Bayesian models in context

In this concluding chapter, we describe how Bayesian models of behavior connect to bigger themes in the field of behavioral research and beyond. Along the way, we will discuss current limitations and promising future directions for the field.

Plan of the chapter

We will first suggest that Bayesian behaviors can plausibly be expected to emerge over lifetime and evolutionary time scales. However, it is also reasonable to expect that not all Bayesian behaviour is optimal and indeed that not all behavior is Bayesian. We will discuss how Bayesian models of behavior differ from and relate to other classes of popular models within cognitive science and neuroscience. We will discuss how Bayesian models differ from models of learning how how they can be experimentally distinguished. We will conclude with several open areas of research: dealing with real-world complexity, and approximate inference.

15.1 Bayesian versus optimal behavior

Animals, including humans, are faced with consequential decision problems throughout their lives. Accordingly, We expect that adaptive behaviors will tend to be selected over both lifetime and evolutionary timescales. Among the decision problems that animals encounter in the real world, a great many will involve uncertainty. For this large set of problems, the optimal solutions are Bayesian. We therefore expect that humans and other animals will tend to use Bayesian-based approaches and that many behaviors will tend to conform to the predictions of Bayesian models.

The use of Bayesian inference is not synonymous with optimal behavior. On the one hand, some optimization problems do not require a Bayesian approach. Such problems include finding an unbeatable way to play tic-tac-toe, a fast way to sort a list, or the way to pack the largest number of apples into a crate. However, such problems are often relatively artificial computer science problems without great ecological relevance to organisms. In most ecologically important problems, uncertainty about world states is not far away.

On the other hand, a decision-maker may be fully Bayesian but suboptimal. For example, when faced with an unfamiliar task, they might apply incorrect priors or likelihood calculations, leading to model mismatch (see Section 3.5). In such cases, the decision-maker's behavior will typically be suboptimal with respect to the generative model characterizing the current task. It would still be optimal with respect to the counterfactual assumed generative model. A Bayesian model is still useful in this case, e.g. to attribute the suboptimality to underlying factors, for example, a wrong belief about the width or shape of the prior based on past experience.

It can also happen that behavior is not optimal with respect to the generative model of a laboratory task, but it is with respect to a generative model in the natural world. In those cases, it would be misleading to call the behavior suboptimal.

Exercise 15.1 Take three of your favorite papers in the broader space of Bayesian models of behavior and clarify for yourself which exactly are its assumptions. What is assumed optimal? How is the model justified? ■

15.2 Overly strong claims of optimality

We might plausibly expect behavior to be close to optimal for the kinds of problems that most mattered over evolutionary or life-time scales. Many studies have addressed the question of whether a particular behavior is close to the optimal Bayesian solution. To claim that a behaviour is close to optimal is often attractive, because this provides a crisp, principled account of the behavior, whereas dissecting suboptimality tends to be a messier business. Thus, it is perhaps not surprising that authors occasionally play somewhat loose with the notion of optimality. This could take the form of adding ad-hoc mechanisms in order to "salvage" optimality, or of only performing a qualitative comparison between data and model. Bayesian models are inherently quantitative, and every effort should be made to make detailed quantitative comparisons. A special kind of questionable optimality claim is when the prior distribution is fit to the data rather than obtained from real-world statistics. The strongest justification for a model occurs when the priors and likelihoods derive from either natural statistics or experimental statistics that the participants have been provided the opportunity to learn.

The specification of the task should also include a meaningful cost function. In models of categorical perceptual tasks, such as the ones considered in Chapters 7 to 12, the universal and, arguably, natural, assumption is that the cost function is (negative) proportion correct. This is called the 0-1 cost function and optimizing it gives rise to MAP estimation. In perceptual tasks that require continuous estimation, there is already more flexibility, but the dominant assumption is that the cost is the squared estimation error, which gives rise to posterior mean estimation. In reality, the choice of the squared error is often one of convenience and should be questioned. In other tasks, explicit monetary rewards are provided and a valid model would assume maximization of expected reward; a complication here is a potential nonlinearity between reward and subjective utility (Section 13.4.2). Adding to these difficulties, many situations exist in which the cost function is unclear; for example, what is the cost of moving your body in a certain way when catching a ball, or the cost of misunderstanding someone in a conversation? An area for future research is to investigate the cost function for realistic tasks.

Given occasional overly strong claims of optimality, it is not surprising that there has been some skepticism of such claims, and by extension, of Bayesian models [30, 61]. "Bayesian papers can make everything sound optimal." While this is true about Bayesian models without any constraints, it simply is not true for Bayesian models that derive from a specification of the task that people actually solve. To the extent that investigators can identify the actual likelihoods, priors, and cost function for a task, the degrees of freedom in a Bayesian model diminish and can even disappear altogether.

Exercise 15.2 Do you know of any models that you feel make excessive claims of optimality? How would their message be affected if they removed those claims? ■

15.3 Understanding why some behaviors are optimal and others are not

If behavior is close to optimal in some cases and not in others, what distinguishes these situations? A common notion is that behavior in perceptual and motor tasks tends to be optimal, whereas behavior in cognitive tasks tends to be suboptimal.

Arguments for the suboptimality of cognition typically revolve around the length of the list of cognitive biases and fallacies that humans exhibit [7, 8], from the gambler's fallacy to the anchoring fallacy to confirmation bias to the availability heuristic. On very general grounds, such a distinction between perception and cognition seems plausible. After all, perceptual systems are evolutionarily older and therefore have had more time to become optimized. On closer look, however, the notion of optimality in these cognitive effects are very different in many ways than ones on the perceptual and motor side. For example, the cognitive biases typically involve explicit manipulations of probabilities. Replacing probabilities by frequencies reduces or eliminates some biases; in addition, the specific wording sometimes matters a lot. Also, cognitive fallacies are only crudely measured at the individual-subject level: a subject either gives the “biased" or the “unbiased" answer, yielding no insight into the individual's decision-making process.

By contrast, using parametric measurements in more robust experimental paradigms, human reasoning and learning are often found to be well accounted for by Bayesian models at least at a qualitative level (e.g. [21, 2, 43, 53, 54, 90, 123]). Optimality, however, is much harder to establish in the context of high-level cognitive tasks.

An interesting approach is to package what is essentially the same task in both a perceptual and a cognitive framing, and comparing human behavior between them. Julia Trommershäuser and her colleagues showed that if a lottery tasks in which people exhibit forms of suboptimality gets translated into a rapid movement tasks, people are close to optimal in the sense of maximizing expected reward [105]. Similarly, Stephanie Chen and colleagues found that in a task in which optimal behavior required marginalizing over categories, people would fail to do so when giving verbal reports but not when catching an object [37].

15.4 Bayesian models are not mechanistic models of brain function

Bayesian models have been criticized for being “as if" models, in the sense that humans might follow something close to the Bayesian decision rule, but the Bayesian computations leading up to that rule might not have any grounding in reality. This statement ultimately is a statement about what scientists want to explain.

Bayesian models of behavior are not mechanistic, in the sense that they do not specify the neural operations that implement the Bayesian computations. That being said, there is an abundance of work on the neural implementation of Bayesian inference. We already briefly discussed some in Chapter 14. Most likely, we are dealing with *multiple realizability*, in the sense that the same Bayesian computation can be implemented by neurons in multiple (and potentially infinitely many) different ways. This is related to the concept of *separation of scales* in physics: the macro scale (the inference computation) can be modeled separately from the micro scale (the neural implementation).

One place where Bayesian models might intersect with mechanism is in *resource-rational models*. This is a category of models that deviate from purely Bayesian models in that the cost of representation or computations is taken into account. In some cases, this cost is biologically motivated, for example as corresponding to the number of neurons devoted to a computation or to the total activity of a population of neurons.

15.5 Probabilistic computation and hybrid models

Bayesian transfer studies of the likelihood demonstrate that uncertainty is computed by the brain and taken into account in decisions. To distinguish this from Bayesian inference more generally, we have previously named such computation *probabilistic computation*. Combining the need to compare Bayesian models against suboptimal alternatives with the notion of probabilistic computation, we arrive at an important category of hybrid models, namely models of suboptimal probabilistic computation. These are models in which stimulus uncertainty is represented by the brain on a trial-by-trial basis, but then used suboptimally in subsequent computations. For example, in Section 8.6, we derived a Bayesian decision rule for a categorization rule; this rule, one form of which is Eq. (8.38), depends on the level of sensory uncertainty σ. However, it is possible that the brain takes uncertainty into account but according to a different rule. In several visual decision-making paradigms, evidence for suboptimal probabilistic computation has been found. The suboptimality may or may not be Bayesian under a different generative model than the one dictated by the task structure – this is usually difficult to establish. As long as it is not established that the decision rule is Bayesian, the model should be thought of as hybrid, with a Bayesian front end and a non-Bayesian back end. In neural studies, one could in principle trace the propagation of the likelihood function across brain regions involved in different parts of a computation, but to our knowledge, this has not been done.

15.6 Learning to be Bayesian

In this book, we have generally modeled the brain as *already Bayesian* and have not focused on the question of how the brain learns to be Bayesian. This is a hard question. In Section 6.3, we discussed a Bayesian approach to learning the parameter of a Bayesian model, which is philosophically consistent. However, this approach had limitations. First, it was only about learning a single variable. Second, it relied on trial-by-trial feedback, whereas in the real world, generative models often have to be learned without such supervisory signals.

At a neural level, training a neural network to become Bayesian is most simply done with trial-by-trial feedback but that is certainly unrealistic. Many neuroscientists would like to see learning implemented through biologically plausible learning rules, specifically rules that only make use of local connections between pairs of neurons. This is an active area of research.

Box 15.1 — Artificial Neural Networks as an alternative philosophy to Bayesian modeling. A challenge to Bayesian modeling is that the world is sufficiently complex that we may not be able to specify a satisfactory generative model for a particular task. The field of Artificial Neural Networks follows an entirely different approach. Instead of having a human defined generative model of the world where we do inference about the parameters, deep learning uses a general purpose function (see 6.4) which can approximate inference on a large class of potential worlds. It then optimizes the parameters of that model. The focus is thus not on inference regarding the right parameters and their posterior distribution but on finding a model that works.

Most application of learning boil down to the question of estimating a quantity y based on some observed quantities x. In Bayesian systems we typically estimate $p(y|x)$. If our model is sufficiently simple then we can solve these equations. However, $p(y|x)$ may be an arbitrarily complex function whose form we do not know. Instead of choosing a function we know, we can use a large number of parameters θ and a class of functions with many free parameters $p(y|x) = f_\theta(x)$ and optimize the parameters to obtain the best possible estimates. In practice, choosing a hierarchical, continuous, differentiable function, an artificial neural network, often produces particularly good results. ■

15.7 Bayesian transfer

Optimal behavior for a task at hand will depend on likelihoods, priors, and cost functions, simply because these determine which behaviors will be successful. But are these constructs meaningful and actually used by the individual? A strategy for demonstrating that the internal constructs of a Bayesian model are meaningful is what Maloney and Mamassian call *Bayesian transfer* [81]. The idea is that if prior, likelihood, and cost function are meaningful, then they should be flexible and generalize across tasks or conditions. For example, if a participant is near-optimal with the combination of likelihood 1 and prior 1, as well as with the combination of likelihood 2 and prior 2, then the generalization test would consist of presenting a condition that calls for the use of likelihood 1 and prior 2, or likelihood 2 and prior 1. In such a new condition, relearning a rule through trial and error would be slow and require many trials, whereas recombining “computational modules" should be fast. Of course, it is possible that some constructs of Bayesian computations are flexible and generalizable with respect to variations of one component but not to others. Ma and Jazayeri proposed a hierarchy of Bayesian flexibility [76]. Bayesian flexibility is the main way of obtaining evidence about the degree to which priors and likelihoods can be combined in arbitrary ways.

Taken to the extreme, the adjustment in a Bayesian transfer experiment should be instantaneous: varying the prior and likelihood on a trial-by-trial basis should produce near-optimal behavior. In other words, we could use every trial as a generalization test from other trials. In practice, few studies vary both prior and likelihood at the same time, but varying the likelihood from trial to trial is very common. For instance, in cue combination studies, the reliability of at least one sensory cue may be varied from trial to trial [1]. To strengthen the test, trial-by-trial feedback should be withheld, so as to make trial-and-error rule learning nearly impossible. There could be initial training only at the start of the session or of the entire experiment, but perhaps not even for all likelihood conditions – forcing the participants to generalize from the limited training conditions combined with pre-experiment experience with similar stimuli. It is less common to test the generalizations of priors or cost functions than of likelihoods, but a number of relevant studies have been conducted. For example, the light-from-above prior and a prior over faces seem to be task-general. Acerbi 2014 [15] varied the prior from trial to trial, and Whiteley Sahani [117] varied the cost function from trial to trial. In all these situations, it seems that people are capable of Bayesian transfer in those situations as well. Altogether, we think that as-if criticisms have been adequately addressed empirically, and we do not consider them major challenges of Bayesian models of behavior.

Bayesian transfer studies, when successful, often emphasize near-optimality of human behavior. However, the stronger message is that they provide evidence that the likelihood (or the prior) is a meaningful construct to the brain, a result that argues against the as-if criticism.

15.8 Real-world complexity

A main challenge to Bayesian models of behavior consists of scaling up to complex and/or naturalistic problems. The vast majority of tasks studied in this book and in the field are simple, involving one or a few variables, each of which is binary or one-dimensional. This matches well with most laboratory tasks but poorly with naturalistic tasks, in which there are more variables, those variables are higher-dimensional, and they can take on more values. We saw a glimpse of this in Section 12.2.2, where we calculated a likelihood function over a change point vector that could take 2^T values (also see Problem 11.8). Even for modest T, this yields a very large number of hypotheses to keep in mind. Another example is top-level nuisance parameters as discussed in Chapter 9. In a real scene, there are many such parameters for every object: viewing angle,

[1] In conditions with sensory noise, the likelihood will vary from trial to trial even when the stimulus is held fixed; however, experimenters cannot control this variation so they have to resort to changing the stimulus.

viewing distance, lighting, etc. Moreover, these nuisance parameters themselves have strong spatial structure, with objects also playing a special role. Properly marginalizing over all these parameters, analogous to our simple examples in Chapter 9, is a daunting task.

In addition, real-world behavioral output is nothing like the response options in laboratory tasks. For example, when hiking in the forest, a behavior to predict is where one is going to move one's foot next, at which angle, at what time, and with what force. Another example of real-world behavior is to conduct a conversation. There, the behavior to be predicted is not a simple binary choice or one-dimensional estimate, but it is natural language. Adding to this the complexity of the inference – for example, of the intentions of one's conversation partner – and the complexity of the cost function – which likely involves multiple terms such as the accuracy of one's understanding, the time a conversation takes, and the impression one makes on the other – it is not surprising that Bayesian modelers have not tried to tackle this problem.

It is easy to imagine that if the environment or the task becomes too complex, humans may cease to follow the predictions of Bayesian models. However, it is exactly under such circumstances that it is not even clear how to formulate a Bayesian model – for example because we cannot reasonably know the distributions involved in the generative model. We could call this the *Bayesian tragedy*: the phenomenon that we can specify Bayesian models in regimes where they are likely to work, and we have a hard time specifying them in regimes where they may be unlikely to work.

15.9 Approximate inference

Not all hope is lost when it comes to scaling up the complexity of Bayesian computations. Whereas in this book, we have mostly studied exact Bayesian inference, there is a whole world out there of approximate Bayesian computation [28]. Some approximate algorithms have made inroads in cognitive science and neuroscience.

Some algorithms allow calculating exact posteriors when given enough time, regardless of the number of variables involved. Some message passing algorithms, such as the forward-backward algorithm can solve a large set of probabilistic problems, e.g. when variables all have Gaussian distributions. Alternatively, there are sampling algorithms that, under certain assumptions, converge to the correct solutions over time. Other algorithms are designed to approximate the correct results in a certain sense. While these approaches are beyond the scope of this book, we briefly describe them here.

Numerical Integration. When the number of relevant variables is reasonably small, it is often feasible to numerically solve the relevant integrals for marginalization. For this we may use standard packages for integration. This approach is direct but quickly becomes slow as the number of variables to be integrated increases.

Belief propagation. In this class of algorithm we start with some variables, propagate messages through the system, and apply potential corrections. We often assume that the relevant variables have distributions for which the integrals can be analytically solved or that approximations can be had. We also assume that the relations between variables are such that they can be described as a (generally) sparse graph. In certain kinds of graphs, belief propagation provides an efficient way of solving many Bayesian problems. Sometimes it is approximately correct even when the assumptions are not satisfied.

Variational Bayesian approaches. For complex probability distributions, belief propagation can not work in a fully analytical fashion. In fact, the probability distribution of each unobserved variable given the observed variables may be arbitrarily complicated. Variational Bayes uses the following trick. We approximate the probability distribution of p by a probability distribution (q_θ) that depends on a number of parameters θ. Subsequently we optimize the parameters of q_θ so that q_θ becomes as similar as possible to the real probability distribution p. Making them the same is not generally possible, so variational methods usually optimize an approximation, the so-called Evidence Lower Bound (ELBo).

Markov Chain approaches. A different approach towards the solution of similar problems are Markov Chain methods. These initialize a vector of all variables to be inferred, and change it iteratively to produce a sequence of vectors. This is set up in such a way that, over time, the expected number of times of being in a state is proportional to the actual probability of the state given the observed variables. These methods can deal with arbitrary probability distributions. They are particularly useful if the probability distribution is concentrated in small areas of high probability. However, all the proofs for correctness only hold in the limit of large, often astronomical, numbers of iterations. In practice, Markov Chain methods are usually limited by the difficulty of problems they can solve by available runtime. In cognitive science, it has been proposed that human decision-makers behave like Markov Chains with a limited number of iterations [111].

15.10 Summary

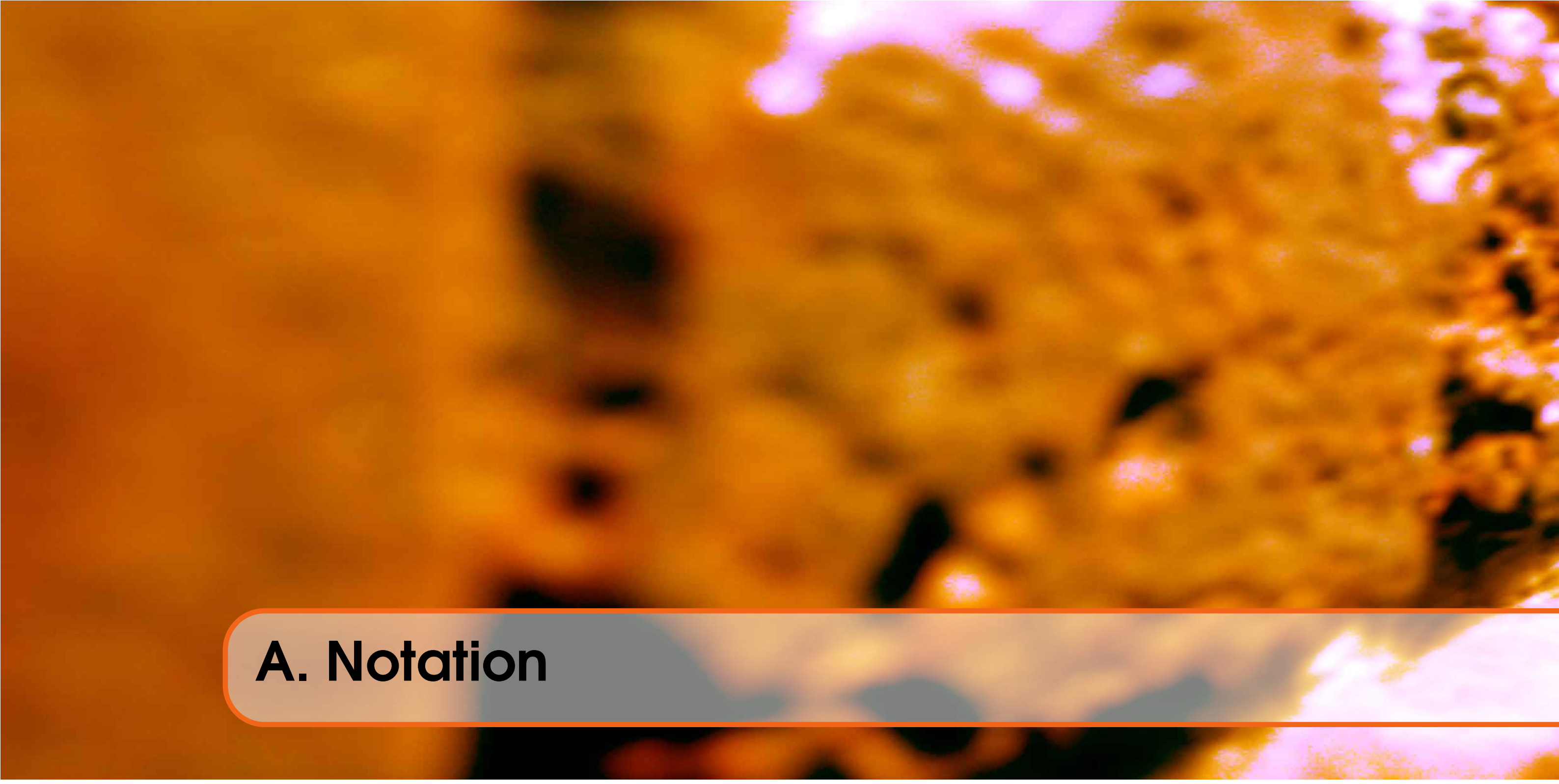
Bayesian statistics is a way of modeling behavior that starts with a specification of the nature and parameters of an underlying model. We learned:

- Bayesian models are more meaningful (in a normative sense) if they derive from an understanding of the world around us.
- Bayesian models can in many situation be relatively uniquely defined.
- Bayesian models often have high predictive value.
- Neural implementations of Bayesian models of behavior can relatively easily be found.
- The internal constructs of Bayesian models – likelihoods, priors, and cost functions – can be given validity through Bayesian transfer studies.
- Machine learning provides the field with tools for approximate Bayesian inference; however, their relevance to human reasoning remains unclear.

15.11 Suggested readings

- Wendy J Adams. “A common light-prior for visual search, shape, and reflectance judgments”. In: *Journal of Vision* 7.11 (2007), pages 11–11
- John R Anderson. “The adaptive nature of human categorization.” In: *Psychological review* 98.3 (1991), page 409
- John R Anderson. *The adaptive character of thought*. Psychology Press, 2013
- Jeffrey S Bowers and Colin J Davis. “Bayesian just-so stories in psychology and neuroscience.” In: *Psychological bulletin* 138.3 (2012), page 389
- Nick Chater, Mike Oaksford, et al. *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford University Press, USA, 2008
- Nick Chater et al. “The imaginary fundamentalists: The unshocking truth about Bayesian cognitive science”. In: *Behavioral and Brain Sciences* 34.4 (2011), page 194
- Stephanie Y Chen, Brian H Ross, and Gregory L Murphy. “Implicit and explicit processes in category-based induction: Is induction best when we don’t think?” In: *Journal of Experimental Psychology: General* 143.1 (2014), page 227
- Gerd Gigerenzer. “On narrow norms and vague heuristics: A reply to Kahneman and Tversky.” In: (1996)
- Matt Jones and Bradley C Love. “Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition”. In: *Behavioral and brain sciences* 34.4 (2011), page 169
- Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011

- Charles Kemp, Andrew Perfors, and Joshua B Tenenbaum. “Learning overhypotheses with hierarchical Bayesian models”. In: *Developmental science* 10.3 (2007), pages 307–321
- Falk Lieder and Thomas L Griffiths. “Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources”. In: *Behavioral and Brain Sciences* 43 (2020)
- Wei Ji Ma. “Organizing probabilistic models of perception”. In: *Trends in cognitive sciences* 16.10 (2012), pages 511–518
- Wei Ji Ma and Mehrdad Jazayeri. “Neural coding of uncertainty and probability”. In: *Annual review of neuroscience* 37 (2014), pages 205–220
- Laurence T Maloney and Pascal Mamassian. “Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer”. In: *Visual neuroscience* 26.1 (2009), pages 147–155
- Barbara Mellers, Ralph Hertwig, and Daniel Kahneman. “Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration”. In: *Psychological Science* 12.4 (2001), pages 269–275
- Daniel J Navarro and Amy F Perfors. “Similarity, feature discovery, and the size principle”. In: *Acta Psychologica* 133.3 (2010), pages 256–268
- Dobromir Rahnev and Rachel N Denison. “Suboptimality in perceptual decision making”. In: *Behavioral and Brain Sciences* 41 (2018)
- Julia Trommershäuser, Laurence T Maloney, and Michael S Landy. “Decision making, movement planning and statistical decision theory”. In: *Trends in cognitive sciences* 12.8 (2008), pages 291–297
- Louise Whiteley and Maneesh Sahani. “Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes”. In: *Journal of vision* 8.3 (2008), pages 2–2
- Fei Xu and Joshua B Tenenbaum. “Word learning as Bayesian inference.” In: *Psychological review* 114.2 (2007), page 245



A. Notation

Notation for probabilities is confusing in more than one way. One source of confusion is the distinction between the probability of an event or assertion, e.g. $P(\text{it will rain today})$, and the probability density of a continuous variable x e.g. a position, which we denote by a lowercase $p(x)$. Within the latter category, it is confusing that the same letter $p(x)$ is used to denote different functions. If we had been dealing with ordinary functions, we would typically give different functions different names, e.g. $f(x)$, $g(x)$, etc. But for probability distributions, although they are (special) functions, that is not common. Instead, it is common to use a subscript on the $p(x)$ to denote the random variable that the probability distribution belongs to, i.e. $p_X(x)$. In mathematics text books, this is standard notation.

This, however, is not a mathematics textbook. We find the subscripts cumbersome and sometimes redundant. Therefore, we try to avoid them where possible, without compromising accuracy. How do we do that? First, for discrete random variables taking on specific values, we prefer the equivalent “event notation" with an uppercase P. For example, the expression $P(X = 3)$ is easier to read than $p_X(3)$, even though they mean the same thing. We can always use event notation for a discrete random variable taking on a specific value. Second, if we describe the probability distribution of a random variable X with a generic argument x (instead of a specific value), we write $p(x)$ instead of $p_X(x)$; thus, the letter used for the argument implies the name of random variable.

Avoiding subscripts is not always possible. Let’s say we have a continuous random variable X evaluated at a specific value, say 3. Then, we cannot replace $p_X(3)$ by $P(X = 3)$, because the latter is equal to 0 not just for the value 3 but for any other specific value. For example, the probability that you hit the exact center point of a dart board with an (infinitely) sharp dart is 0. By contrast, $p_X(3)$ is a probability density function, which is well-defined.

These considerations give rise to the following conventions:

1. “Event notation": If X is a statement or event, then we use $P(X)$ to denote the probability of X.
2. We use lowercase p to denote probability mass functions over discrete variables and probability density functions over continuous variables.
3. If X is a discrete random variable, then the most accurate notation for its probability mass

function would be $p_X(x)$, which is the same as $P(X = x)$, the probability that the random variable X takes the value x. If x is a generic rather than a specific value, the notation $p_X(x)$ feels redundant and we simply write $p(x)$. If the value of X is specific, say 3, then we write $P(X = 3)$.

4. If X is a continuous random variable and its value is generic, say x, then we will again write $p(x)$ instead of $p_X(x)$. If its value is specific, say 3, then we cannot write $P(X = 3)$, but we have to write $p_X(3)$.
5. How we denote joint and conditional probabilities of discrete random variables X and Y depends again on whether its values are generic or specific. If their values are both generic, say x and y, we write $p(x,y)$ instead of $p_{X,Y}(x,y)$ for their joint probability mass function. If only one variable has a specific value, we write $p(X = 3, y)$ or $p(x, Y = 2)$. If the values are both specific, we write $P(X = 3, Y = 2)$. We denote the corresponding conditional probabilities by $p(x|y)$, $P(X = 3|y)$, $p(x|Y = 2)$, and $P(X = 3|Y = 2)$; note that we can only use uppercase P if what appears before the conditional sign is an event rather than the generic value of a variable. In all, if all variables involved are discrete, we do not need subscripts.
6. We next consider the case that X and Y are both continuous random variables. If their values are both generic, say x and y, we write $p(x,y)$ instead of $p_{X,Y}(x,y)$ for their joint probability density function. If at least one value is specific, we do use the subscripts, writing, for example $p_{X,Y}(x,2)$, $p_{X,Y}(3,y)$, or $p_{X,Y}(3,2)$ for joint probabilities; in the first two cases, we typically do not even bother to use a different symbol for the random variable and its generic value. We denote conditional probabilities by $p(x|y)$ if both values are generic, by $p(x|Y = 2)$ if the variable after the conditional sign has a specific value, by $p_{X|Y}(3,y)$ if the variable before the conditional sign has a specific value, and by the slightly odd but entirely correct $p_{X|Y}(3|2)$ if both values are specific. In other words, for a continuous variable taking on a specific value, we can use “event notation" only if the variable appears after the conditional sign.
7. Mixed distributions of discrete and continuous random variables are trickiest notation-wise. Say that X is discrete and Y continuous. If their values are both generic, we can still write $p(x,y)$, which is interpreted as a probability mass function for X and a probability density function for Y. If at least one value is specific, we write, for example, $p_{X,Y}(3,y)$, $p_{X,Y}(x,2)$, or $p_{X,Y}(3,2)$ for joint probabilities; this notation is the same as when X and Y are both continuous. We denote conditional probabilities by $p(x|y)$ or $p(y|x)$ if both values are generic, by $p(y|X = 3)$ or $p(x|Y = 2)$ if the variable after the conditional sign takes a specific value (regardless of whether that variable is the discrete or the continuous one), by $P(X = 3|y)$ if only the discrete variable has a specific value and appears before the conditional sign, by $p_{Y|X}(2|x)$ if only the continuous variable has a specific value and appears before the conditional sign, by $P(X = 3|y = 2)$ if both values are specific and the discrete variable appears before the conditional sign, and finally, by $p_{Y|X}(2|3)$ if both values are specific and the continuous variable appears before the conditional sign.

Now let’s talk about expected values:

1. If $f(X)$ is a function of only one random variable, X, then $\mathbb{E}[f(X)]$ is automatically the expected value with respect to X, i.e. $\sum_{i=1}^{n} f(x_i)p(x_i)$ (discrete) or $\int f(x)p(x)dx$ (continuous).
2. If $f(X,Y,...)$ is a function of two or more random variables, X, Y, etc., then we use subscripts to indicate which variable(s) the expected value is taken over. For example, in the continuous

case, the following expected values are different:

$$\mathbb{E}_X[f(X,Y)] \equiv \int f(x,y)p(x)dx \tag{A.1}$$

$$\mathbb{E}_{X,Y}[f(X,Y)] \equiv \int f(x,y)p(x,y)dx \tag{A.2}$$

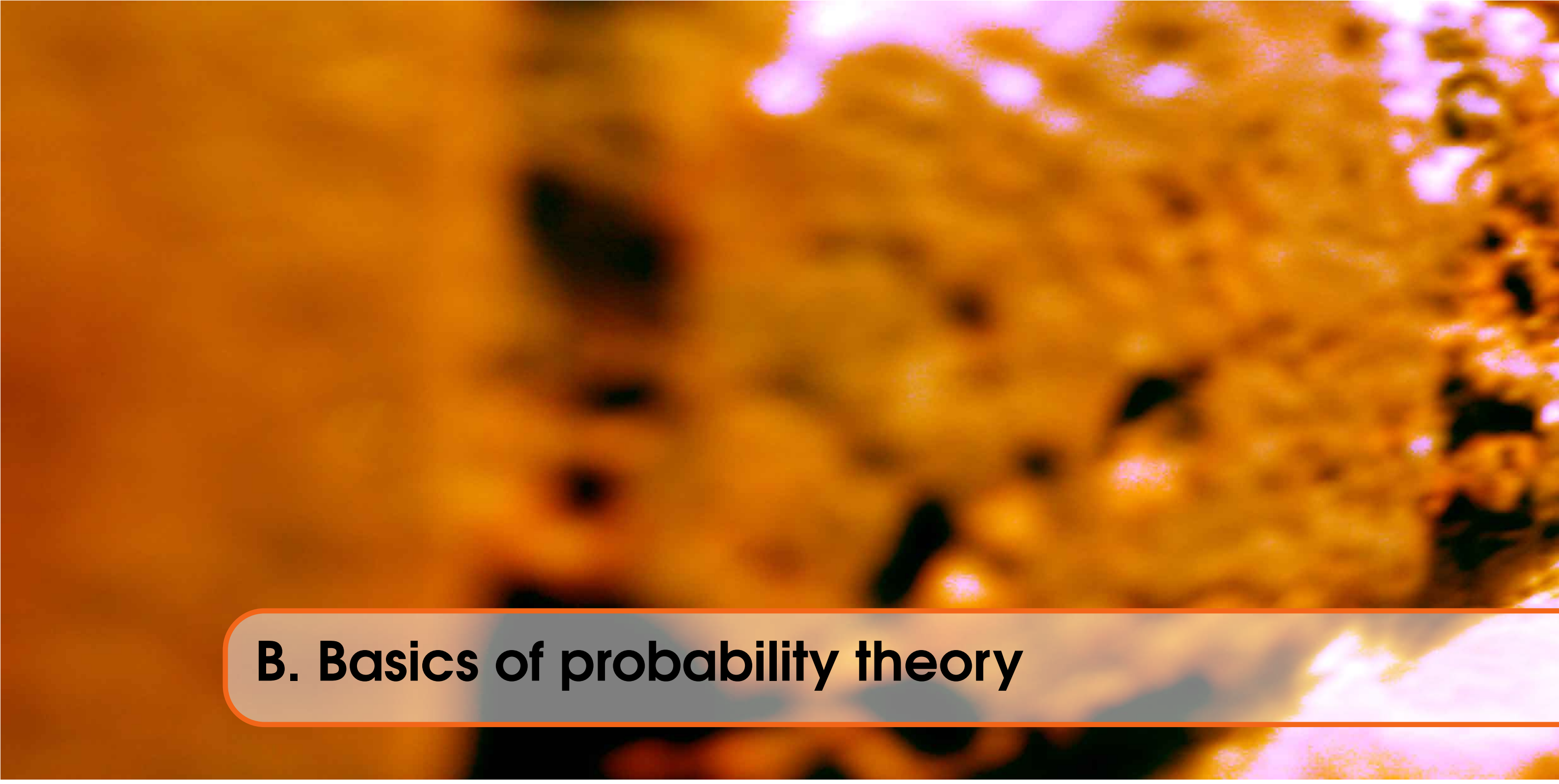
$$\tag{A.3}$$

3. When using a conditional distribution to calculate the expectation, we write the random variable being conditioned on in the subscript, and the value of that variable inside the square brackets:

$$\mathbb{E}_{X|Y}[f(X)|y] \equiv \int f(x)p(x|y)dx \tag{A.4}$$

$$\mathbb{E}_{X|Y}[f(X,Y)|y] \equiv \int f(x,y)p(x|y)dx \tag{A.5}$$

   (in the continuous case). Here, y could be a generic or a specific value. If y is a generic value, we abbreviate $\mathbb{E}_{X|Y}[f(X)|y]$ to $\mathbb{E}[f(X)|y]$ to avoid redundancy, but we keep the subscript in any more complicated case.
4. Conventions for variance and standard deviation are as for expected values. We will denote them by Var$[X]$ and Std$[X]$, respectively.



B. Basics of probability theory

We believe that calculus helps you understand math, but probability helps you understand life. Nevertheless, probability theory is not a standard component of most undergraduate science curricula. We hope that this will change, but in the meantime, this Appendix provides some basics of probability theory. It is by no means an exhaustive introduction as one would find in a textbook on probability theory. Instead, it is a tutorial that focuses only on the concepts and calculations used in the book.

B.1 Objective and subjective probability

Probability is degree of possibility. In its most restrictive sense, probability can be defined as the expected outcome frequency of a repeatable event, such as the probability that a coin will come up heads or the probability that someone rolls a 5 on a die. These events can be repeated an arbitrarily large number of times, and the long-run outcome frequencies tallied. If the proportion of tosses on which a coin lands heads converges to 0.5 as the number of tosses approaches infinity, we can state that the coin has a 0.5 probability of landing heads. This type of probability is sometimes called *objective probability*, and it is the only valid type of probability according to a strict frequentist view of probability.

A much broader – and, we believe, much more useful – conceptualization of probability is as degree of belief in a possibility. This is sometimes called *subjective probability*. The everyday terms *confidence* and *uncertainty* refer to subjective probabilities. If I know that a die has 1/6 chance of landing 5, then my confidence in the proposition that it will land 5 is 1/6. This particular example is trivial, because it involves simply converting an objective probability (a long-run outcome frequency) into a belief statement. However, the vastly wider applicability of the subjective conceptualization becomes clear when we consider degrees of belief in outcomes that cannot be repeated, for example the probability that candidate A will beat candidate B in the next election. This is not a probability that can be obtained by repeating the same event many times, but we may nevertheless have a strong prediction regarding the outcome. Indeed, examples of subjective probabilities that cannot be phrased as long-run outcome frequencies abound in daily life: What is the probability that it will rain today? What is the probability that I will enjoy the course taught by professor X? Many scientific questions also can be phrased only in terms of subjective and not in terms of objective probabilities: What is the probability that Saturn's mass lies between 10^{25} and 10^{26} kg? What is the probability that disease X is caused by a virus?

The distinction between objective and subjective probability is not always clear. For example, to determine the probability that it will rain today, a forecaster might run a large number of simulations starting from the current state of the atmosphere, each with a different instantiation of the stochastic factors in the model, and record the frequency of rain among these runs. While the resulting probability is subjective, it has been obtained in an "objective" way, namely by counting. Similarly, if I observe dark clouds in the sky and express my opinion that there is a high probability of rain, I am expressing a subjective probability judgment, but I am basing this judgment on a large number of previously observed, similar (though not identical) skies.

A great advantage of Bayesian inference is that it treats both subjective and objective probabilities in the same way. Bayesian inference is therefore extremely widely applicable. The same mathematical relationships (Bayes' rule, marginalization, etc.) apply identically to both types of probability. Bayesian models of perception, however, are grounded fundamentally in subjective probability: there is only one true world state, but from the point of view of an organism trying to infer it, there are many possibilities, and degrees of belief can be assigned to these possibilities.

B.2 The intuitive notion of probability

We call the set of all possibilities under consideration the *sample space*. An *event* or *hypothesis* is a subset of the sample space. The term "event" is commonly used when discussing objective probability, and the term "hypothesis" when discussing subjective probability. The sample space could be "all possible numbers I can roll on a die" or "all possible weather patterns that can occur today". Given the former sample space, an event could be "I will roll an even number". Given the second sample space, a hypothesis could be "It will rain today". The probability of an event or hypothesis is a real number between 0 and 1, indicating the degree of possibility of the event or hypothesis. An event that is certain has a probability of 1, and an impossible event has a probability of 0. For events, one can think of probability as the frequency that the event happens among a very large number of random samples from the sample space. For instance, the probability that I will roll an even number on a 6-sided die is 0.5. As explained above, this can also be conceptualized as a degree of belief. For a hypothesis, the frequency concept does not generally apply but probability still represents a degree of belief; for example, the degree of belief that it will rain today could be 0.35. The probability of an event or hypothesis X is denoted $\Pr(X)$. For example, $\Pr(\text{it will rain today}) = 0.35$, or $\Pr(\text{coin comes up heads}) = 0.50$.

B.3 Complementary event

Given an event or hypothesis, its complementary event or hypothesis is that the first event or hypothesis does not occur or is false. For example, the complementary event to "rolling a 1 on a die" is "rolling a 2, 3, 4, 5, or 6 on a die". If the event or hypothesis is denoted X, then its complementary event or hypothesis, or complement, is denoted $\neg X$ (read: "not X"). Here, $\neg$ is the symbol for a logical negation. The probability of the complementary event or hypothesis is 1 minus the probability of the event or hypothesis:

$$\Pr(\neg X) = 1 - \Pr(X). \tag{B.1}$$

In some situations, it is easier to calculate the probability of the complement of an event or hypothesis than of an event or hypothesis itself. For example, if you are asked to calculate the probability that the sum of the eyes on two dice is at least 3, it is easiest to first calculate the probability that the sum is lower than 3, and subtract that from 1 (the answer is 35/36).

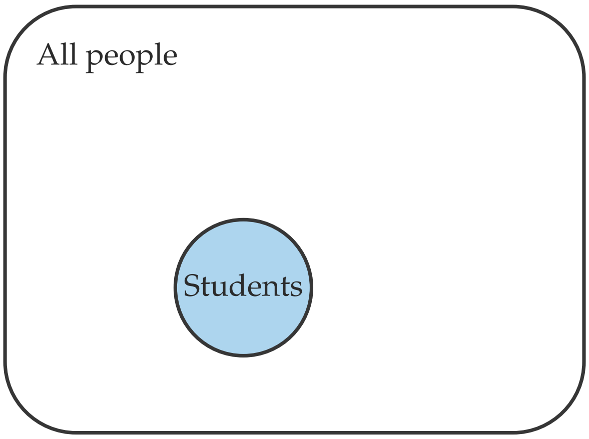


Figure B.1: Example of Venn diagram used to represent probabilities.

B.4 Venn diagram representation

Events and hypotheses can be represented graphically through *Venn diagrams*[1]. First draw a large rectangle whose interior represents all possible outcomes, i.e. the sample space. Assign the area of this rectangle a value of 1, representing a total probability of 1. Then draw inside this rectangle a circle that represents all outcomes consistent with a particular event or hypothesis.

For example, the rectangle could represent all people in a group, and the circle all students among them. The area enclosed by the circle is a fraction of the area enclosed by the rectangle; this fraction represents the probability of the event or hypothesis – in our example, the probability that a randomly selected person is a student. The complement of the event or hypothesis is represented by the points that are inside the rectangle but outside the circle. Its area divided by the area of the rectangle represents the probability that a randomly selected person in the group is not a student. We will make use of the Venn diagram representation in later sections.

B.5 Random variables and their distributions

A random variable is a variable whose values cannot be known with certainty. Examples include the number rolled on a die, the date of birth of a person, the shoe size of a random person on your street, the time it takes to travel from home to work, the number of voters who will participate in an upcoming election, or the price of a stock tomorrow. The opposite of a random variable is a variable whose value is known with certainty. Examples of non-random variables are the number of planets between us and the Sun (2), the number of days in a week (7), the ratio of the circumference to the diameter of a circle (π), and the distance between two adjacent cm marks on a ruler (1 cm).

This is not an airtight distinction. Variables that appear non-random might be subject to measurement or production noise, which makes them random. For example, the distance between two adjacent cm marks on a ruler might vary, since the machine that produced the rulers was probably programmed by a computer that set the centimeter marks. However, computer-generated numbers have only a finite number of decimals, perhaps 10. As a consequence, the centimeter marks will never reach femtometer precision. In addition, the paint used for the marks will not attach itself to the surface in an identical way every time a ruler is produced. Therefore, one can

[1] Strictly speaking, the areal diagrams that we show in this book are *Euler diagrams*, as they represent only relationships that actually occur; Venn diagrams, strictly defined, represent all possible logical relationships between sets. However, we follow the prevalent convention of naming any areal probability diagram a Venn diagram.

think of the distance between two adjacent marks as a random variable. Therefore, certainty about this variable does not exist. For reasons such as these, it might be useful to think of all variables as random, just with some having very low uncertainty.

Randomness, also called variability, noise, or stochasticity, is often a consequence of a lack of knowledge. When I roll a die, if you could somehow record exactly the position, direction, and speed with which the die left my hand, and you were able to simulate exactly the interactions the die had with air and table, then you would be able to predict with certainty the outcome of the roll. Since nobody knows the values of all these variables, the outcome of the die roll is considered to be random. Whether true randomness exists is a philosophical question that is beyond the scope of this book.

B.5.1 Discrete versus continuous random variables

Random variables can be distinguished based on the values they can take. The most important distinction is between discrete and continuous random variables.

A *discrete random variable* takes on a set of values that can be counted, even though there might be infinitely many. Examples are the number of children in a household, the number of dots we draw on a piece of paper, the number of action potentials fired by a neuron, in fact any "number of..." variable: the number of moves in a chess game, the age of a person when counted in whole years, the price of a movie ticket, the number of ingredients that go into a recipe, or the identity of a spoken word. A discrete random variable that takes on only two possible values is called *binary*.

A *continuous random variable* takes on values on a continuum. Examples are the length of a line segment, the direction one can walk in an open field, the amount of an ingredient in a recipe, the waiting time in front of a red light, the speed of a car, and the frequency of a musical note. One can think of a continuous variable as discrete but with values that come in infinitesimally small increments. For example, distance is a continuous variable, but when it is measured in whole millimeters, it is a discrete variable. In a computer program, truly continuous variables do not exist; their domain must always be discretized.

B.5.2 Total probability = 1

The total probability of all possible values of a random variable equals 1. This total is a reflection of the fact that the possibilities are mutually exclusive. If one were to increase the probability of one value, the probability of at least one other value has to decrease.

B.5.3 Discrete probability distributions

Discrete probability distributions are functions that assign a probability to each possible value of a discrete random variable. The probability distribution over a discrete random variable is also called a *probability mass function*. A discrete random variable X taking a particular value x is an event, denoted $\Pr(X=x)$. As we now vary x over all its possible values, we obtain a function of x. This is the probability mass function, denoted $p_X(x)$.

$$p_X(x) = \Pr(X=x). \tag{B.2}$$

(Throughout this Appendix, we will denote random variables by capitals and their values by lowercase letters. We do not use this convention throughout the book, though.) This means that the probability mass function evaluated at x is equal to the probability that the random variable takes this value. We use the subscript X (uppercase) to refer to a random variable, and the argument x (lowercase) to refer to a specific value of this random variable. The term "mass" is borrowed from physics. Roughly speaking, it is based on using matter as a metaphor for a possibility (a point in the sample space). The larger the probability of an event or hypothesis, the larger the mass of the piece of matter in the metaphor.

For example, the random variable X is the number rolled on a die. Its possible values x are 1 to 6. If the die is fair, the probability of each of these values is 1/6, i.e. $\Pr(X=x)=\frac{1}{6}$ for all x. This is an example of a discrete uniform distribution. If the die is not fair, the probability of at least two of the values differs from 1/6, and the distribution will no longer be uniform.

For discrete random variables, total probability is computed by summing over all possible values; it should return 1. This is denoted as follows:

$$\sum_x p_X(x) = 1. \tag{B.3}$$

In a specific case where the possible values of X are given, we can put those above and below the $\sum$ sign. For example, the total probability of a die roll would be

$$\sum_{x=1}^{6} p_X(x) = \frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}. \tag{B.4}$$

Binary random variables are a special case of discrete random variables. Suppose a binary random variable X can take values x_1 and x_2. We know that total probability equals 1. Therefore, the probability of x_2 is 1 minus the probability of x_1, i.e., $p_X(x_2)=1-p_X(x_1)$.

B.5.4 Continuous probability distributions

What is the probability that someone is exactly 160 cm tall? It is zero, since "exactly" means that the length is accurate to an infinite number of decimal places. This problem is characteristic of continuous random variables, and illustrates that probability mass functions, which worked well for discrete distributions, have to be replaced by a different concept in order to accommodate continuous variables.

Suppose we are interested in the probability distribution of the height of an adult (**Fig. B.2**). As a first approximation, we could consider possible heights in bins of 10 cm increments: between 120 and 130 cm, between 130 and 140 cm, etc. Each bin has an associated probability, and in this way we can build up a probability mass function. However, we might want to describe height more finely, say in bins of 5 cm: between 122.5 and 127.5 cm, etc. Each original bin is thus replaced by 2 new bins, each of which has on average one half the probability mass of the original bin. Thus, the new probability mass function is scaled to about half the height of the original one (**Fig. B.2**). As we keep decreasing bin width in order to increase precision, the probability mass per bin keeps decreasing as well – it can become arbitrarily small. This is not very satisfactory. Is there a way to prevent the probability mass function from "disappearing"? Yes, this can be accomplished by dividing the probability mass in a bin by the width of the bin. By doing so, the function does not change much as we decrease bin width – it only becomes more precise. The *probability density function* is the result of this process as the bin width approaches zero (red curve). Again, there is an analogy with physics: if the probability in a bin is regarded as mass, then dividing this probability by bin width is analogous to computing a linear density: mass per unit length of the x-axis.

The similarity in notation between the probability mass function for discrete variables and the probability density function for continuous variables, both denoted $p_X(x)$, is misleading as there are some important conceptual differences between the two. For a discrete distribution, the probability mass of a single point never exceeds 1, since the probability mass values have to sum to 1. For a continuous distribution, the probability density at a single point is meaningless and can take arbitrarily large values. Consider, for example, a uniform distribution on the interval $[0,0.01]$. It will have a probability density of 100 at every point. Only the integral over an interval will always be less than or equal to 1. Stated differently, for a discrete distribution, the probability $\Pr(X=x)$ is a meaningful number that can take any value between 0 and 1, and is in fact identical to $p(x)$. For a

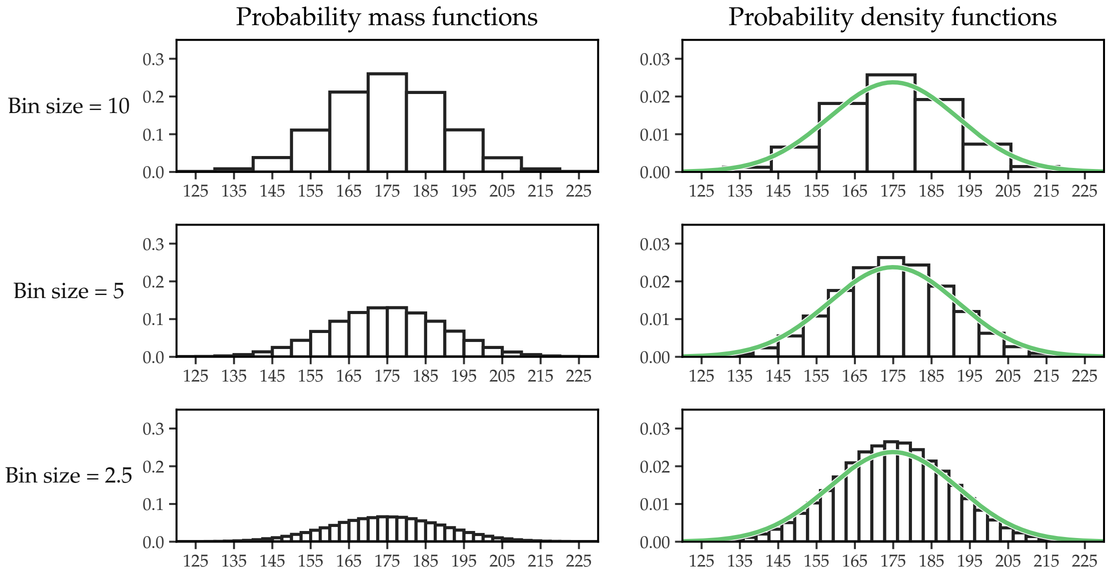


Figure B.2: When a random variable can take on a continuum of values (x-axis), a probability mass function is only defined when outcomes are binned. The values of the function will decrease with decreasing bin size (left column). Probability density functions are obtained by dividing the probability mass values by bin size. This yields values that are independent of the bin size. The process of making the bin size smaller can be continued until the bins are infinitesimally small. This produces a continuous probability density function, overlaid in green.

continuous distribution, $\Pr(X = x)$ is always 0, and only probabilities of the form $\Pr(a \leq X \leq b)$, with a and b arbitrary numbers, are meaningful.

We will use the terms *probability distribution function (pdf)*, *probability distribution*, or simply *distribution* to refer to the probability mass function of a discrete random variable or the probability density function of a continuous random variable.

Just as for discrete variables, the total probability of all values of a continuous variable equals 1. Total probability for a continuous variable is computed not as a sum, but as an *integral*. The integral of a continuous probability density function, as defined above, is the width of a bin multiplied by the function value in that bin, summed over all bins, in the limit that bin width approaches zero. Calculus provides recipes to compute integrals of certain functions. In this chapter, we familiarize ourselves with various integrals over probability density functions, especially because they directly parallel expressions with sums over probability mass functions; however, we will not evaluate these integrals, so no calculus is needed. The rule of total probability for a continuous variable X is written as

$$\int p_X(x)dx = 1 \tag{B.5}$$

The "dx" is in essence the width of a very small bin, and the integral sign $\int$ is a deformed "S" for sum.

The most important continuous distribution is the normal distribution, which we discuss in detail below. Another important one is the uniform distribution. The uniform distribution on an interval $[a, b]$ has a constant pdf

$$p(x) = \frac{1}{b-a}. \tag{B.6}$$

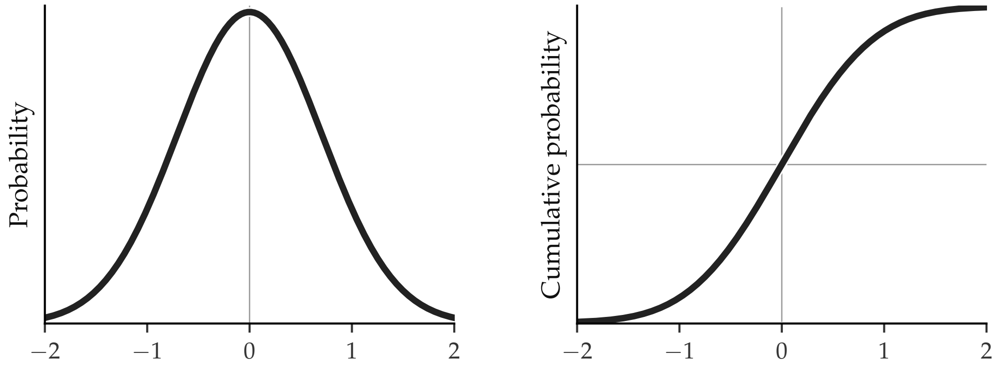


Figure B.3: A probability density function and a cumulative distribution function.

The following continuous distributions are also common in applications of probability theory. The exponential distribution is given by $p(x) = \lambda e^{-\lambda x}$, with λ a constant and x defined on the positive real line. The power law distribution is given by $p(x) \propto x^{-a}$, with a a constant and x again defined on the positive real line. We will discuss the normalizations of these distributions shortly.

B.5.5 Formal definition of the probability density function

Consider a continuous random variable X, such as the waiting time in a queue. The probability that the value of X is less than or equal to x is denoted by $\Pr(X \leq x)$. This is the *cumulative distribution function* (cdf) of X at x, denoted $P_X(x)$:

$$P_X(x) = \Pr(X \leq x). \tag{B.7}$$

By definition, this is a monotonically increasing function that takes values between 0 (as x approaches ∞) and 1 (as x approaches ∞). The *probability density function* (pdf) of X is now the derivative of this function:

$$p_X(x) = \frac{dP_X}{dx}. \tag{B.8}$$

We use uppercase letters to denote cumulative distribution functions, and lowercase letters to denote probability density and mass functions.

Fig. B.3 shows an example of a cumulative distribution function and a probability density function. For discrete random variables, the cdf can be defined in the same way, but it is not a necessary step in defining the probability mass function.

To go back from the pdf to the cdf, one integrates:

$$P_X(x) = \Pr(X \leq x) = \int_{-\infty}^{x} p_X(y)dy. \tag{B.9}$$

The physics equivalent of this statement is that the integral over a density is a mass. It immediately follows that the probability that X takes values in an interval $(x_1, x_2]$ can be obtained by integrating $p_X(x)$ between x_1 and x_2:

$$\Pr(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p_X(y)dy. \tag{B.10}$$

It also follows from the definition that $\int_{-\infty}^{\infty} p_X(y)dy = 1$.

B.5.6 Normalization

A function can be made into a probability distribution by dividing each value by the total value on the entire domain, provided that this total value is finite. As a result, the probability distribution will integrate (or sum) to 1. This process is called *normalization*. If the total value on the entire domain is infinite, normalization is not possible.

Exercise B.1 Prove that the exponential distribution is normalized. ■

Exercise B.2 Normalize the power law distribution and find a condition on a for which normalization is possible. ■

B.6 Mean, variance, and expected value

For a discrete random variable X, the mean or expected value of X is

$$\mathbb{E}[X] = \sum_x x p_X(x). \tag{B.11}$$

The variance, which is a measure of the spread around the mean, is defined as

$$\mathrm{Var}[X] = \sum_x (x - \mathbb{E}[X])^2 p_X(x). \tag{B.12}$$

The standard deviation is the square root of the variance. Mean and variance are special cases of the *expected value* of any function of a random variable. If we denote the function by f, then the expected value of f is

$$\mathbb{E}[f(X)] = \sum_x f(x) p_X(x). \tag{B.13}$$

Thus, the mean is the expected value of X, and the variance is $\mathrm{Var}[X] = \mathbb{E}[(x - \mathbb{E}[X])^2]$.

For a continuous random variable X with probability density $p(x)$, the analogous expressions are obtained by replacing sums with integrals:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x p_X(x) dx \tag{B.14}$$

$$\mathrm{Var}[X] = \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 p_X(x) \tag{B.15}$$

$$\mathbb{E}[f(X)] = \int_{-\infty}^{\infty} f(x) p_X(x) dx \tag{B.16}$$

Both for discrete and continuous variables, there is a common alternative expression for the variance, namely the difference between the mean of the squares and the square of the mean:

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X])^2. \tag{B.17}$$

Exercise B.3 Prove this. ■

B.7 The normal distribution

B.7.1 Definition

The most important continuous distribution in applications of probability theory is the normal or Gaussian distribution. Its probability density function is

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{B.18}$$

We sometimes use $p(x) = \mathcal{N}(x; \mu, \sigma^2)$ as short-hand notation. The parameters μ and σ^2 do not have an a-priori meaning (they are just denoted suggestively), but they, of course turn out to be equal to the mean and variance of the distribution, respectively. The factor $\frac{1}{\sqrt{2\pi\sigma^2}}$ is needed for normalization. The *standard normal* distribution is a normal distribution with mean 0 and standard deviation 1.

B.7.2 Central limit theorem

The importance of the normal distribution derives mainly from the central limit theorem. Roughly, the central limit theorem states that the mean of a large number of independent random variables with identical probability distributions will follow an approximately normal distribution, regardless of the distribution of the original variables. This theorem is most powerful because of its last part: the distribution of the original variables is irrelevant. The theorem can be relaxed to allow for independent, but not identically distributed variables.

In mathematical models of perception, the central limit theorem always plays a role in the background: whenever we assume that the noise corrupting a stimulus is normally distributed, we are essentially motivating this using the central limit theorem. The random variable describing the noise corrupting a stimulus might be the sum of a large number of independent noise processes.

B.7.3 Multiplying two normal distributions

Let's consider the product of two Gaussian probability distributions over the same random variable X. One has mean μ_1 and variance σ_1^2, and the other mean μ_2 and variance σ_2^2. We will multiply these two distributions just as we would multiply regular functions, and then we will normalize the result (since the product is not automatically normalized). What is the resulting probability distribution?

We first write down the expressions for the two probability density functions:

$$p_1(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_2^2}} \tag{B.19}$$

$$p_2(x) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_2^2}} \tag{B.20}$$

Multiplying these two functions comes down to summing the exponents. We will do that first:

$$\text{sum of exponents} = -\frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{(x-\mu_2)^2}{2\sigma_2^2} \tag{B.21}$$

$$= -\frac{1}{2}\left(\frac{x^2 - 2\mu_1 x + \mu_1^2}{2\sigma_1^2} + \frac{x^2 - 2\mu_2 x + \mu_2^2}{2\sigma_2^2}\right). \tag{B.22}$$

We reorganize by collecting all terms containing x^2, and all containing x:

$$\text{sum of exponents} = -\frac{1}{2}\left(\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)x^2 - 2x\left(\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}\right) + \ldots\right) \tag{B.23}$$

$$= -\frac{1}{2}\left((J_1 + J_2)x^2 - 2x(J_1\mu_1 + J_2\mu_2) + \ldots\right), \tag{B.24}$$

where we used *precision* notation,

$$J_1 = \frac{1}{\sigma_1^2} \tag{B.25}$$

$$J_2 = \frac{1}{\sigma_2^2} \tag{B.26}$$

Moreover, here and in the following, the dots represent all terms that do not depend on x. When exponentiated, these terms become a multiplicative constant that is independent of x. Since the resulting product of distributions must be normalized at the end of the calculation anyhow, any multiplicative constant that we insert or leave out until that point is irrelevant. The sum of exponents can be written as

$$\text{sum of exponents} = -\frac{1}{2(J_1+J_2)^{-1}}\left(x^2 - 2x\frac{J_1\mu_1 + J_2\mu_2}{J_1+J_2} + \ldots\right) \tag{B.27}$$

$$= -\frac{1}{2(J_1+J_2)^{-1}}\left(x - \frac{J_1\mu_1 + J_2\mu_2}{J_1+J_2}\right)^2 + \ldots. \tag{B.28}$$

Thus, the product of the distributions in Eq. (B.20) is

$$p_1(x)p_2(x) \propto e^{-\frac{1}{2(J_1+J_2)^{-1}}\left(x^2 - 2x\frac{J_1\mu_1+J_2\mu_2}{J_1+J_2} + \ldots\right)} \tag{B.29}$$

$$= e^{-\frac{1}{2(J_1+J_2)^{-1}}\left(x - \frac{J_1\mu_1+J_2\mu_2}{J_1+J_2}\right)^2}, \tag{B.30}$$

where the proportionality sign absorbs all factors that are independent of x. We recognize this as another normal distribution, now with mean $\frac{J_1\mu_1+J_2\mu_2}{J_1+J_2}$ and variance $\frac{1}{J_1+J_2}$.

Exercise B.4 What is the correct normalization constant in Eq. (B.30)? ■

B.7.4 Multiplying N normal distributions

We now generalize the previous section to N normal distributions. This is used in Problem 10.9. Consider a set of N normal distributions over the same variable x The i^{th} distribution has mean μ_i and variance σ_i^2. The (unnormalized) product of these distributions is equal to

$$\sqrt{\prod_i \frac{J_i}{2\pi}}e^{-\frac{\sum_i J_i}{2}\left(\mu - \frac{\sum_i J_i x_i}{\sum_i J_i}\right)^2} e^{-\frac{1}{2}\sum_i J_i x_i^2} e^{-\frac{(\sum_i J_i x_i)^2}{2\sum_i J_i}}. \tag{B.31}$$

B.7.5 The cumulative normal distribution

The cumulative distribution function of a Gaussian distribution is not an elementary function (i.e. one built from exponentials, logarithms, and powers using addition, subtraction, multiplication, and division). However, there are two standard ways of expressing it: one is as an *error function*; the other, which we will prefer, in terms of the cumulative distribution of the standard normal density:

$$\Phi_{\text{standard}}(y) = \frac{1}{\sqrt{2\pi}}\int_0^y e^{-\frac{x^2}{2}}dx. \tag{B.32}$$

This function takes values between 0 and 1.

Exercise B.5 Show that $\Phi_{\text{standard}}(0) = 0.5$. ■

Exercise B.6 Show that the integral of a Gaussian distribution can be expressed in terms of the cumulative normal distribution as follows:

$$\int_{-\infty}^{x} \mathcal{N}(x;\mu,\sigma^2) = \Phi_{\text{standard}}\left(\frac{x-\mu}{\sigma}\right). \tag{B.33}$$

This is possible because any normal distribution can be shifted (by -μ) and scaled (by dividing

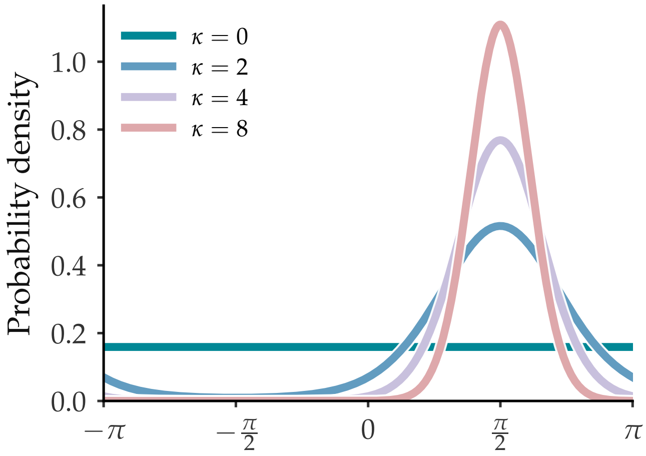


Figure B.4: Examples of Von Mises distributions. The domain of the stimulus/measurement is periodic.

> by σ) to obtain a standard normal distribution. ■

Two other integrals of the Gaussian distribution are useful:

$$\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \sqrt{2\pi\sigma^2} \tag{B.34}$$

$$\int_a^b \mathcal{N}(x;\mu,\sigma^2) = \Phi_{\text{standard}}\left(\frac{b-\mu}{\sigma}\right) - \Phi_{\text{standard}}\left(\frac{a-\mu}{\sigma}\right) \tag{B.35}$$

B.7.6 The Von Mises distribution

Some variables, such as orientation and motion direction naturally have a circular (periodic) domain. For such variables, a Gaussian distributions does not make sense. One solution is to choose a Von Mises distribution (see **Fig. B.4**). This can be regarded as the circular analog of a Gaussian distribution. It has two parameters: a circular mean, and a concentration parameter, which is similar to the reciprocal of the variance of a Gaussian. A Von Mises distribution over a circular variable s with domain $[0, 2\pi)$, circular mean μ_s, and concentration parameter κ_s is

$$p(s) = \frac{1}{2\pi I_0(\kappa_s)} e^{\kappa_s \cos(s-\mu_s)}. \tag{B.36}$$

Here, I_0 is the modified Bessel function of the first kind of order 0. This is a so-called *special function*, a function that is defined in terms of an integral or infinite series. Its precise definition is not important here; for us, all that matters is that $2\pi I_0(\kappa)$ normalizes the Von Mises distribution. When $\kappa = 0$, $I_0(\kappa) = 1$, and the Von Mises distribution becomes a uniform distribution on the circle. The higher κ, the more similar the Von Mises distribution becomes to a normal distribution. This is illustrated in **Fig. B.4**.

> **Exercise B.7** We said that μ is the circular mean, but how would you define the mean of a circular variable? ■

> **Exercise B.8** Show analytically that in the limit of large κ, the Von Mises distribution becomes

the normal distribution. (Hint: use the Taylor expansion of the cosine.) Moreover, show that the precision of that normal distribution is κ. ■

The distribution of a measurement given a stimulus can be Von Mises with circular mean s and concentration parameter κ:

$$p(x|s) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(x-s)}. \tag{B.37}$$

B.8 The delta distribution

A special type of random variable that we will encounter quite often is one that takes only one possible value. There is a special notation for the probability distribution of such a random variable. If X is a continuous random variable that always takes the value $X = a$, then we write for its distribution,

$$p(x) = \delta(x-a). \tag{B.38}$$

Here, δ is the Dirac delta function. It returns 0 unless the argument equals 0, in which case it returns infinity. Of course, infinity is not a number and therefore the Dirac delta function is strictly speaking not an ordinary function. This is not of practical concern, since the only place in which we will use the delta function is inside an integral. There, the following property holds for any function $f(x)$:

$$\int \delta(x-a) f(x) dx = f(a), \tag{B.39}$$

where we assumed that the region of integration contains a. In fact, Eq. (B.39) is the defining property of the Dirac delta. The delta function has the effect of evaluating the function f inside the integral at a single point, a.

We find it convenient to use the same notation for discrete as for continuous variables, i.e. write $\delta(x-a)$ instead of δ_{xa}. Then, the discrete analog of Eq. (B.39) is

$$\sum_x \delta(x-a) f(x) = f(a), \tag{B.40}$$

B.9 The Poisson distribution

A discrete probability distribution that we use to describe neural activity (Section 14.1) is the Poisson distribution. The possible values of a Poisson random variable are $0, 1, 2, 3, \ldots$ (there is no upper limit). The Poisson distribution has a free parameter, which we will call λ. The probability distribution of X is given by

$$p(x) = \Pr(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \tag{B.41}$$

where $x! = 1 \cdot 2 \cdot 3 \cdots x$ is the factorial operation. The Poisson distribution is discrete and only defined on the non-negative integers. The factor $e^{-\lambda}$ acts as a normalization factor. Unlike x, λ does not have to be an integer. The mean and the variance of a Poisson-distributed variable are both equal to λ.

B.10 Drawing from a probability distribution

In probabilistic modeling, we often have to draw random numbers according to a specified probability distribution. These draws are also called samples. Drawing random numbers is by no means trivial, but fortunately, most software packages have built-in random number generators for the most common probability distributions. We can then use these functions to custom-write code for drawing from probability distributions that are not pre-programmed.

B.11 Distributions involving multiple variables

Random variables can depend on each other in interesting ways. This is formalized in joint and conditional probability distributions, and in Bayes' rule, which we derive here formally. The concepts discussed in this section apply to both continuous and discrete variables. Thus, probability or p can refer to either probability mass or probability density. Since we consider multiple variables at the same time, we will generally use a subscript on p to denote which random variable(s) the probability distribution belongs to.

B.11.1 Joint probability

The *joint probability distribution* of random variables X and Y is denoted $p_{X,Y}(x,y)$, or in shorthand, $p(x,y)$. It is the probability of the values x and y as a pair. Summing over both x and y gives 1:

$$\sum_x \sum_y p(x,y) = 1. \tag{B.42}$$

For continuous variables, the integral over both variables is 1:

$$\iint p(x,y)dxdy = 1. \tag{B.43}$$

Joint probability is symmetric:

$$p(x,y) = p(y,x). \tag{B.44}$$

If X and Y represent events, the joint probability of X and Y is the probability that both occur, denoted $p(X,Y)$ or $p(X \cap Y)$. In the Venn diagram representation (**Fig. B.5**), we represent Y by another circle, intersecting the first one. The joint probability of X and Y is equal to the area of the intersection. It is always less than or equal to the area of each individual circle. This expresses the relations $p(X,Y) \leq p(X)$ and $p(X,Y) \leq p(Y)$. For example, the probability that it rains on a given day and you will be at work on time is smaller than the probability that it rains. These relations only hold for discrete variables.

B.11.2 Marginalization

Marginalization is the operation of obtaining from a joint distribution over multiple variables the distribution over a subset of those variables. For example, if $p(x,y)$ is the joint distribution of X and Y, then summing over Y produces the distribution of X alone:

$$\sum_y p(x,y) = p(x). \tag{B.45}$$

Marginalization is an important, frequently used procedure in applications of probability. A daily-life example: Imagine you have a cat and a dog. You carefully track what the probabilities are during the day that only your cat is present in the living room, only your dog, neither, or both. These probabilities are shown in Table B.1; this table, called a *contingency table*, represents the

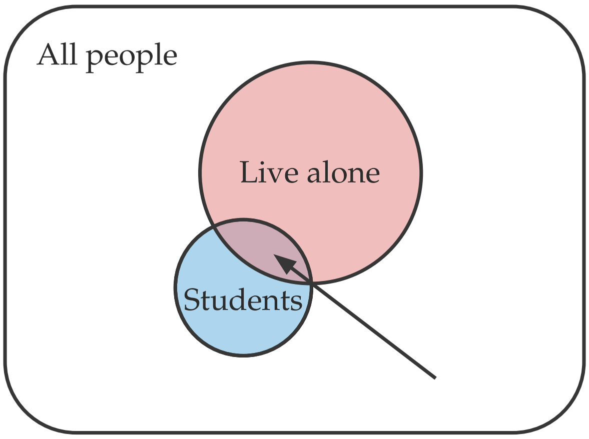


Figure B.5: The joint probability of the events "being a university student" and "living alone" is represented by the area of the intersection, indicated by the arrow.

	cat absent	cat present	total
dog absent	0.40	0.05	0.45
dog present	0.30	0.25	0.55
total	0.70	0.30	1

Table B.1: Frequencies of combinations of two random variables

probabilities of joint outcomes. The marginal probabilities are the probabilities that the cat is present or absent regardless of the dog, and the probabilities that the dog is present or absent regardless of the cat.

The continuous analog is obtained by replacing the sum by an integral:

$$\int p(x,y)dy = p(x). \tag{B.46}$$

This summation or integration is called "marginalization" because $p(x)$ and $p(y)$ are called the marginals of $p(x,y)$. If you think of (x,y) as a point in two-dimensional space, and the joint distribution providing z-values in this space, then the marginals are the distributions obtained by summing in either dimension (**Fig. B.6**). This results in two one-dimensional distributions that live in the "margins" of the original two-dimensional distribution.

B.11.3 Conditional probability

If X and Y are random variables, the probability distribution of X given Y is denoted by $p_{X|Y}(x|y)$. The "|" sign is read as "given" or "conditioned on". It is defined as the probability of x and y as a pair, divided by the probability of y:

$$p(x|y) = \frac{p(x,y)}{p(y)} \tag{B.47}$$

Consider these three examples of conditional probability when X and Y are discrete random variables:

- If the probability that it rains today and you arrive at work on time is equal to 0.4, and the probability that it rains today is 0.5, then the probability that you arrive at work on time given that it rains is 0.4/0.5 = 0.8.

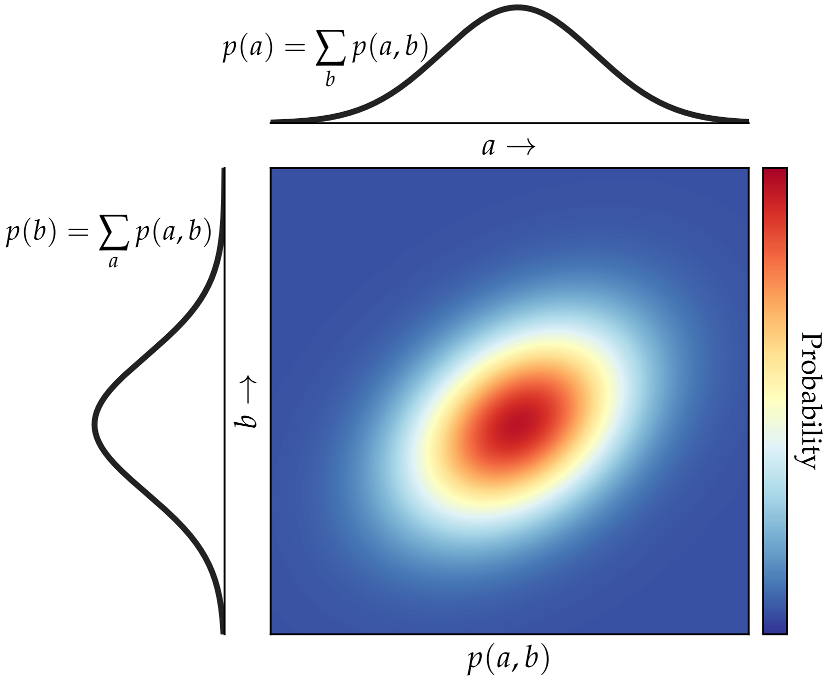


Figure B.6: The color plot represents the joint probability distribution of two random variables A and B, the black curves the two marginals, obtained by summing the joint across of the the two variables.

- The probability that I roll a 6 on a die given that I roll an even number is equal to (1/6)/(1/2) = 1/3.
- In a given country, each state has a different proportion of taxi drivers. The probability that a person randomly selected from a particular state is a taxi driver, $p(x|y)$, is equal to the proportion of people in the country who live in that state *and* are taxi drivers, $p(x,y)$, divided by the proportion of people living in that state, $p(y)$.

From the contingency table in Table B.1, conditional probabilities can readily be computed. For example, the probability that the cat is present given that the dog is present is 0.25 (cat and dog both present) divided by $0.30 + 0.25 = 0.55$ (dog present).

The conditional probability $p(X|Y)$ is the answer to the question: "of all outcomes that are consistent with event Y, what fraction is also consistent with event X?" Conditional probability of an event always lies between 0 and 1. In the Venn diagram representation (**Fig. B.9**), $p(X|Y)$ is equal to the area of the intersection divided by the area of the second circle. Similarly, the probability that Y occurs given that X occurs is equal to the area of the intersection divided by the area of the first circle.

It is easy to see, either from equation or from the figure, that $p(X|Y)$ is not equal to $p(Y|X)$. Mistakenly equating the two is known as the conditional probability fallacy or the prosecutor's fallacy (see Section 2.3). For example, the probability is very high that a professional basketball player is tall. However, the probability is very low that a tall person is a professional basketball player. If X is "being a basketball pro" and Y is "being tall", then this example illustrates that $p(X|Y)$ is not equal to $p(Y|X)$. The same holds for conditional probability distributions over continuous variables.

A (discrete or continuous) conditional probability distribution $p(x|y)$ is a probability distribution over x, but not over y. The distinction arises from the fact that when summed over x, $p(x|Y)$ adds

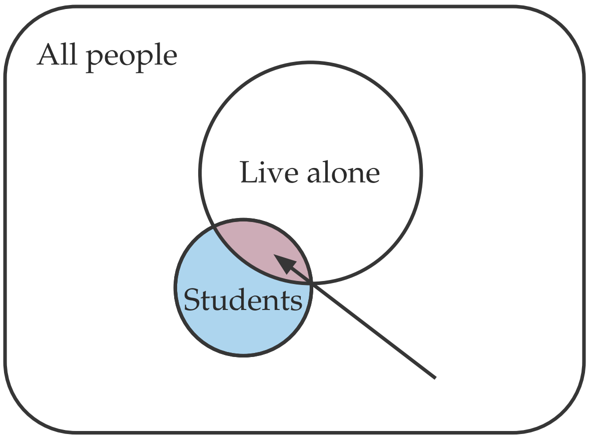


Figure B.7: The conditional probability of "living alone" given "being a university student" is represented by the area of the intersection divided by the area of the blue disc.

up to 1, but when summed over y it does not necessarily add up to 1.

Exercise B.9 a) Show formally that $p(x|y)$ is normalized as a function of x.
b) Give a counterexample that shows that $p(x|y)$ is not normalized as a function of y. ■

Although $p(x|y)$ is not normalized as a function of y, it is still a function of y. In the context of inference, it is called the *likelihood function* of y. Remember: likelihood functions are not probability distributions, because they are not normalized, and they are always functions of the variable *after* the "|" sign. It would be wrong to talk about $p(x|y)$ as "the likelihood of x".

We will now combine the notion of marginalization with the definition of conditional probability.

Exercise B.10 Show that:

$$p(x) = \sum_y p(x|y)p(y). \tag{B.48}$$

■

Eq. (B.48) and its continuous analog, $p(x) = \int p(x|y)p(y)dy$, are rules that we use throughout the book. Continuing on the taxi driver example: suppose you are interested in the probability that a randomly selected citizen is a taxi driver. You know for each state the proportion of taxi drivers. I also know the proportion of all citizens living in each state. To obtain my answer, you multiply those two proportions for every state and then sum over all provinces.

We can condition every probability in Eqs. (B.45) and (B.48) on a third random variable, z (this can be done with any rule in probability calculus). Then we get

$$p(x|z) = \sum_y p(x,y|z) \tag{B.49}$$

$$= \sum_y p(x|y,z)p(y|z) \tag{B.50}$$

Or in its integral form

$$p(x|z) = \int p(x|y,z)p(y|z)dy \tag{B.51}$$

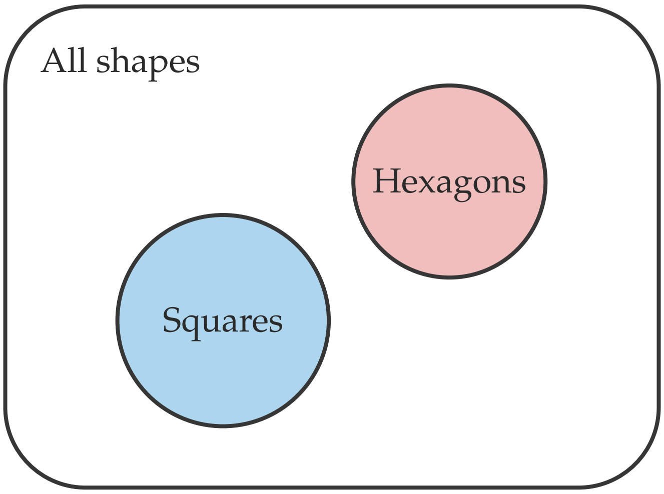


Figure B.8: Venn diagram depiction of the independence of two random variables.

> **Exercise B.11** Prove this formally using the definition of conditional probability and the marginalization rule. ■

Conditional distributions are not limited to a single random variable before and after the given sign. For example, one could consider the distribution of X and Y given Z and W, denoted by $p(x,y|z,w)$.

B.11.4 Independence

Two random variables X and Y are called independent if their joint distribution factorizes into the marginals, i.e., if

$$p(x,y) = p(x)p(y) \tag{B.52}$$

for all x and y. For example, the probability that I roll a 6 on a die and toss heads on a coin is the product of both events taken separately. Independence can be depicted graphically as in **Fig. B.8**: one can reconstruct the joint distribution by multiplying the marginals. The notion of independence is closely related to that of correlation: two independent random variables are also uncorrelated. The opposite is not true.

> **Exercise B.12** Why not? ■

> **Exercise B.13** If X and Y are independent, what can one say about the conditional distributions $p(x|y)$ and $p(y|x)$? ■

If X, Y, and Z denote three random variables, then X and Y are *conditionally independent* given Z if

$$p(x,y|z) = p(x|z)p(y|z) \tag{B.53}$$

for any values x, y, and z. We use this in Chapter 5. Be careful to not confuse conditional independence with independence!

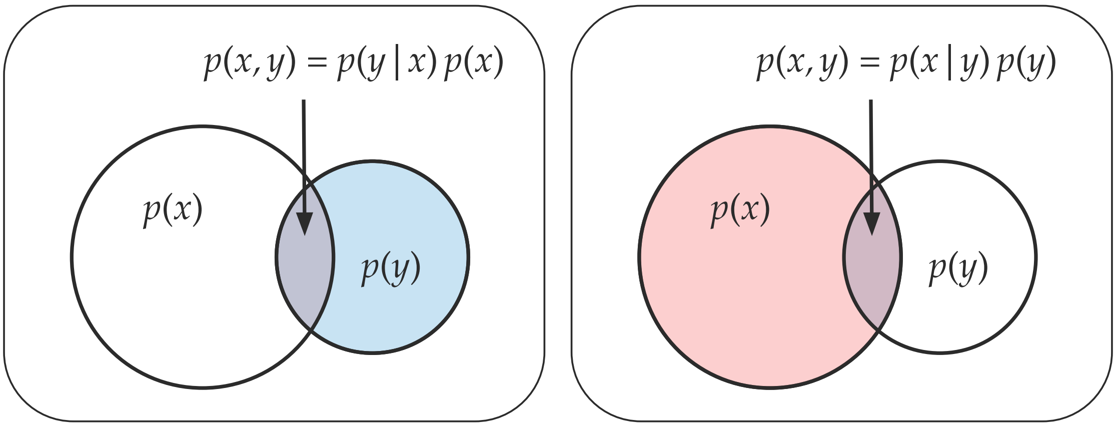


Figure B.9: Bayes' rule is obtained by writing the intersection area in two different ways and equating the two expressions.

B.11.5 Bayes' rule

We saw before that the conditional probabilities $p(x|y)$ and $p(y|x)$ are not equal. Bayes' rule relates them to one other:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}. \tag{B.54}$$

Here, $p(y|x)$ as a function of x is the *likelihood function* over x, $p(x)$ is the *prior distribution* over x, and $p(x|y)$ is the *posterior distribution* over x.

Exercise B.14 Before reading on, try to prove Bayes' rule using the equations in the preceding sections. ■

Here is how the proof goes: From Eq. (B.47), we know that $p(x,y) = p(x|y)p(y)$. By renaming x and y, we also obtain $p(y,x) = p(y|x)p(x)$. Joint probability is symmetric, $p(x,y) = p(y,x)$. From these three equations, it follows that $p(x|y)p(y) = p(y|x)p(x)$. Dividing both sides by $p(y)$ gives Bayes' rule.

Exercise B.15 Prove that the right-hand side of Eq. (B.55) is normalized over x. ■

The Venn diagram interpretation of Bayes' rule for events X and Y is that the area of overlap can be calculated in two ways (**Fig. B.9**): as a fraction of the X-circle area times the X-circle area, or as a fraction of the Y-circle area times the Y-circle area. Since the outcomes should be identical, this means that the two fractions can be expressed in terms of each other if one knows the ratio of areas of the X- and Y-circles.

Suppose that 1 in 100,000 people are professional basketball players, that 1 in 100 people are tall, and that 95% of basketball pros are tall. What is the probability that a tall person is a professional basketball player? We solve this problem using a direct application. of Bayes' rule: If X is "being a basketball pro" and Y is "being tall", then $p(X) = 0.00001$, $p(Y) = 0.01$, and $p(Y|X) = 0.95$. It follows that $p(X|Y) = \frac{0.95 \cdot 0.00001}{0.01} = 0.0095$, or about 1 in 1000.

Exercise B.16 Prove a different form of Bayes' rule:

$$p(x|y) = \frac{p(y|x)p(x)}{\sum_x p(y|x)p(x)}. \tag{B.55}$$

■

Since $p(y)$ does not depend on x, it is a constant multiplicative factor in Eq. (B.55), and acts as a normalization factor. Since the normalization is determined by integrating the numerator, $p(y|x)p(x)$, it is often left out for convenience and replaced by a proportionality sign:

$$p(x|y) \propto p(y|x)p(x) \tag{B.56}$$

Here, it is understood that the constant of proportionality is such that the left-hand side is normalized. See also the Box "Why the proportionality sign?" in Section 3.6. For binary variables, the normalization is usually written out, since it contains only two terms then.

Bayes' rule in the taxi driver example: you know what proportion of people in each state is a taxi driver. You know what proportion of the population lives in each state. You are told a particular individual is a taxi driver. What is your best guess of which state they are from? This is computed by multiplying the probability that someone lives in a certain state by the probability that someone in that state is a taxi driver, and comparing this product (the numerator of Bayes' rule) across states.

B.12 Functions of random variables

B.12.1 Functions of one variable: changing variables

In this section, we discuss the frequently occurring problem of transforming the distribution of a continuous random variable. The question is as follows. If X is a random variable with probability distribution $p_X(x)$, and $Y = f(X)$ is a new random variable obtained by applying the function or transformation f to X, what is the distribution of Y? In this section, we will use subscripts, such as X in $p_X(x)$, to avoid confusion, since there are multiple random variables.

An example: X is a random variable following a uniform distribution on $[0,1]$. $Y = X^2$ is a new random variable obtained by squaring outcomes of X. What is the distribution of Y?

An easy but wrong answer would be that because X follows a uniform distribution, Y does as well. It can be understood intuitively that this answer is wrong. When a number x lies between 0 and 1, squaring it will always make it smaller. Thus, even though the values of Y will also lie between 0 and 1, lower values in this range will have greater probability density than higher values do. The question can be answered correctly by considering the cumulative distribution functions of X and Y, which we will denote $P_X(x)$ and $P_Y(y)$, respectively:

$$\begin{aligned}
P_Y(y) &= \Pr(Y \leq y) && \text{(B.57)}\\
&= \Pr(X^2 \leq y) && \text{(B.58)}\\
&= \Pr(X \leq \sqrt{y}) && \text{(B.59)}\\
&= P_X(\sqrt{y}) && \text{(B.60)}
\end{aligned}$$

Now, we use the fact that the probability density function is the derivative of the cdf, Eq. (B.8), to find the pdf of y, denoted $p_Y(y)$:

$$\begin{aligned}
p_Y(y) &= \frac{dP_Y}{dy} && \text{(B.61)}\\
&= \frac{d}{dy}P_X(\sqrt{y}) && \text{(B.62)}\\
&= \left.\frac{dP_X}{dx}\right|_{x=\sqrt{y}} \frac{d}{dy}\sqrt{y} \quad \text{(chain rule)} && \text{(B.63)}\\
&= p_X(\sqrt{y})\frac{1}{2\sqrt{y}} && \text{(B.64)}\\
&= 1\cdot\frac{1}{2\sqrt{y}} && \text{(B.65)}\\
&= \frac{1}{2\sqrt{y}}. && \text{(B.66)}
\end{aligned}$$

The resulting distribution, $p_Y(y)$, is normalized (verify this) and conforms to our intuition: the probability density is higher for lower values of y.

We can verify the result through simulation: draw many random numbers from a uniform distribution between 0 and 1, square them, plot a finely binned, normalized histogram of the squares, and plot the function $\frac{1}{2\sqrt{y}}$ on top of it.

We could have stated the same problem with $p_X(x)$ being any distribution (instead of a uniform distribution). The calculation is then identical except for the last step. Then, we find

$$p_Y(y) = p_X(\sqrt{y})\frac{1}{2\sqrt{y}}. \tag{B.67}$$

Thus, the distribution of the squared variable is a product of the original distribution evaluated at the value of x that maps to y, $p_X(\sqrt{y})$, and an extra factor. The extra factor, called the *Jacobian*, is equal to the derivative of the mapping from y to x. It would be wrong to leave out the Jacobian and to assert that $p_Y(y) = p_X(\sqrt{y})$, It would also be wrong to assert that the distribution of a squared variable is given by the square of the distribution, $p_Y(y) = p_X(y)^2$.

The Jacobian appears not just in this example of squaring a random variable, but in our original, general problem. Suppose X is a random variable with probability distribution $p_X(x)$, and $Y = f(X)$, where f is a monotonically increasing function. What is the distribution of Y? We first define the inverse function f^{-1} as the function of y that "undoes" the effect of f, in other words, $f^{-1}(f(x)) = x$. This inverse function is well-defined because we assumed that f is monotonically increasing. Tempting but incorrect ways to obtain the distribution of Y would be to substitute the inverse function into $p(X)$, $p_Y(y) = p(X)(f^{-1}(y))$, or to assume that an operation applied to a distribution is the same as the distribution applied to the operation, $p_Y(y) = f(p_X(y))$. The correct approach is again to calculate the cumulative distribution of Y,

$$\begin{aligned}
P_Y(y) &= \Pr(Y \le y) && \text{(B.68)}\\
&= \Pr(f(X) \le y) && \text{(B.69)}\\
&= \Pr(X \le f^{-1}(y)) && \text{(B.70)}\\
&= P_X(f^{-1}(y)) && \text{(B.71)}
\end{aligned}$$

and from that the probability density function of y,

$$\begin{aligned} p_Y(y) &= \frac{dP_Y}{dy} && \text{(B.72)} \\ &= \frac{d}{dy} P_X(f^{-1}(y)) && \text{(B.73)} \\ &= \frac{dP_X}{dx}\bigg|_{x=f^{-1}(y)} \frac{df^{-1}}{dy} \quad \text{(chain rule)} && \text{(B.74)} \\ &= p_X(f^{-1}(y)) \frac{df^{-1}}{dy} && \text{(B.75)} \end{aligned}$$

So far, we considered a monotonically increasing function f. When f is instead monotonically decreasing, the final expression for $p_Y(y)$ acquires an extra minus sign.

Exercise B.17 Show this. ■

We can summarize both cases – monotonically increasing and decreasing – in a single equation:

$$p_Y(y) = p_X(f^{-1}(y)) \left| \frac{df^{-1}}{dy} \right|. \tag{B.76}$$

An informal but intuitive way of writing Eq. B.76 is $p_Y(y) = p_X(x) \left|\frac{dx}{dy}\right|$, where it is understood that $x = f^{-1}(y)$. How can the Jacobian, the factor $\left|\frac{dx}{dy}\right|$, be interpreted and why can it not be left out? Recall that for a continuous variable, the probability of an event is defined as the integral of the probability density function between two points. The integral can be approximated (and is in fact defined as) a sum of areas of very narrow rectangles that together fill the interval. The factor $\left|\frac{dx}{dy}\right|$ stems from the fact that a rectangle of width dy in the domain of the transformed variable Y does not always correspond to a rectangle of the same width in the domain of the original variable, X. The derivative can be regarded as a "magnification factor". Equivalently, we can state that $p_Y(y)|dy| = p_X(x)|dx|$, because the left side of this equation is the probability that $Y = f(X)$ lies between the values $y - \frac{dy}{2}$ and $y + \frac{dy}{2}$, which it does when x lies between $x - \frac{dx}{2}$ and $x + \frac{dx}{2}$. Note that transforming discrete random variables does not involve a Jacobian, since the probability mass is concentrated in discrete points.

B.12.2 Apple example

As a final illustration of the change-of-variables procedure, let's suppose that you are going to visit an apple orchard. You know very little about how fast apples grow or the duration of the growing season in the area, and you don't know the type of apples in the orchard. If a friend asked you what you thought the size of the apples in the orchard was going to be, you might initially respond that you have no idea. Upon more careful consideration, drawing upon your limited knowledge of apples in general, suppose you state that you have a uniform prior density over the diameter of apples in the orchard, from 3 to 13 cm. What, then, is your prior density over apple *volume*?

Before we derive the answer, let's appreciate the problem. Your uniform prior over apple diameter means that, for example, you consider it equally probable that an apple's diameter will lie between 5 and 6 cm as between 10 and 11 cm. If we approximate apples as spheres, then the volume of an apple is

$$v = \frac{4\pi r^3}{3} = \frac{\pi w^3}{6}, \tag{B.77}$$

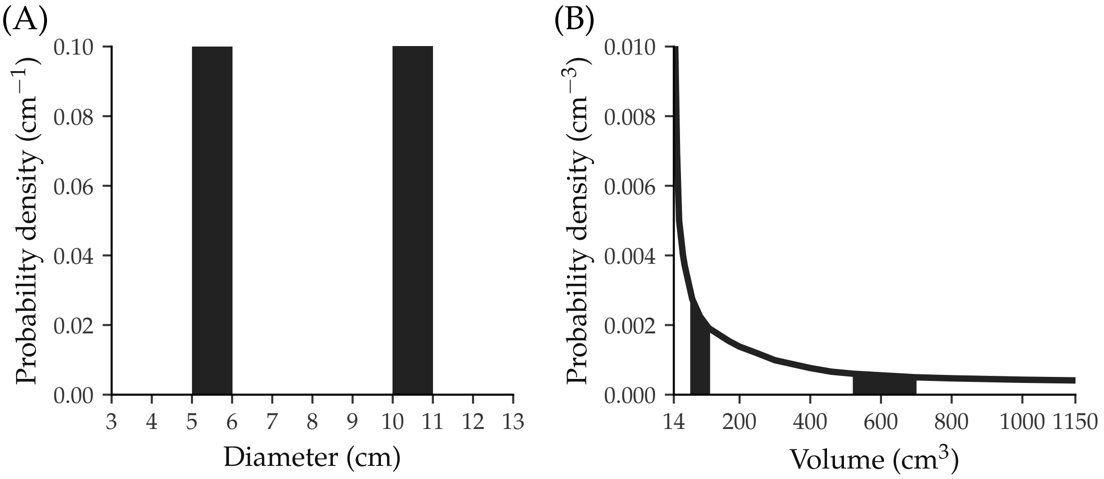


Figure B.10: Change of variables. **(A)** A uniform prior over apple diameter, from 3 to 13 cm (a range of 10cm). The prior density is a line at height 0.1 cm^{-1}, because the total area under the density must equal 1. The probability that the diameter lies between 5 and 6 cm is 0.1, as is the probability the diameter lies between 10 and 11 cm (filled rectangular areas). **(B)** The prior over apple volume. Each filled rectangular area is the probability that apple volume lies in the range corresponding to the apple diameters covered by the filled rectangles in A. Again, the total area under the density is 1, and the area of each filled rectangle is 0.1. (Note the differences in y-axis scales).

where r is the radius and $w = 2r$ is the diameter of the apple. The volumes corresponding to diameters of 5, 6, 10, and 11 cm are therefore (to the nearest integer) 65, 113, 524, and 697 cm^3, respectively. This means that you believe it is equally probable (10% probable, to be exact) for the volume of an apple to lie between 65 and 113 cm^3 – a range of 48 cm^3 – as it is to lie between 524 and 697 cm^3 – a range of 173 cm^3. Your prior density over apple volume, then, is clearly not flat. Rather, the density will be higher at smaller volumes (see **Fig. B.10**).

To derive the probability density over volume, we note that:

$$w = \sqrt[3]{\frac{6v}{\pi}} \tag{B.78}$$

For the derivative, we find that:

$$\frac{dw}{dv} = \sqrt[3]{\frac{2}{9\pi v^2}}. \tag{B.79}$$

Therefore,

$$\begin{aligned} p_V(v) &= p_W(w)\left|\frac{dw}{dv}\right| && \text{(B.80)} \\ &= p_W(w)\sqrt[3]{\frac{2}{9\pi v^2}} && \text{(B.81)} \\ &= \frac{1}{10\text{cm}}\sqrt[3]{\frac{2}{9\pi v^2}}. && \text{(B.82)} \end{aligned}$$

This is the curve plotted in **Fig. B.10B**.

B.12.3 On ignorance

An interesting consequence of the change-of-variables procedures, illustrated by the apple orchard example, is that it is not possible to be ignorant about every feature of a problem. For instance, it is not possible to be fully ignorant about apple size, generally defined. As we have just seen, if we are ignorant about apple diameter, in the sense that we consider a wide range of diameters to be equally probable, then we are consequently not ignorant about apple volume! When doing Bayesian statistical analysis, a researcher may want to incorporate as little prior opinion as possible into an analysis about which they feel they have almost no relevant background knowledge. How can they best do this, if by specifying their ignorance about a parameter, they are consequently specifying knowledge about related parameters? For instance, if a researcher has “no knowledge” of the standard deviation, σ, of a random variable, they may choose to use a flat prior over a very wide range of σ, but then they are implicitly specifying a non-uniform prior over the variance, σ^2. The search to develop appropriate default or reference priors for such situations is an interesting topic in the field of Bayesian statistical analysis.

In Bayesian models of behavior, this is usually not an issue, since the prior is assumed to be either derived from the experimental statistics or from natural statistics.

B.12.4 Marginalization formulation

It is instructive to phrase the problem of transforming the distribution of a random variable as a formal problem of marginalization. This formulation is equivalent, but gives more insight in some ways. We assume again that X is a random variable with probability distribution $p_X(x)$, and $Y = f(X)$, where f is a monotonically increasing function. As we discussed in Section B.8, a deterministic mapping such as f can be expressed as a delta distribution. Here, this distribution would take the form

$$p_{Y|X}(y|x) = \delta(y - f(x)). \tag{B.83}$$

Now we can compute the probability density at y formally using the marginalization identity from Eq. (B.48):

$$\begin{aligned} p_Y(y) &= \int_{-\infty}^{\infty} p_{Y|X}(y|x)p_X(x)dx & \text{(B.84)} \\ &= \int_{-\infty}^{\infty} \delta(y - f(x))p_X(x)dx & \text{(B.85)} \end{aligned}$$

In words, the probability density of y is the total probability of all values of x that get mapped to y by f. We can evaluate this expression by making a transformation of variables: $x = f^{-1}(t)$, so that $dx = \frac{df^{-1}}{dt}dt$. Substituting, we find

$$p_Y(y) = \int_{-\infty}^{\infty} \delta(y - t)p_X(f^{-1}(t))\frac{df^{-1}}{dt}dt. \tag{B.86}$$

We can now use Eq. (B.39) to evaluate the integral:

$$p_Y(y) = p_X(f^{-1}(y))\frac{df^{-1}}{dy}dy, \tag{B.87}$$

which is the same as Eq. (B.75). Again, when f is monotonically decreasing instead of increasing, we obtain the same result but with a minus sign.

Exercise B.18 Where does the minus sign come from in this formulation? ■

The advantage of this integral formulation is that the first equality in Eq. (B.84) is general and not limited to deterministic mappings from X to Y. Thus, the problem of transforming a random variable is simply a special case of a probabilistic mapping from Y to X, and the first equality in Eq. (B.84) can be applied for *any* conditional distribution $p(y|x)$.

A second advantage is that the expected value of any function $g(Y)$ of a random variable $Y = f(X)$ is now easy to transform:

$$\begin{aligned}
\mathbb{E}[g(Y)] &= \int g(y) p_Y(y) dy && \text{(B.88)}\\
&= \int g(y) \left(\int_{-\infty}^{\infty} \delta(y - f(x)) p_X(x) dx \right) dy && \text{(B.89)}\\
&= \int \left(\int_{-\infty}^{\infty} g(y) \delta(y - f(x)) dy \right) p_X(x) dx \quad \text{(swapped order of integration)} && \text{(B.90)}\\
&= \int g(f(x)) p_X(x) dx && \text{(B.91)}\\
&= \mathbb{E}[g(f(X))]. && \text{(B.92)}
\end{aligned}$$

In other words, the combination $p_Y(y)dy$ inside an integral is identical to $p_X(x)dx$, as long as $y = f(x)$ is substituted elsewhere in the integral. (The integration limits might also change accordingly.)

Exercise B.19 Use this to show that the mean of $aX + b$ is $a\mathbb{E}[X] + b$, and that its variance is $a^2\text{Var}[X]$. ■

A third advantage of the marginalization formulation is that it directly generalizes to functions of multiple variables, as we will now examine.

B.12.5 Functions of multiple variables

Suppose you roll two fair dice and add the outcomes. What is the probability distribution of the sum? Simple counting gives the answer: the outcome 2 can be reached in only one way (1+1) and therefore has probability $\frac{1}{36}$. The outcome 3 can be reached in two ways (1+2 and 2+1) and therefore has probability of $\frac{2}{36}$, etc. This results in the probability distribution shown in **Fig. B.11**. How do we calculate this distribution formally?

We call the random variables corresponding to both die rolls X and Y. Their sum is a new random variable, $Z = X + Y$. In other words,

$$p_{Z|X,Y}(z|x,y) = \delta(z - x - y). \qquad \text{(B.93)}$$

To calculate the distribution of Z, denoted $p_Z(z)$, we apply the discrete analog of Eq. B.84:

$$\begin{aligned}
p_Z(z) = \sum x = 1^6 \sum_{y=1}^{6} p_{Z|X,Y}(z|x,y) p_{X,Y}(x,y) && \text{(B.94)}\\
\sum x = 1^6 \sum_{y=1}^{6} p_{Z|X,Y}(z|x,y) p_X(x) p_Y(y) && \text{(B.95)}\\
\sum x = 1^6 p_X(x) \sum_{y=1}^{6} \delta(z - x - y) p_Y(y). && \text{(B.96)}
\end{aligned}$$

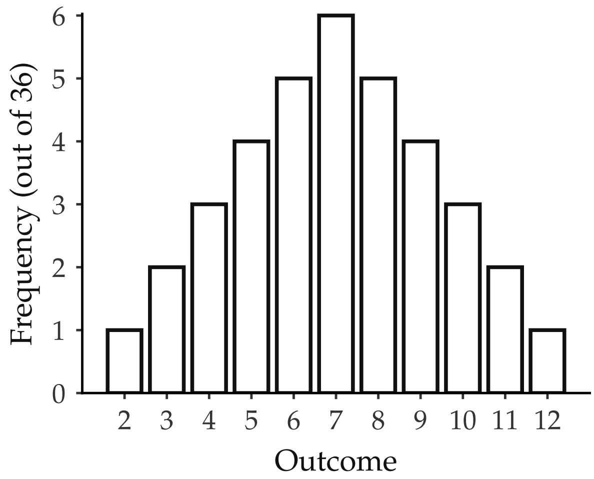


Figure B.11: The probability distribution of the sum of two die rolls.

We now use the property of the delta function, Eq. (B.40), as well as the condition that for $p_Y(y)$ to be nonzero, we must have $1 \le y \le 6$, therefore $1 \le z - x \le 6$, and therefore $z - 6 \le x \le z - 1$. Then,

$$p_Z(z)\sum x = \max(1,z-6)^{\min(z-1,6)} p_X(x)p_Y(z-x) \tag{B.97}$$

$$\sum x = \max(1,z-6)^{\min(z-1,6)} \frac{1}{6}\cdot\frac{1}{6} \tag{B.98}$$

$$\frac{1}{36}(\min(z-1,6)-\max(1,z-6)+1). \tag{B.99}$$

The same logic can be applied to a continuous distribution. Let X and Y be independent continuous variables with respective pdfs $p_X(x)$ and $p_Y(y)$. We define a new variable $Z = f(X,Y)$, with f any function, and denote by f_x^{-1} the inverse function of f for given x: $Y = f_x^{-1}(Z)$. (Such an inverse function does not always exist, but in the examples in this book, it does.) Then the distribution of Z is

$$p_Z(z) = \iint p_{Z|X,Y}(z|x,y)p_X(x)p_Y(y)dxdy \tag{B.100}$$

$$= \iint \delta(z - f(x,y))p_X(x)p_Y(y)dxdy \tag{B.101}$$

$$= \int dx p_X(x)\left(\delta(z-t)p_Y(f_x^{-1}(t))\left|\frac{df_x^{-1}}{dt}\right|dt\right) \tag{B.102}$$

$$= \int dx p_X(x)p_Y(f_x^{-1}(z))\left|\frac{df_x^{-1}}{dz}\right|, \tag{B.103}$$

where from the second to the third line we have made the transformation of variables $y = f_x^{-1}(t)$. One can think of the delta function as selecting a region of N-dimensional space – namely all points that map onto y – and of the integral as the total probability under $p(X)$ in that region.

Exercise B.20 If X and Y are independent and have a uniform distribution on $[0,1]$, compute the distribution of $Z = X + Y$. The answer is a special case of the *Irwin-Hall distribution*. ■

Exercise B.21 If X and Y are independent and have normal distributions, prove that $Z = X + Y$ also has a normal distribution. ■

So far, we have computed the distribution of a sum random variable. We can also use Eq. (B.103) to compute the distribution of nonlinear combinations of random variables, such as a product or quotient. The distribution of the product (or quotient) of two variables is not equal to the product (or quotient) of their distributions, and often very different. We will examine this in a problem.

B.13 Problems

Problem B.1 In a flower garden, 20% of flowers are tulips. Of those, one quarter are red. What is the probability that a random flower in this garden is a tulip but not red? Although it is easily possible to solve this problem using intuition, we would like you to formally apply the rules of probability.

Problem B.2 Four players, sitting around a table, are about to play a game. To determine who starts, one person throws two dice. The sum of the two numbers determines who starts, with the counting going clockwise starting with the roller being 1 (so with a sum of 5 or 9, the roller starts). What is the probability of starting for each player?

Problem B.3 You and I each roll a die once.

a) What is the probability that one of us rolls a 6 and the other an odd number?
b) What is the probability that at least one of us rolls a 6?
c) What is the probability that you roll higher than me?
d) What is the probability that our total is higher than 8?

Problem B.4 You are a student in a class of 30.

a) What is the probability that a particular classmate shares your birthday?
b) What is the probability that any classmate shares your birthday?
c) What is the probability that any two students share the same birthday?

Problem B.5 Let X, Y, and Z be random variables with possible values x, y, and z. Prove the following.

a) Conditional marginal:

$$p(x|z) = \sum_y p(x|y,z)p(y|z). \tag{B.104}$$

b) Conditional Bayes' rule:

$$p(x|y,z) = \frac{p(y|x,z)p(x|z)}{p(y|z)}. \tag{B.105}$$

Problem B.6 You and I alternately toss a coin. You start. If you toss heads, you win instantly. If you toss tails, it is my turn. If I toss tails, I win instantly. If I toss heads, it's your turn again. This repeats until one of us has won. What is your probability of winning this game?

Problem B.7 Email programs automatically classify emails as spam or not spam based on the words in the email. To do this, they use a Bayesian algorithm that works very much like the Bayesian inference that we discussed for medical diagnosis. Assume that 70% of all email is spam. Suppose that if an email is not spam, it has a 0.1% probability of including the word "bargain". Suppose further that if an email is spam, it has a 1% probability of including the word "bargain".

a) Draw the generative model diagram.

b) Fill in the associated probabilities:

$$p(\text{spam}) = \ldots \tag{B.106}$$

$$p(\text{no spam}) = \ldots \tag{B.107}$$

$$p(\text{contains “bargain"}|\text{spam}) = \ldots \tag{B.108}$$

$$p(\text{does not contain “bargain"}|\text{spam}) = \ldots \tag{B.109}$$

$$p(\text{contains “bargain"}|\text{not spam}) = \ldots \tag{B.110}$$

$$p(\text{does not contain “bargain"}|\text{not spam}) = \ldots \tag{B.111}$$

c) What is the prior probability that a random email is spam? What is the prior probability that a random email is not spam?

d) Suppose that a particular email contains the word “bargain”. What is the likelihood that it is spam? What is the likelihood that it is not spam?

e) Multiply the prior over “spam” by the likelihood of “spam”.

f) Multiply the prior over “not spam” by the likelihood of “not spam”.

g) Now you have what we called the “protoposterior”. Divide each the answers to (e) and (f) by the sum of both answers. What is the posterior probability that this particular email is spam?

Problem B.8 (Monty Hall problem) You are on a game show. The host shows you three doors. Behind one of them, a prize is hidden. You choose one door. The host, who knows behind which door the prize lies, opens a remaining door that does not contain the prize. The host then gives you the opportunity to switch your choice to the remaining unopened door, or stay with your original choice. Your door of choice gets opened and you receive the prize if it is there.

a) To maximize the probability of receiving the prize, what should you do?

b) If there are N doors and the host opens m of them (where $m < n-1$, what is the probability of receiving the prize under the best strategy?

c) Would the answer to (a) change if the host did not know which of the two remaining doors contained a prize, but the one he opens just happens not to contain the prize? Explain.

d) Speculate on why most people believe it does not matter whether you stay or switch.

Problem B.9 The probability density function of a Gaussian random variable X is given by Eq. (B.18). Show that the mean and variance of this random variable are equal to μ and σ^2, respectively.

Problem B.10 Use Eq. (B.76) to prove that if X is normally distributed with mean μ and variance σ^2, that $aX+b$ is normally distributed with mean $a\mu$ and variance $a^2\sigma^2$.

Problem B.11 If X is an exponentially distributed random variable where X has the positive real line as domain, what are the domain and distribution of $Y = e^X$?

Problem B.12 a) Someone tosses a fair coin three times. You observe the outcome of one toss, which is heads. What is the probability that heads are more common than tails among all three tosses?

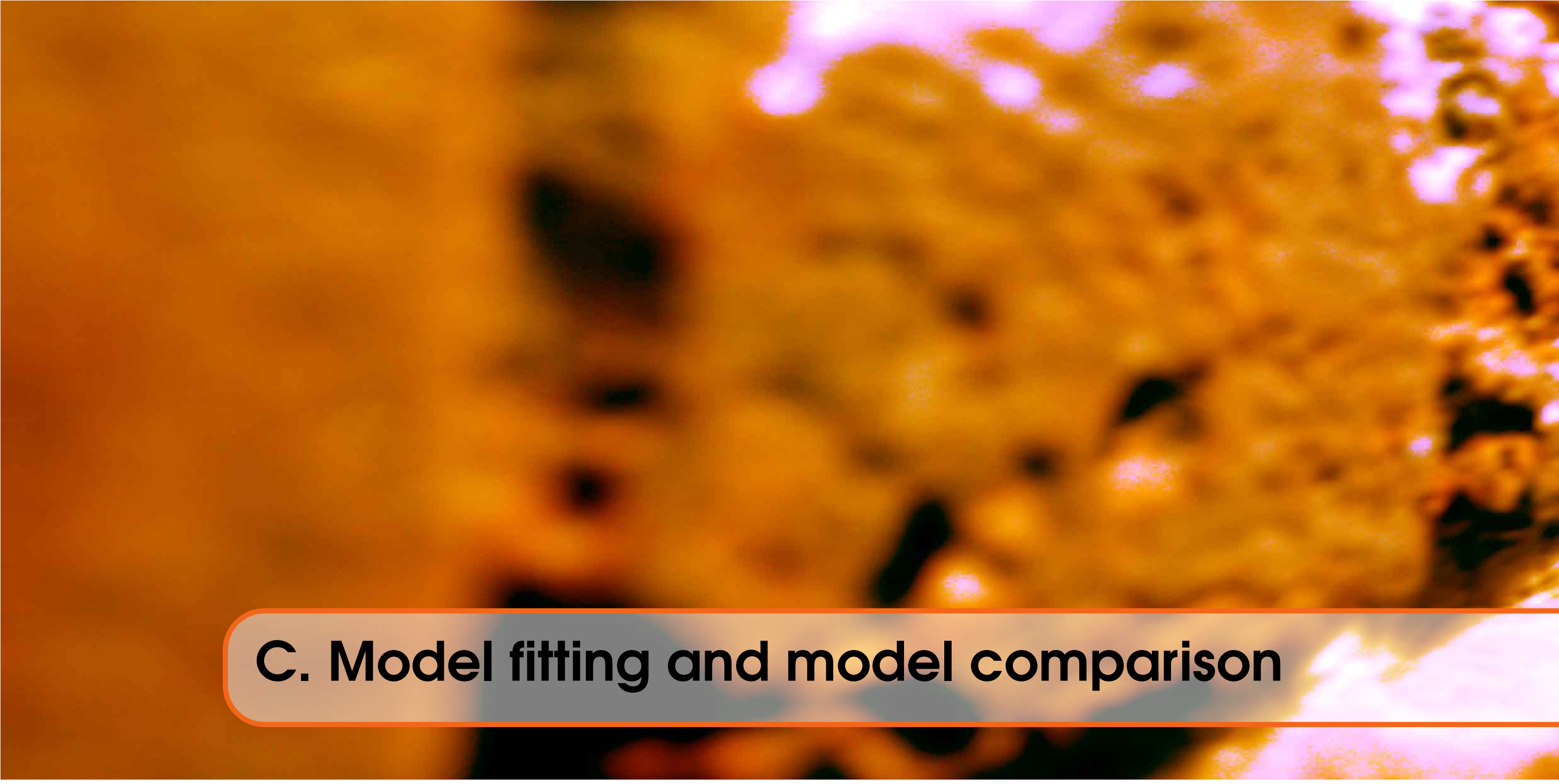
b) Someone tosses a fair coin N times. You observe the outcome of one toss, which is heads. What is the probability that heads are more common than tails among all N tosses? (For even N, interpret “more” as “strictly more”.)

Problem B.13 If X and Y are independent and have a uniform distribution on $[0,1]$ $(0 <= x <= 1)$, show that the product random variable $Z = XY$ has distribution $p_Z(z) = -\log z$. Verify that this distribution is normalized even though the density at 0 is infinity. This example illustrates how the distribution of a product is wildly different from the distributions of each of the factors.

Problem B.14 Consider two Von Mises distributions for the same random variable, one with mean μ_1 and concentration parameter κ_1, the other with mean μ_2 and concentration parameter κ_2. Show that the normalized product of these distributions is again a Von Mises distribution, and compute its mean and concentration parameter.

Problem B.15 If X and Y are independent standard normal variables, show that the quotient random variable $Z = \frac{Y}{X}$ has a Cauchy distribution, i.e.,

$$p_Z(z) = \frac{1}{\pi(1+z^2)} \tag{B.112}$$



C. Model fitting and model comparison

In this appendix, we describe how to fit a model to data. This is shorthand for "fitting the parameters of a model to data", in other words, adjusting the model's unknown parameters (such as σ) so that the data are accounted for as well as possible. The standard method for parameter fitting is *maximum-likelihood estimation (MLE)*.

The methods in this appendix do not apply exclusively to the Bayesian formalism. *Any* mathematical model of observer behavior can be fitted to data. One can also fit non-Bayesian models using the methods described in this appendix.

C.1 What is a model?

For a Bayesian model, Step 3 of the recipe outputs an estimate distribution or *response distribution*. This is a probabilistic mapping from stimuli to responses: $p(\text{subject responses}|\text{stimuli})$. Such a distribution, however, is not only what characterizes a Bayesian model but it is what characterizes really any model of behavior. Stronger yet, we can *define* a model of behavior as a response distribution $p(\text{subject responses}|\text{stimuli})$. Non-Bayesian models will generally not be built using the same three-step recipe; their $p(\text{subject responses}|\text{stimuli})$ might be arrived at in a different way.

C.2 Free parameters

Most models of behavior have *free parameters* (or simply parameters): variables of unknown value that are constant throughout an experiment. Most of them characterize the properties or beliefs of an observer. An example is the sensory noise level σ. For example, we may believe that a subject has a given σ that characterizes their visual acuity, and can not be changed over the course of our experiment. Other examples arise when the observer uses for their inference (Step 2) a generative model that is different from the true generative model (Step 1). We described such *model mismatch* in Chapter 3. For example, if in the decision rule of that chapter, the observer uses an assumed stimulus distribution width $\sigma_{s,\text{assumed}}$, then this is a free parameter. Similarly, in Chapters 7 and 8, if the observer uses a prior over class that is different from the true prevalence of the classes (which is typically 0.5/0.5), then the assumed prior probability of one of the classes is a free parameter (the prior probability of the other class is 1 minus that parameter).

We will denote the set of parameters of a model collectively by θ, and the model itself by M. We will make the dependence of the response distribution on the parameters and the model explicit by writing $p(\text{subject response}|\text{stimuli};\theta,M)$. The semicolon serves to separate the variables that vary from trial to trial (response and stimuli) from the model identity and the model parameters, which do *not* vary from trial to trial. When it is unambiguous which model we are talking about, we will leave out M.

C.3 The parameter likelihood

Free parameters have to be adjusted to best describe the data; this is called *parameter fitting*, or *model fitting*. The standard method for parameter fitting is *maximum-likelihood estimation*. The *likelihood* of a parameter combination θ under a model M is defined as the probability of the observed subject data given the stimuli experienced by the subject, the parameter combination θ, and the model:

$$\mathscr{L}_M(\theta) = p(\text{subject responses across all trials}|\text{stimuli across all trials};\theta,M) \tag{C.1}$$

In other words, the likelihood of a parameter combination is high when the model with that parameter combination, applied to the stimuli experienced by the subject, would relatively often produce the subject responses.

To make progress, we make a conditional independence assumption: on a given trial, the probability of a particular subject response only depends on the stimuli on that trial, the parameter combination, and the model. It does not depend on the subject responses on previous trials or on the stimuli on previous trials. As sequential dependencies between trials are well-documented in psychophysics, this assumption is often violated[1]. However, relaxing the assumption of conditional independent requires a model for sequential dependencies, which is beyond the scope of this book. The conditional independence assumption can be formulated as

$$\mathscr{L}_M(\theta) = \prod_{i=1}^{n_{\text{trials}}} p(\text{subject response on trial } i|\text{stimuli on trial } i;\theta,M), \tag{C.2}$$

where n_{trials} is the number of trials. The conditional probability in Eq. (C.3), is, for general responses and stimuli, exactly the specification of the predictions of a model; for example, in a Bayesian model, it would be the estimate distribution produced by Step 3 of the recipe, which defines the probability of an estimate of the world state of interest given a set of stimuli and a parameter combination. In Eq. (C.3), however, we substitute the *actual* subject response and the *actual* stimuli on the i^{th} trial.

In most of this book, the subject's response is equal to the maximum likelihood estimate of the world state of interest (this is not true in Chapter 13, where the action can be different from a world state estimate). In those cases,

$$\mathscr{L}_M(\theta) = \prod_{i=1}^{n_{\text{trials}}} p(\text{world state estimate on trial } i|\text{stimuli on trial } i;\theta,M), \tag{C.3}$$

[1] We want to warn the reader, however, to be mindful of this. If trial-trial correlations are strong then it means that trials are not independent. As trials become dependent, the statistical effects are often similar to having fewer trials. Consequently, the resulting p-values may be too low. When trial-trial correlations are strong then p-values need nontrivial corrections.

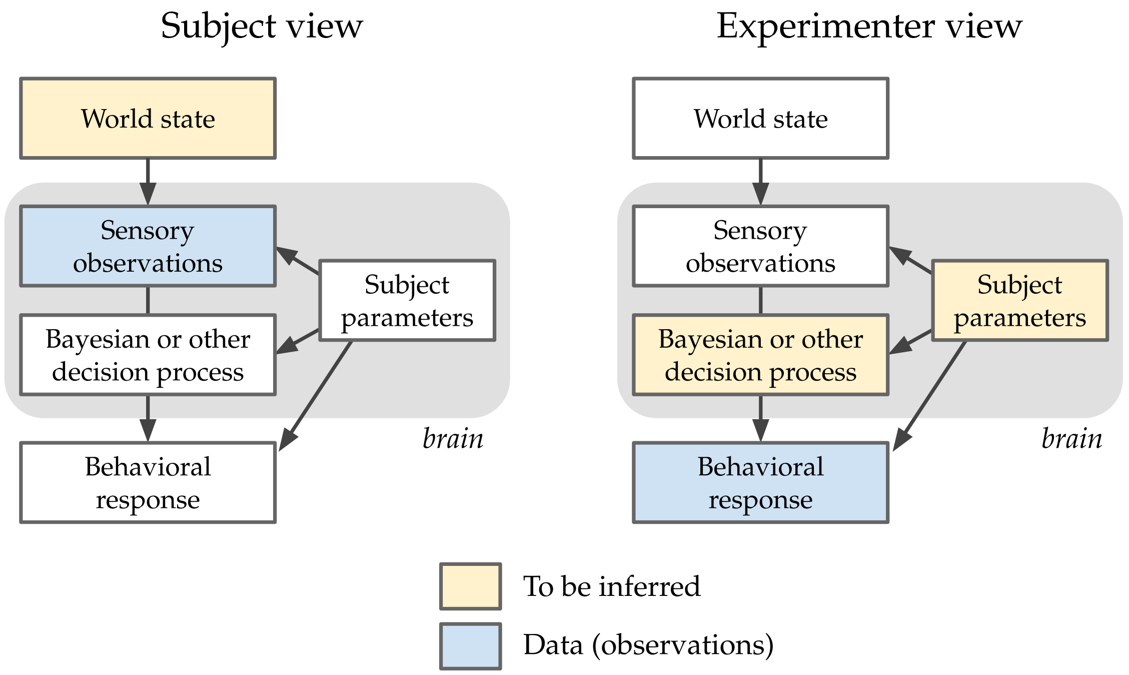


Figure C.1: Subject view versus experimenter view.

For example, in Chapter 4, the world state of interest is s, the subject's response is $\hat{s}$, and the likelihood of σ would be

$$\mathscr{L}(\sigma;\text{data}) = \prod_{i=1}^{n_{\text{trials}}} p(\hat{s}_i|s_i,\sigma). \tag{C.4}$$

Note that the estimate and the stimulus have an index i to label the i^{th} trial, but σ does not because it is common to all trials.

C.4 Maximum-likelihood estimation

Maximum-likelihood estimation of the parameters θ means finding the values of θ such that $\mathscr{L}_M(\theta)$ is highest. This is equivalent to maximizing $\log \mathscr{L}_M(\theta)$, because the logarithm is a monotonically increasing function. It is often more convenient to maximize the log likelihood than the likelihood itself. From Eq. (C.3), we know that the log likelihood is

$$\log \mathscr{L}_M(\theta) = \sum_{i=1}^{n_{\text{trials}}} \log p(\text{world state estimate on trial } i|\text{stimuli on trial } i;\theta,M), \tag{C.5}$$

It bears emphasizing that the parameter likelihood maximized in model fitting is conceptually similar to but different from any likelihood we encountered in previous chapters: it is over the parameters of the model (which are unknown to the experimenter), not over a relevant state of the world (which is typically known to the experimenter but unknown to the subject). Thus is illustrated in **Fig. C.1**. Throughout this Chapter and in Bayesian modeling practice in general, it is important to distinguish the observer's decision process (Step 2) and the experimenter's model of the observer (Step 4).

Sometimes, priors over parameters are used in model fitting. This makes the model fitting process Bayesian. However, choosing a prior over parameters requires additional justification and

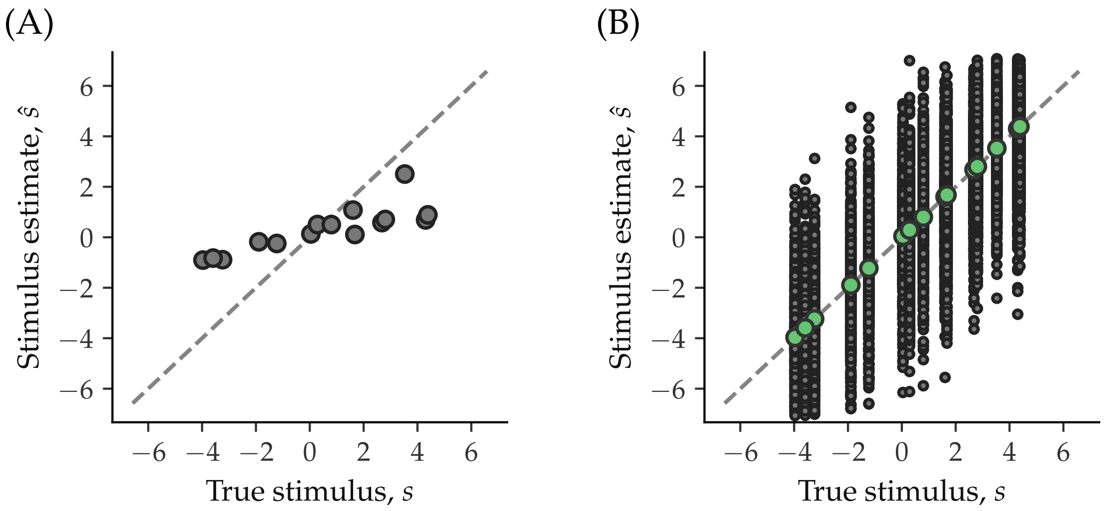


Figure C.2: (A) Example data set used in Section 8.4. The subject estimates a real-valued stimulus. **(B)** Model checking: prediction of the simple model after fitting σ. Black dots: individual simulations. Green circles: average across simulations for the same stimulus. The model does not fit well.

is not often necessary for fitting models, as long as the data set is sufficiently large (many trials). We will not discuss priors over parameters here.

In the following sections, we will revisit several models from previous chapters, and perform maximum-likelihood estimation of the model's parameters based on hypothetical data sets.

C.5 Fitting data from the estimation task in Chapter 3

The first step in fitting a model is always to specify the nature of your data. Suppose you perform an estimation experiment like in Chapter 3, where you draw a stimulus from a distribution $p(s)$ and the observer estimates the stimulus. Over 15 trials, you collect the following data:

Trial	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
s	1.61	5.50	-6.78	2.59	0.96	-3.92	-1.30	1.03	10.74	8.31	-4.05	9.10	2.18	-0.19	2.14
$\hat{s}$	0.37	1.62	-1.17	1.66	1.17	-0.79	-1.14	0.76	4.31	2.86	-0.61	3.25	0.48	0.12	0.18

Table C.1: Example data

These data are shown in **Fig. C.2A**.

C.5.1 Simple model

Even based on this small data set, it is already possible to fit a model. Let's first consider a model in which the observer does not use a prior and simply reports the measurement. Then, $\hat{s} = x$ and the distribution of $\hat{s}$ given s is

$$p(\hat{s}|s,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\hat{s}-s)^2}{2\sigma^2}}. \tag{C.6}$$

This fully specifies the model. We now want to estimate the parameter σ. The log likelihood of σ is, from Eq. (C.4),

$$\log \mathscr{L}(\sigma;\text{data}) = \sum_{i=1}^{n_\text{trials}} \log p(\hat{s}_i|s_i,\sigma) \tag{C.7}$$

$$= \sum_{i=1}^{n_\text{trials}} \left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(\hat{s}_i - s_i)^2}{2\sigma^2} \right] \tag{C.8}$$

Our goal is to find the maximum-likelihood of σ, which we will denote by $sig\hat{m}a$; this is the value of σ for which the log likelihood is highest.

Method 1: Analytical calculation

In this case, we can maximize the log likelihood analytically, by setting the derivative of $\log L$ with respect to σ to 0:

$$0 = \frac{d}{d\sigma} \log \mathscr{L}(\sigma;\text{data}) \tag{C.9}$$

$$= -\frac{n_\text{trials}}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n_\text{trials}} (\hat{s}_i - s_i)^2. \tag{C.10}$$

Solving for σ, we find the maximum-likelihood estimate of σ:

$$\hat{\sigma} = \sqrt{\frac{1}{n_\text{trials}} \sum_{i=1}^{n_\text{trials}} (\hat{s}_i - s_i)^2} \tag{C.11}$$

Since $\hat{s}_i - s_i$ is the observer's estimation error on the i^th trial, the right-hand side is the *root-mean-square error*. This answer is not surprising: under a model in which estimates are normally distributed around the true value, as in Eq. (C.6), the standard deviation of that normal distribution is estimated as the empirical standard deviation. In the data set above, the answer is $\hat{\sigma} = 3.49$.

There are no disadvantages to deriving the minimum analytically, but the option is rarely available: setting the derivative to 0 will yield a solvable equation in only very few simple cases. In all other cases, the log likelihood in Eq. (C.5) has to be maximized numerically. We will now explore methods for numerical maximum-likelihood estimation.

Method 2: Grid search

The simplest numerical method is to define a grid of possible values of σ, for example from 0.5 to 10 in steps 0.01 , and simply evaluate the log likelihood on this grid. This has the advantage that we can plot the log likelihood function (or the likelihood function), which we did in **Fig. C.3**.

Exercise C.1 Remark how in **Fig. C.3** the log likelihood looks quite flat and the Likelihood looks very peaked and the maximal value is very small (10^-14). What do these statements mean about what we can know about σ? ■

Exercise C.2 Choosing a suitable "base" vector of hypothesized values is a matter of intuition and experience, and of course you can always modify it.

a) What happens if we include values lower than 0.5 in sigma_hyp_vec?

b) Why would it be pointless to extend the range beyond 10?

■

In **Fig. C.3**, we first observe that the likelihood function is clearly not Gaussian, or equivalently, the log likelihood function is clearly not a parabola. In the observer models of previous chapters,

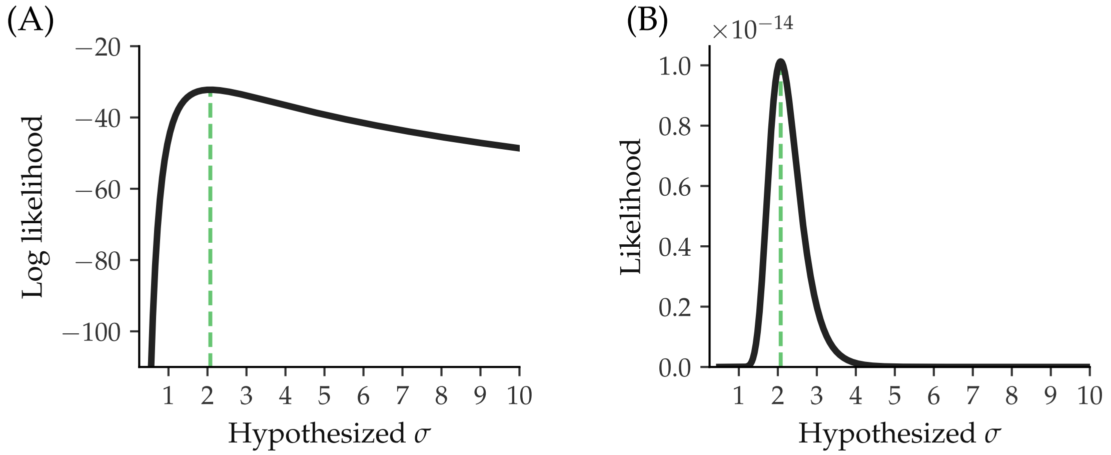


Figure C.3: Log likelihood and likelihood of σ. The dashed lines mark the maximum-likelihood estimate of σ, which is $\hat{\sigma} = 2.07$. The value of the log likelihood at the maximum is $\log \mathscr{L}^* = -32.2$.

the likelihood of a continuous variable was usually Gaussian. However, it is not surprising that this cannot be the case here: since σ can only take positive values, a Gaussian likelihood (which stretches to negative infinity) would be impossible.

Second, we note that the log likelihood takes values in the negative hundreds. In behavioral experiments, it is common for log likelihoods to be in the negative hundreds or thousands. This happens because probability mass functions always, and probability density functions often take only values smaller than 1. That means that each individual probability is smaller than 1, and its log is negative. The sum in Eq. (C.5) will grow negative more or less in proportion to the number of trials.

We can now simply find the value on the grid for which the log likelihood is maximal. This returns an estimate of 3.49, consistent with our analytical calculation.

It is useful to reflect on the advantages and disadvantages of a grid search. The advantages are: 1) Transparency. In a grid search, you know exactly what you are doing. 2) Visualization. Plotting the likelihood is useful so that you are more confident that it does not have a strange shape, and so that you know its width. The disadvantages are: 1) one is limited to the values on the grid and can never find any value in between; 2) the method does not scale up if you have more than a few parameters. For example, 1000 values per parameter for 4 parameters would be 10^{12} parameter combinations, which would take a very long time to compute. These two problems interact: if you use a finer grid to counter the first problem, you make the second problem even worse.

Method 3: Numerical optimization algorithm

Because of this combinatorial problem, it is often necessary to look beyond grid search. All ecosystems for scientific computation, including Python, Matlab, R, and Mathematica, have algorithms for minimizing arbitrary functions. Optimization algorithms are usually *iterative*: starting from an initial point, a particular routine is executed repeatedly until a termination criterion is satisfied. Many optimization algorithms are based on some form of gradient descent, where, from the current point, the next point is chosen nearby, in the direction of the steepest slope (see 6.4). Others are *global optimization algorithms*, that do not always choose the next point to be nearby and that sometimes try to estimate the large-scale structure of the function to be optimized. Many optimizers that work well in that space are Bayesian by themselves. Many have been developed in the context of machine learning under the term of Bayesian Hyperparameter Optimization. We recommend trying a range of algorithms to be sure - such optimization problems are hard and the specific algorithm used does make a difference. You can feel more confident that you found the true maximum if multiple different methods return the same result.

Optimization algorithms do not always find the global optimum. This could because there is a "ridge" of parameter combinations at which the log likelihood is close to the maximum log likelihood, or because the log likelihood landscape is irregular with many local maxima. A common way to alleviate this problem is "multistart", i.e. to initialize the optimizer with many different parameter combinations (either chosen randomly or systematically but sufficiently far apart), and pick the result of the best run.

Model checking

So, now you have found parameter estimates using one of the three methods. In modeling behavioral data, a common mistake is to take estimates of parameters seriously without verifying at least superficially that the model actually fits the data well. Perhaps the root of this problem is a confusion of "best" with "good": even though maximum-likelihood estimation provides the highest likelihood within the context of the model that is being fitted, that highest likelihood might still be low. This type of mistake has consequences. Parameter estimates of a poorly fitting model are meaningless: it is like estimating the size of the earth while assuming that the earth is flat.

The easiest way (and in many cases sufficient) to check whether a model fits well, is to plot the data along with the corresponding representation of the fitted model. One way to do that in our current example is to draw stimulus estimates from the normal distribution in Eq. (C.6), using $\hat{\sigma} = 3.49$ in place of σ. In **Fig. C.3B**, we did that for 500 repetitions of each presented stimulus. Clearly, the trend in the model is very different from the data, and we have to look for a new model.

In short, the act of fitting a model does not make the model good! Always verify that a model fits well before attaching any importance to its parameter estimates. If a model does not fit well, find a better model.

C.5.2 A better model

For the data in **Fig. C.2**, we can easily come up with a better model: it looks like the observer's estimates are affected by a prior centered at 0. Therefore, we also fit the full Bayesian model from Chapter 3: the observer uses a Gaussian prior with mean 0 and standard deviation σ_s. Then, the distribution of $\hat{s}$ given s is, from Eq. (4.6)

$$p(\hat{s}|s;\sigma,\sigma_s) = \mathcal{N}\left(\hat{s};\frac{\frac{s}{\sigma^2}+\frac{\mu}{\sigma_s^2}}{\frac{1}{\sigma^2}+\frac{1}{\sigma_s^2}},\frac{\frac{1}{\sigma^2}}{(\frac{1}{\sigma^2}+\frac{1}{\sigma_s^2})^2}\right). \tag{C.12}$$

We will not assume that we know σ_s. Then, this model has two parameters, σ and σ_s. The log likelihood of both parameters is:

$$\log \mathcal{L}(\sigma,\sigma_s;\text{data}) = \sum_{i=1}^{n_{\text{trials}}} \log p(\hat{s}_i|s_i;\sigma,\sigma_s). \tag{C.13}$$

An analytical approach is not feasible in this case. Therefore, the maximum-likelihood estimates of σ and σ_s have to be calculated numerically. Here, we do this for Method 2, grid search. We define a grid for σ and one for σ_s. We choose both vectors to range from 0.02 to 10 in 100 steps. Then, we can loop over both vectors (double for-loop), and for each combination of σ and σ_s on the grid evaluate the log likelihood from Eq. (C.13). After completing this, we can plot the log likelihood as a function of σ and σ_s; since there are two independent variables, one way to make this plot is using a heat map (**Fig. C.4A**). We can find the maximum numerically: $\hat{\sigma} = 1.76$ and $\hat{\sigma}_s = 1.24$; these values are close to the ones we used to generate the data, which were 1.5 and 1, respectively.

(Because of noise in the data, we never expect to find the exact values that generated the data.) The value of the maximum is $\log \mathscr{L}^* = -13.1$, whereas in the simple model, it was much worse, -40.0.

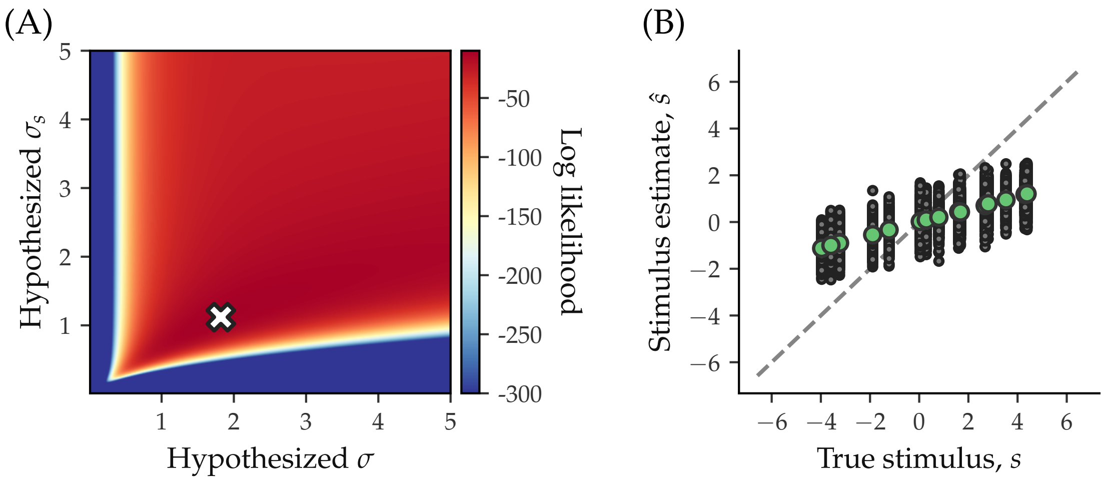


Figure C.4: A better model of the same data. **(A)** Log likelihood landscape as a function of σ and σ_s for the data in **Fig. C.2A**. A more red color represents a higher log likelihood. Log likelihood values smaller than -300 were set to -300, because otherwise, the differences in log likelihood near the peak would not be visible. The maximum is shown as a white X; the value at the maximum is $\log \mathscr{L}^* = -10.2$. **(B)** Model checking: predictions of the model with the parameter estimates from part **(A)**. Black dots: individual simulations. Green circles: average across simulations for the same stimulus.

To check the model, we again plot the model's prediction for the scatterplot of stimulus estimate versus true stimulus (**Fig. C.4B**). The prediction is now much better than in **Fig. C.2B**. We now witness the consequences of blindly believing parameter estimates in poorly fitting models: if we had believed the parameter estimate $\hat{\sigma} = 3.49$ from Section C.5, we would have been completely off from the true value of 1.5. This confirms the importance of checking a model, and if it fits poorly, rejecting it and looking for a better one.

C.5.3 Model comparison

In the simple model, the log likelihood at the maximum was roughly $\log \mathscr{L}^* = -40.0$. In the more complex model, this was roughly $\log \mathscr{L}^* = -13.1$. Is a difference of 26.9 big enough that we can reject the simple model? The more complex model has one parameter extra, which makes it fit better, but is the gain in maximum log likelihood "worth it"? This is an important and very common question in model fitting: how to compare models after fitting? Two common methods for answering these questions are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Both take the maximum log likelihood, $\log \mathscr{L}^*$, as the starting point, but then penalize the model for the number of free parameters. The nature of the penalty differs between AIC and BIC. In AIC, the number of free parameters gets subtracted from $\log \mathscr{L}^*$, then the result gets multiplied by -2. The factor does not do anything meaningful, it is just a convention so that for a Gaussian model, the leading term of AIC and BIC is the sum of squared errors. In BIC, the number of free parameters multiplied by half the logarithm of the number of trials gets subtracted from $\log \mathscr{L}^*$, then the result gets multiplied by -2. In equations,

$$\mathrm{AIC} = -2\left(\log \mathscr{L}^* - n_{\text{parameters}}\right) \tag{C.14}$$

$$\mathrm{BIC} = -2\left(\log \mathscr{L}^* - \frac{\log n_{\text{trials}}}{2} n_{\text{parameters}}\right) \tag{C.15}$$

Higher AIC or BIC means that a model is *worse*. However, both AIC and BIC derive from strong assumptions about the involved data or models and care is thus recommended in the interpretation of their results.

C.5.4 Cross-validation

In K-fold cross-validation, the data are divided into K parts. These parts have to be statistically equivalent – for example, it cannot be the case that all stimuli of one type are in the same part. In turn, one of the parts is left out of the data; the models are than fitted to the remaining data. The goodness of fit of the model is evaluated by computing the log likelihood of the fitted parameters on the left-out part of the data. By cycling through the K parts, K values are obtained, which are then averaged; the result is called the cross-validated log likelihood. The cross-validated log likelihood can be used directly to compare models. Good values for K are 5 or 10; if K is too low, the result will too much depend on the random composition of the parts, whereas if K is too high, there will be too few trials in each part. The philosophy of cross-validation is that a good model should be able to make good predictions for unseen data. The underlying assumptions of cross-validation are thus comparably simple.

C.5.5 Comparing model comparison methods

The origin of these metrics is beyond the scope of this book, but it is worth pointing out that they have different roots: AIC and cross-validation are meant to measure a model's (in)ability to *predict* new data, BIC to *explain* existing data – a subtle but important conceptual difference. In practice, the most important difference between AIC and BIC is that the BIC penalty for an extra parameter is almost always greater than the BIC penalty. However, conceptually, cross-validation is conceptually simpler and, arguably, requires weaker assumptions than AIC and BIC.

Exercise C.3 Compare AIC and BIC. Why are they different? ■

It is believed that AIC might underpenalize free parameters whereas BIC might overpenalize free parameters, but these issues are subtle and subject to debate. Another debate is whether cross-validation is universally preferable over using either AIC and BIC.

Pragmatically, we recommend to calculate AIC, BIC, as well as the cross-validated log likelihood, and only be confident in the results if they are highly consistent across the three metric. Drawing strong conclusions when their opposite would be obtained by changing the model comparison metric is bad science.

Besides consistency, magnitude matters. How big of a difference in model comparison metric is large? Of course, any classification is arbitrary, but that hasn't stopped authors from postulating ones anyhow. For example, Jeffreys came up with a narrative classification that we can apply to a difference in AIC, a difference in BIC, or to -2 times a difference in cross-validated log likelihood. In this classification, a difference greater than 4.6 is "strong", between 2.3 and 4.6 is "substantial", and below 2.3 "barely worth mentioning". Such labels implicitly refer to the distribution of strength of evidence found in experiments in a particular field, but this distribution is hard to establish and field-dependent. Playing along, in the context of psychology and neuroscience, we prefer to be more cautious and call differences greater than approximately 10 "substantial" and larger than approximately 20 "large". Better than playing along would be to completely extract oneself from narrative labels, but that is difficult for human scientists.

C.5.6 Likelihood ratio test

The likelihood ratio test is a method that is a bit different in nature from AIC, BIC, and cross-validation. This method can be applied when one model is a subset of another model. For example, a model with a guessing rate can reduce to a model with that guessing rate set to 0. To compare two such models, one can use a *likelihood ratio test*. When M_1 is the more specific (restricted; null) model and M_2 is the more general (unrestricted) one, the test statistic is

$$D = 2\left(\log \mathscr{L}^*_{M_2} - \log \mathscr{L}^*_{M_1}\right). \tag{C.16}$$

If the true model were M_1 (null model), then D would approximately follow a chi squared distribution with number of degrees of freedom equal to the difference in the numbers of free parameters of M_2 and M_1. The p-value of the test is the probability that D under the null model is greater than the observed value of D. When one model is not "nested" inside another model, the likelihood ratio cannot be used.

C.5.7 Parameter recovery and model recovery

In practice, maximum-likelihood estimation and model comparison can get complicated when there are several models, each having multiple parameters, and the parameter likelihood is not easily computed. Therefore, programming mistakes are easily made. A very useful debugging check is to create "fake data sets" from each of the models under consideration, by simulating measurements and observer decisions according to the model, like we have done several times in previous chapters, and subsequently run the maximum-likelihood fitting and Bayesian model comparison codes on these fake data sets. On each data set, the ML fitting should return parameter values close to the ones with which the data set was created. To verify this properly, it is best not create a single data set per model, but multiple, with different parameter combinations. Moreover, on each data set, Bayesian model comparison should show the true underlying model as the winner. These checks are contingent on the models being sufficiently distinguishable and the fake data sets being large enough. However, if systematically model B wins on data sets generated from model A, then you know for sure something is wrong in your model comparison code. Besides debugging, this model recovery process also serves to determine whether two models are in practice distinguishable.

C.5.8 Limitations of model comparison

Different scientists want to answer distinct questions, and hence model comparison does not have a simple correct answer. For some questions, some scientists might only look for the best-fitting model. For other questions scientists may want them to encapsulate concepts that make sense as descriptions of perceptual, cognitive, or motor processes. Just adding parameters to a model to achieve a better model comparison metric does not make that model or those parameters psychologically meaningful. Many scientists try to make every component of the model justified by one of the following:

- An understanding of the ecological niche or the problem
- Previous models that have that parameter or independent experiments
- A narrative hypothesis about a psychological process

For example, when we work towards normative models, we may want our models to start with a description of the problem, maybe uncertainty in perception and a loss function. Or if we work with reductionist models we may want to describe behavior in terms of what neurons do. As such, for the normative models we would like them to be cast in terms of variables describing the problem and for reductionist models we may want them to be cast in terms of variables describing parts of the nervous system.

Bayesian models usually have strong justification in terms of a narrative hypothesis (the brain has evolved to perform certain task close to optimally, apart from noise). But also in Bayesian models it is possible to make arbitrary, poorly justified assumptions just to fit the data. Any inclinations to do so should be avoided if at all possible.

C.6 Absolute goodness of fit

A technical limitation of model comparison is that the best model can still be a bad model. It would be useful to have a sense of good a model is in an absolute sense, i.e. compared to the best possible model. The best possible model of the data is not the data itself, in view of intrinsic stochasticity (noise) in the data. For example, on a given trial, you might predict that a subject chooses option A with probability p and option B with probability $1-p$, but you will not be able to predict the choice on that trial perfectly. The best you can do is to make p match the empirical probability of A responses across trials of this type.

The log likelihood of such an ideal model contributed by one trial of that type will then be $\log p$ when the response is A, i.e. with probability p, and $\log(1-p)$ when the response is B, i.e. with probability $1-p$. Combining the two, the expected contribution of this trial to the log likelihood of this ideal model will be

$$\log \mathscr{L}(\text{ideal model}; \text{data}) = p\log p + (1-p)\log(1-p). \tag{C.17}$$

This is exactly the *negative entropy* of the data. The entropy of a (discrete) probability distribution is a measure of uncertainty: a deterministic mapping has the lowest entropy (0) and a uniform distribution has the highest possible entropy ($\frac{1}{N}$, where N is the number of alternatives).

The relationship between model log likelihood and negative entropy turns out to be true in general:

$$\log \mathscr{L}(\text{any model}; data) \leq -\text{Entropy}(\text{data}). \tag{C.18}$$

The noisier the data, the higher the entropy, and the lower the upper bound on the log likelihood of any model. While Eq. (C.18) is theoretically universal, it is in practice sometimes difficult to estimate the negative entropy of the data, in particular if trial types do not get repeated within the experiment.

C.7 Fitting data from a discrimination task

We conclude this Appendix with two more model fitting examples. We first consider the following hypothetical data from a discrimination task like in Chapter 7, with $\Delta s = 1$. The data can then be fully summarized as in **Table C.2**. The two stimulus values are equally frequent (in this case, occurring on 200 trials each).

True stimulus	Observer responds s_+	Observer responds s_-
s_+	138	62
s_-	90	110

Table C.2: Example data

C.7.1 Simple model

Our goal is again to fit a noise parameter σ. In this experiment, since both s and $\hat{s}$ can only take on two values, the model prediction $p(\hat{s}|s,\sigma)$ consists of just four numbers, namely $\Pr(\hat{s}=s_+|s=s_+,\sigma)$, $\Pr(\hat{s}=s_-|s=s_+,\sigma)$, $\Pr(\hat{s}=s_+|s=s_-,\sigma)$, and $\Pr(\hat{s}=s_-|s=s_-,\sigma)$

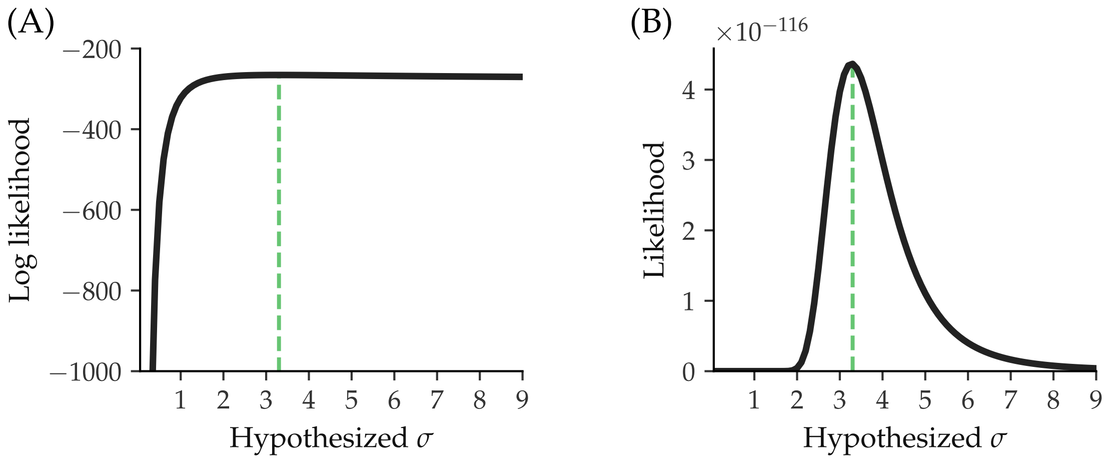


Figure C.5: Log likelihood function and likelihood function over σ obtained from the model in Section C.7.1.

If the observer is optimal and therefore knows that the two stimuli are equally frequent, we use Eqs. (7.20)-(7.21) to evaluate these four probabilities as

$$P(\hat{s}=s_+|s=s_+,\sigma)=\Phi_{\text{standard}}\left(\frac{\Delta s}{2\sigma}\right) \tag{C.19}$$

$$P(\hat{s}=s_-|s=s_+,\sigma)=1-\Phi_{\text{standard}}\left(\frac{\Delta s}{2\sigma}\right) \tag{C.20}$$

$$P(\hat{s}=s_+|s=s_-,\sigma)=1-\Phi_{\text{standard}}\left(\frac{\Delta s}{2\sigma}\right) \tag{C.21}$$

$$P(\hat{s}=s_-|s=s_-,\sigma)=\Phi_{\text{standard}}\left(\frac{\Delta s}{2\sigma}\right) \tag{C.22}$$

The only free parameter is σ. The log likelihood of σ is, from Eq. (C.5),

$$\log\mathscr{L}(\sigma;\text{data})=\sum_{i=1}^{n_{\text{trials}}}\log p(\hat{s}_i|s_i,\sigma). \tag{C.23}$$

The trials can be grouped together into the four combinations of stimulus and response. Then, the sum in Eq. (C.23) becomes a sum of four terms:

$$\log\mathscr{L}(\sigma;\text{data}) = n_{++}\log\Pr(\hat{s}_i=s_+|s_i=s_+,\sigma)+n_{+-}\log\Pr(\hat{s}_i=s_-|s_i=s_+,\sigma) \tag{C.24}$$
$$+n_{-+}\log\Pr(\hat{s}_i=s_+|s_i=s_-,\sigma)+n_{--}\log\Pr(\hat{s}_i=s_-|s_i=s_-,\sigma), \tag{C.25}$$

where we introduced the notation

n_{++}	number of trials where $s=s_+$ and $\hat{s}=s_+$
n_{+-}	number of trials where $s=s_+$ and $\hat{s}=s_-$
n_{-+}	number of trials where $s=s_-$ and $\hat{s}=s_+$
n_{--}	number of trials where $s=s_-$ and $\hat{s}=s_+$.

Both the log likelihood and the likelihood are plotted as a function of σ in **Fig. C.5**. The likelihood was obtained by exponentiating the log likelihood.

To find the maximum-likelihood estimate of σ, the analytical method (Method 1) is feasible here. We take the derivative of Eq. (C.25), set it to 0, and solve for σ. The answer is

$$\hat{\sigma} = \frac{\Delta}{2\Phi^{-1}_{\text{standard}}\left(\frac{n_{\text{correct}}}{n_{\text{correct}}+n_{\text{incorrect}}}\right)}, \tag{C.26}$$

where n_{correct} and $n_{\text{incorrect}}$ are the number of correct and incorrect trials, respectively. This expression depends on the fact that in this particular model, $\Pr(\hat{s}_i = s_+|s_i = s_+,\sigma)$ and $\Pr(\hat{s}_i = s_-|s_i = s_-,\sigma)$ are equal. In the example given, Eq. (C.26) returns $\hat{\sigma} = 1.64$. The corresponding maximum log likelihood is $\log\mathscr{L}^* = -265.6$.

True stimulus	Model responds s_+	Model responds s_-
s_+	124	76
s_-	76	124

Table C.3: Checking the simple model.

Like in Section C.5.1, it is important to check the model. Since the probability correct is 0.62 according to the model, the predicted numbers of responses are as in **Table C.3**. Thus, this model does not account for the unequal proportions correct between s_+ and s_- seen in the data (**Table C.2**). Of course, that inequality could be due to chance. Therefore, it is worth comparing to a more flexible model.

C.7.2 A better model?

In Chapter 7, we also considered a more flexible model, in which the observer's prior is not equal to 0.5. Then, the model has two parameters, σ and that prior. Since our data in this very simple experiment also amount to two numbers (proportion correct when the stimulus is s_+ and when the stimulus is s_-, we would be fitting two data points with two parameters. Without even fitting those parameters, we know that this can be done perfectly (see Problem). That does not mean that every trial can be predicted perfectly, just that the predicted *probabilities* of the responses exactly match the empirical probabilities.

Per Eq. (C.25), the log likelihood again consists of four terms. The maximum log likelihood of this two-parameter model is (see Problem)

$$\log\mathscr{L}^* = n_{++}\log\frac{n_{++}}{n_{++}+n_{+-}} + n_{+-}\log\frac{n_{+-}}{n_{++}+n_{+-}} + n_{-+}\log\frac{n_{-+}}{n_{-+}+n_{--}} + n_{--}\log\frac{n_{--}}{n_{-+}+n_{--}}. \tag{C.27}$$

This evaluates to -261.4, which is 4.2 higher than in the simple model of Section C.7.1. This difference is large enough that this model wins in both AIC and BIC. Nevertheless, this is not a particularly insightful model. The fact that the winning model is equivalent to a mere description of each condition is a sign that the data in this experiment are too simple to draw the relevant conclusions.

C.8 Fitting data from a classification task

In Chapter 8, we considered binary classification tasks. Data from such a task are much richer than the previously considered discrimination task. Example data are shown in **Table C.4**.

The model in this task is a probability distribution $p(\hat{C}|s,\theta)$, where θ represents all parameters. Thus, the log likelihood function takes the form

$$\log\mathscr{L}(\theta;\text{data}) = \sum_{i=1}^{n_{\text{trials}}} \log p(\hat{C}_i|s_i,\theta). \tag{C.28}$$

Trial number	Stimulus s	Subject response, $\hat{s}$
1	-4.3	-1
2	-1.0	1
3	2.1	1
4	0.7	-1
5	4.5	1
6	-3.5	-1
...	...	...

Table C.4: Example data from the classification task.

We work this out in Problem C.4. In some cases, the large sum over trials can be split up in a sum over unique stimulus-response combinations. Then, per combination, we have to multiply the summand by the number of instances of each combination, similar to Eq. (C.25).

Box C.1 — Good experimental design for Bayesian modeling. To allow you to successfully build a Bayesian model of your behavioral data, the first requirement is to design an experiment well. There are well-known general guidelines for this, and in addition some that are specific to Bayesian models. In general, one would like to control as many of the parameters that are not of interest to the scientific question. For example, to study how humans perform discrimination, we want to present stimuli for a short time (a few tens of milliseconds), to avoid complications associated with eye movements, the time course of attention, and the integration of information over time, all of which can effect the quality of encoding (i.e. the standard deviation of the noise distribution) in a potentially complex way. Reaction time experiments are typically more complex to model than accuracy experiments with short presentation times. Therefore, if your scientific question allows to do an accuracy version of the same experiment, it will likely save you work and computation time during modeling.

Similarly, we want to keep attributes of the stimuli that are not of interest as much the same between stimuli. Specifically, make sure to carefully control the reliability/precision/noise level of the stimuli. For example, if you arrange multiple items in a display, arranging them on a circle around the fixation point instead of in a rectangular grid ensures that the eccentricity (distance from the fixation point) is the same and therefore encoding precision is at least approximately the same, allowing to model reliability with a single parameter.

Furthermore, one needs to be well aware of domain-specific effects that can influence performance in the task. For example, when two stimuli are brought close together, an effect known as crowding can occur, in which the internal representations of both stimuli influence each other. If crowding is not of primary interest, it is best to minimize it by placing the stimuli sufficiently far apart from each other. Specific to Bayesian modeling, it is often useful to use stimuli whose feature of interest is one-dimensional or at most two-dimensional. For example, when studying cue combination, it is easier to model a flash and a beep presented on a horizontal line, than to model the integration of the auditory and visual information in speech perception. In the perceptual experiments discussed in this book, we use stimuli that are as simple as possible: they have only a single relevant dimension, for example orientation. Stimuli like letters, line drawings, images of objects, or natural scenes are much more difficult to cast into a model because they have many features and in some cases it is even clear what the relevant features (perceptual building blocks) are. Moreover, large number of features translates into a large number of dimensions, and noise models in high numbers of dimensions have even higher numbers of free parameters. This is not to say that studying complex stimuli not interesting, on the contrary. In a way, we are merely pointing out the limitations of Bayesian modeling.

Finally, we recommend that anyone interested in building a Bayesian model of their task write out the model and simulate it before even starting to collect data. For Bayesian models, since they are based on principles of optimality, this is always possible. This process will usually highlight potential problems in the experimental design. ■

C.9 Summary

In this Appendix, we have focused on the methodology of fitting models to data and of finding out which models work better for a given set of data. We have learned:

- A model is characterized by a probability distribution over the behavioral response given the stimuli and parameters. Modelers need to estimate the parameters.
- In maximum-likelihood estimation, we are asking which set of parameters makes the observations most likely.
- We have seen how the simple models from the previous chapters can be fit to actual behavior.
- If we have few parameters we can do model fitting by grid search – we can basically try any meaningful combination of parameters.
- If the function we have to optimize is well shaped, e.g. is convex, then we can generally use standard optimizers, e.g. those based on simplex methods, to solve our problems.
- If the function is less well-behaved we need to resort to more advanced strategies. Popular nonconvex optimization strategies will generally outperform the simple ideas we may come up with ourselves.
- When models have different numbers of parameters a simple comparison of quality of fit is pointless.
- Instead we need to correct for the number of free parameters, e.g. using AIC, BIC, or cross-validation. Care needs to be taken, as all these techniques have their own caveats.

C.10 Suggested readings

- Luigi Acerbi and Wei Ji Ma. “Practical Bayesian optimization for model fitting with Bayesian adaptive direct search”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, pages 1834–1844
- Sylvain Arlot and Alain Celisse. “A survey of cross-validation procedures for model selection”. In: *Statistics surveys* 4 (2010), pages 40–79
- David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003
- In Jae Myung. “Tutorial on maximum likelihood estimation”. In: *Journal of mathematical Psychology* 47.1 (2003), pages 90–100
- Lionel Rigoux et al. “Bayesian model selection for group studies—revisited”. In: *Neuroimage* 84 (2014), pages 971–985
- Scott I Vrieze. “Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC).”. In: *Psychological methods* 17.2 (2012), page 228
- Robert C Wilson and Anne GE Collins. “Ten simple rules for the computational modeling of behavioral data”. In: *Elife* 8 (2019), e49547

C.11 Problems

Problem C.1 Derive Eq. (C.27).

Problem C.2 Eqs. (7.22)-(7.23) specified the response probabilities in a Bayesian model of discrimination:

$$\Pr(\hat{s}=s_+|s=s_+)=\Phi_{\text{standard}}\left(\frac{\Delta s}{2\sigma}+\frac{\sigma}{\Delta s}\log\frac{p_+}{1-p_+}\right) \tag{C.29}$$

$$\Pr(\hat{s}=s_+|s=s_-)=\Phi_{\text{standard}}\left(-\frac{\Delta s}{2\sigma}+\frac{\sigma}{\Delta s}\log\frac{p_+}{1-p_+}\right), \tag{C.30}$$

where $p_+ = \Pr(s=s_+)$ is the prior probability of s_+. As the experimenter, you know the value of Δs. The subject's responses give you the two proportions on the left-hand sides. Solve for σ and p_+. This shows that using these two parameters, the responses proportions in this task can always be exactly matched.

Problem C.3 This problem is about the two models in Section C.7.

a) Calculate the AIC values of both models. Which model is better according to AIC?
b) Calculate the AIC values of both models. Which model is better according to BIC?
c) Calculate the 10-fold cross-validated log likelihood of both models. Which model is better according to the cross-validated log likelihood?
d) Perform a likelihood ratio test. Is the more complex model significantly better than the simpler model?

Problem C.4 We will fit a simple model of binary classification. Download appendixC4_psychometric_s_presented.csv (which contains $s_{\text{presented}}$ and appendixC4_psychometric_r.csv (which contains r) from https://osf.io/84kpb/. These two variables contains a synthetic data set of a subject performing 500 trials of a left/right classification task. The stimulus took values from -5 to 5 in steps of 1; we will denote the values by s_j, with $j=1,2,\ldots,11$. The vector $s_{\text{presented}}$ contains the stimulus values presented on the 500 trials. On each trial, the subject responded whether the stimulus was positive (to the right; $\hat{C}=1$) or negative (to the left; $\hat{C}=-1$). The vector r contains the subject responses.

a) Plot the psychometric curve, i.e. the proportion "right" responses as a function of s, without a connecting line.
b) We will now fit a special case of the model from Chapter 8:

$$p(\hat{C}=1|s,\sigma)=\Phi_{\text{standard}}\left(\frac{s}{\sigma}\right). \tag{C.31}$$

Explain why the log likelihood of σ takes the form

$$\log\mathscr{L}(\sigma;\text{data})=\sum_{j=1}^{11}n_{j+}\log\Pr(\hat{C}=1|s=s_j,\sigma)+\sum_{j=1}^{11}n_{j-}\log\Pr(\hat{C}=-1|s=s_j,\sigma), \tag{C.32}$$

where n_{j+} and n_{j-} are the numbers of trials on which the stimulus was s_j and the response was $\hat{C}=1$ or $\hat{C}=-1$, respectively.

c) Plot the log likelihood function as a function of σ. Use a grid for σ from 0.1 to 5 in 1000 steps.
d) Plot the likelihood function (no log) over σ using the same grid.
e) Explain why the likelihood values are extremely small.
f) Find the maximum-likelihood estimate of σ on the grid.
g) Instead of using a grid, find the maximum-likelihood estimate of σ using a built-in numerical optimization algorithm.
h) Plot the best model fit as a line in the plot obtained in part (a). Use the ML estimate of σ from part (g), or if you did not complete that part, from part (f).

Problem C.5 This is a follow-up on Problem C.4 and uses the same data file psychometric.mat. We will compare the simple model from Problem C.4 with a more general model in which the observer sometimes guesses randomly. The latter model takes the form

$$p(\hat{C}=1|s,\sigma,\lambda)=(1-\lambda)\Phi_{\text{standard}}\left(\frac{s}{\sigma}\right)+0.5\lambda, \tag{C.33}$$

where λ is an unknown guessing rate. This model has more flexibility due to the extra parameter.

a) Plot the log likelihood landscape of this model as a heat map. For σ, use the same grid as in Problem C.4. For λ, use a grid from 0 to 0.3 in 1000 steps.
b) Find the maximum-likelihood estimates of σ and λ on their respective grids.
c) Now forget about the grids and instead find the maximum-likelihood estimates of σ and λ using a built-in optimization algorithm.
d) Plot the psychometric curve (open circles) along with the best fits of both the simple model and the more complex model (solid lines in different colors). Use the ML estimates from part (c), or if you did not complete that part, from part (b).
e) Calculate the AIC for both models. Calculate the AIC difference. Draw a conclusion in words.
f) Calculate the BIC for both models. Calculate the BIC difference. Draw a conclusion in words.
g) Calculate the 10-fold cross-validated log likelihood for both models. Calculate the cross-validated log likelihood difference. Draw a conclusion in words.

Problem C.6 In this problem, we fit and compare models on estimation data. Download appendixC6_estimation_s_hat.csv (which contains $\hat{s}$ and appendixC6_estimation_s.csv (which contains s) from https://osf.io/84kpb/. s contains stimuli presented to a subject across 500 trials, and $\hat{s}$ the subject's estimates of these stimuli. We assume that the observer's measurement follows a Gaussian distribution with mean equal to the true stimulus and unknown standard deviation σ. We also assume that all trials are independent.

a) Plot the data in a scatter plot of estimate against stimulus, using black dots. Draw a dashed black line to indicate the diagonal. Choose your axis ranges suitably. Label the axes.
b) By eye, would you say that the observer is doing maximum-likelihood estimation or is taking into account a prior? Why?
c) Model 1 is that the observer performs maximum-likelihood estimation. Under this model, write down an equation for the log likelihood function over σ in terms of the stimuli $s_1,\ldots,s_n$ and the estimates $\hat{s}_1,\ldots,\hat{s}_n$. Simplify as much as possible.
d) Plot the log likelihood function of σ on a grid from 0.1 to 10 in steps of 0.02.
e) On this grid, what is the maximum-likelihood estimate of σ?
f) As an alternative to the grid search, use a built-in optimization algorithm to find the maximum-likelihood estimate of σ.
g) Model 2 is that the observer performs MAP estimation using a Gaussian prior with mean 0 and unknown standard deviation σ_{prior}. Under this model, write down an equation for the log likelihood function of the combination $(\sigma,\sigma_{\text{prior}})$. Simplify as much as possible.
h) Using grids from 0.1 to 10 in steps of 0.02 for both σ and σ_{prior}, plot the log likelihood function as a heat map. Add a color legend. Label the axes; make sure to check which axis corresponds to which variable.
i) On this grid, what are the maximum-likelihood estimates of σ and σ_{prior}?
j) Use a built-in optimization algorithm to find the maximum-likelihood estimates of σ and σ_{prior}.
k) Which model wins according to AIC?
l) Which model wins according to BIC?

Problem C.7 In this problem we fit and compare models on data from a real estimation experiment. The experiment was a sound localization experiment in humans conducted in the lab of Wei Ji Ma. On each trial, a sound was presented at one of five locations on a horizontal line: -6, -3, 0, 3, 6 (arbitrary units). The sound sources were hidden and the subject did not know that there were only five locations. The subject reported where they heard the sound coming from; they chose from 21 discrete locations: $-10, -9, -8, \ldots, 8, 9, 10$.

a) **The data.** Download appendixC7_localization.csv from https://osf.io/84kpb/. Each row corresponds to a trial (400 trials in total). The first column indicates the true location, the second column the reported location. Separately for each of the five true locations, calculate the proportion of subject responses at each response location. Plot as five solid lines in a single plot, color-coded to distinguish the five true locations.

b) We first consider a "null model", Model 0 (0 parameters). According to this model, the observer chooses a random location (out of the 21 possibilities), with equal probabilities. Calculate the log likelihood, AIC, and BIC for Model 0. You do not even need the subject responses for this. Explain why not.

We now introduce two more meaningful models:

- Model 1 (1 parameter): The observer performs maximum-likelihood estimation of location, i.e. has a flat prior.
- Model 2 (2 parameters): The observer performs maximum-a-posteriori estimation of location, with a prior that is Gaussian with mean 0 and unknown standard deviation σ_s.

In both models, we make the usual assumption that the observer makes a noisy measurement of sound location and that this measurement follows a Gaussian distribution with mean equal to the true location and unknown standard deviation σ. Parts (c) through (g) are about Models 1 and 2.

c) **Model predictions I.** Ignore for the moment that the observer only makes discrete responses, and instead imagine that they report a continuous estimate of location, $\hat{s}$. For Model 1, write down the equation for the probability density function over $\hat{s}$ when the true stimulus is in location s. Repeat for Model 2.

d) **Model predictions II.** We have to deal with a complication: the response is not continuous, since there are only 21 possible response locations. Therefore, we assume that the observer first computes the continuous estimate $\hat{s}$ and then reports the closest possible response location. Thus, the probability of reporting a particular location is the probability that $\hat{s}$ falls in a bin of size 1 around that location. Exception: for the two extreme response locations (-10 and 10), the bin extends infinitely on one side. Suppose that the true location s is -6, and that $\sigma = 3$ and $\sigma_s = 10$. For both Model 1 and Model 2, compute a vector of 21 numbers that gives the probability of reporting each of the 21 locations as predicted by the model. Plot both model predictions in the same plot with different colors.

e) **Model predictions III.** Repeat part (d), except for the plotting, for each of the five true locations and for each combination of parameter values. Choose the possible values of σ and σ_s to range from 0.1 to 15 in steps of 0.1. For Model 1, save the results in a matrix of size 21 (responses) × 5 (true stimulus locations) × 150 (values of σ), and for Model 2, in a matrix of size $21 \times 5 \times 150 \times 150$. To avoid numerical problems, first set probability values of zero equal to the smallest non-zero value in the matrix[2]. After that, make sure that, separately for every true location and parameter combination, the response probabilities again sum to 1 across the 21 response locations.

f) **Model fitting I.** Assume that trials are independent of each other. Use the "lookup tables" of model predictions from part (e) to calculate the log likelihood of every parameter combination, separately for Models 1 and 2. This should give you a vector of length 150 for Model 1 and a

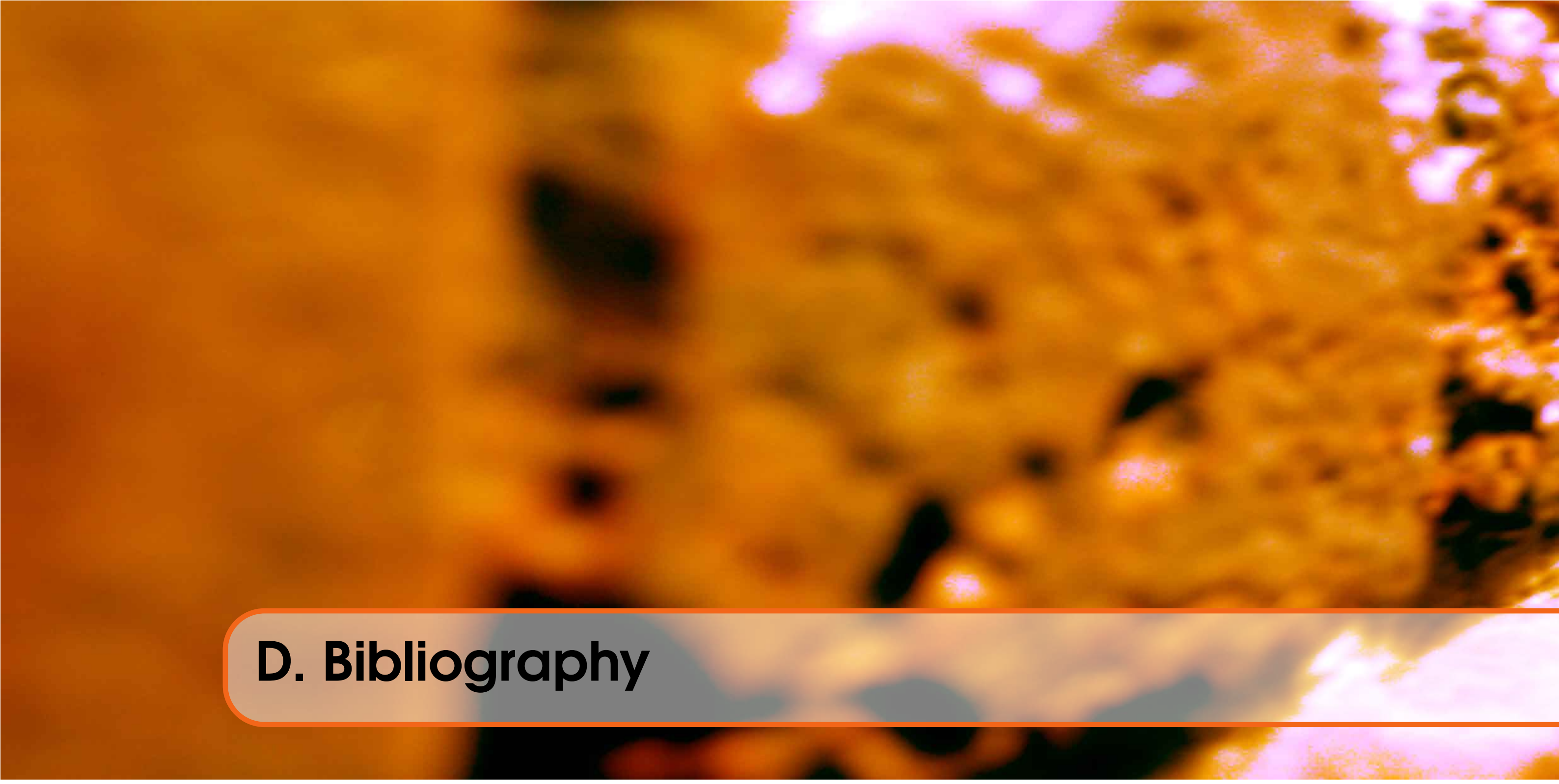
[2] This is an easy solution but not the best one. A better solution is to use Inverse Binomial Sampling (cite Acerbi Ma) throughout.

matrix of size 150 × 150 for Model 2. Plot the former using a line plot and the latter using a heat map.

g) **Model fitting II.** For Models 1 and 2, find the maximum-likelihood estimates of the parameter(s) on their grids.

h) **Model comparison.** Report for Models 1 and 2 the maximum log likelihood, AIC, and BIC. Combining with part (b), which model wins according to AIC? Which model wins according to BIC? State your overall conclusion about the subject's behavior in a sentence in a way you would do in a paper.

i) **Model checking.** Separately for each model, add the model fits to the five curves in part (a). Use dashed lines with colors corresponding to those of the data. Plot each model in a separate plot, so you should get three plots, each containing five solid lines (data) and five corresponding dashed lines (model fits). For Models 1 and 2, use the maximum-likelihood estimates from part (g) and the lookup tables from part (e).



D. Bibliography

Books

[And13] John R Anderson. *The adaptive character of thought*. Psychology Press, 2013 (cited on page 287).

[CO+08] Nick Chater, Mike Oaksford, et al. *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford University Press, USA, 2008 (cited on pages 17, 283, 287).

[Coo87] Gloria Cooper. *Red Tape Holds Up New Bridge, and More Flubs from the Nation's Press*. TarcherPerigee, 1987 (cited on page 37).

[Ges13] George A Gescheider. *Psychophysics: the fundamentals*. Psychology Press, 2013 (cited on page 156).

[GS+66] David Marvin Green, John A Swets, et al. *Signal detection theory and psychophysics*. Volume 1. Wiley New York, 1966 (cited on page 156).

[Jay03] Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003 (cited on page 101).

[Kah11] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011 (cited on pages 283, 287).

[Kah+82] Daniel Kahneman et al. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press, 1982 (cited on page 283).

[Lap12] Pierre-Simon Laplace. *Pierre-Simon Laplace philosophical essay on probabilities: translated from the fifth french edition of 1825 with notes by the translator*. Volume 13. Springer Science & Business Media, 2012 (cited on page 37).

[MM03] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003 (cited on pages 210, 335).

[McG11] Sharon Bertsch McGrayne. *The theory that would not die*. Yale University Press, 2011 (cited on page 37).

[Smi01] A Mark Smith. *Alhacen's Theory of Visual Perception: A Critical Edition, with English Translation and Commentary, of the First Three Books of Alhacen's De Aspectibus, the Medieval Latin Version of Ibn Al-Haytham's Kitab Al-Manazir*. Volume 1. American Philosophical Society, 2001 (cited on page 37).

[TKL11] Julia Trommershauser, Konrad Kording, and Michael S Landy. *Sensory cue integration*. Oxford University Press, 2011 (cited on page 118).

[Wic01] Thomas D Wickens. *Elementary signal detection theory*. Oxford university press, 2001 (cited on page 156).

Articles

[AVW14] Luigi Acerbi, Sethu Vijayakumar, and Daniel M Wolpert. "On the origins of suboptimality in human probabilistic inference". In: *PLoS computational biology* 10.6 (2014), e1003661 (cited on page 285).

[AS10] Daniel E Acuña and Paul Schrater. "Structure learning in human sequential decision-making". In: *PLoS computational biology* 6.12 (2010), e1001003 (cited on page 138).

[AM07] Ryan Prescott Adams and David JC MacKay. "Bayesian online changepoint detection". In: *arXiv preprint arXiv:0710.3742* (2007) (cited on page 237).

[Ada07] Wendy J Adams. "A common light-prior for visual search, shape, and reflectance judgments". In: *Journal of Vision* 7.11 (2007), pages 11–11 (cited on page 287).

[AGE04] Wendy J Adams, Erich W Graf, and Marc O Ernst. "Experience can change the'light-from-above'prior". In: *Nature neuroscience* 7.10 (2004), pages 1057–1058 (cited on page 82).

[AB04] David Alais and David Burr. "The ventriloquist effect results from near-optimal bimodal integration". In: *Current biology* 14.3 (2004), pages 257–262 (cited on page 118).

[And91] John R Anderson. "The adaptive nature of human categorization." In: *Psychological review* 98.3 (1991), page 409 (cited on pages 17, 283, 287).

[Ang07] J Yu Angela. "Adaptive behavior: Humans act as Bayesian learners". In: *Current Biology* 17.22 (2007), R977–R980 (cited on page 237).

[AC10] Sylvain Arlot and Alain Celisse. "A survey of cross-validation procedures for model selection". In: *Statistics surveys* 4 (2010), pages 40–79 (cited on page 335).

[Bah+10] Bahador Bahrami et al. "Optimally interacting minds". In: *Science* 329.5995 (2010), pages 1081–1085 (cited on pages 117, 118).

[Bay58] FRS Bayes. "An essay towards solving a problem in the doctrine of chances". In: *Biometrika* 45.3-4 (1958), pages 296–315 (cited on page 36).

[BSG96] Robert J van Beers, Anne C Sittig, and Jan J van der Gon Denier. "How humans combine simultaneous proprioceptive and visual position information". In: *Experimental brain research* 111.2 (1996), pages 253–261 (cited on page 118).

[Bej+11] Vikranth Rao Bejjanki et al. "Cue integration in categorical tasks: Insights from audio-visual speech perception". In: *PloS one* 6.5 (2011), e19812 (cited on page 118).

[Bis06] Christopher M Bishop. "Pattern recognition". In: *Machine learning* 128.9 (2006) (cited on page 286).

[Bon+15] Kathryn Bonnen et al. "Continuous psychophysics: Target-tracking to measure visual sensitivity". In: *Journal of Vision* 15.3 (2015), pages 14–14 (cited on page 237).

[BD12] Jeffrey S Bowers and Colin J Davis. "Bayesian just-so stories in psychology and neuroscience." In: *Psychological bulletin* 138.3 (2012), page 389 (cited on pages 282, 287).

[BF97] David H Brainard and William T Freeman. "Bayesian color constancy". In: *JOSA A* 14.7 (1997), pages 1393–1411 (cited on page 189).

[BK07] Anne-Marie Brouwer and David C Knill. "The role of memory in visually guided reaching". In: *Journal of vision* 7.5 (2007), pages 6–6 (cited on page 118).

[BB93] Peter Brugger and Susanne Brugger. "The Easter bunny in October: Is it disguised as a duck?" In: *Perceptual and motor skills* 76.2 (1993), pages 577–578 (cited on page 36).

[Bul96] Heinrich H Bulthoff. "Bayesian decision theory and psychophysics". In: *Perception as Bayesian inference* 123 (1996), page 1 (cited on page 118).

[Cha20] Gar Ming Chan. "Bayes' theorem, COVID19, and screening tests". In: *The American Journal of Emergency Medicine* 38.10 (2020), pages 2011–2013 (cited on page 54).

[Cha+11] Nick Chater et al. "The imaginary fundamentalists: The unshocking truth about Bayesian cognitive science". In: *Behavioral and Brain Sciences* 34.4 (2011), page 194 (cited on page 287).

[CRM14] Stephanie Y Chen, Brian H Ross, and Gregory L Murphy. "Implicit and explicit processes in category-based induction: Is induction best when we don't think?" In: *Journal of Experimental Psychology: General* 143.1 (2014), page 227 (cited on pages 283, 287).

[Che97] Patricia W Cheng. "From covariation to causation: A causal power theory." In: *Psychological review* 104.2 (1997), page 367 (cited on page 138).

[CRG15] Anna Coenen, Bob Rehder, and Todd M Gureckis. "Strategies to intervene on causal systems are adaptively selected". In: *Cognitive psychology* 79 (2015), pages 102–133 (cited on page 138).

[Den+18] Rachel N Denison et al. "Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence". In: *Proceedings of the National Academy of Sciences* 115.43 (2018), pages 11090–11095 (cited on page 174).

[Eck11] Miguel P Eckstein. "Visual search: A retrospective". In: *Journal of vision* 11.5 (2011), pages 14–14 (cited on page 224).

[EB02] Marc O Ernst and Martin S Banks. "Humans integrate visual and haptic information in a statistically optimal fashion". In: *Nature* 415.6870 (2002), pages 429–433 (cited on page 118).

[EBP85] J St BT Evans, PG Brooks, and P Pollard. "Prior beliefs and statistical inference". In: *British Journal of Psychology* 76.4 (1985), pages 469–477 (cited on pages 17, 283).

[Fel01] Jacob Feldman. "Bayesian contour integration". In: *Perception & Psychophysics* 63 (2001), pages 1171–1182 (cited on page 210).

[Fen11] Norman Fenton. "Improve statistics in court". In: *Nature* 479.7371 (2011), pages 36–37 (cited on page 54).

[Fet+12] Christopher R Fetsch et al. "Neural correlates of reliability-based cue weighting during multisensory integration". In: *Nature neuroscience* 15.1 (2012), pages 146–154 (cited on page 277).

[FP11] Brian J Fischer and José Luis Peña. "Owl's behavior and neural representation predicted by Bayesian inference". In: *Nature neuroscience* 14.8 (2011), pages 1061–1066 (cited on page 82).

[GD03] Wilson S Geisler and Randy L Diehl. “A Bayesian approach to the evolution of perceptual and cognitive systems”. In: *Cognitive Science* 27.3 (2003), pages 379–402 (cited on page 36).

[GP09] Wilson S Geisler and Jeffrey S Perry. “Contour statistics in natural images: Grouping across occlusions”. In: *Visual neuroscience* 26.1 (2009), pages 109–121 (cited on pages 54, 210).

[Gig96] Gerd Gigerenzer. “On narrow norms and vague heuristics: A reply to Kahneman and Tversky.” In: (1996) (cited on page 287).

[GP12] Daniel Goldreich and Mary A Peterson. “A Bayesian observer replicates convexity context effects in figure–ground perception”. In: *Seeing and Perceiving* 25.3-4 (2012), pages 365–395 (cited on page 210).

[GT13] Daniel Goldreich and Jonathan Tong. “Prediction, postdiction, and perceptual length contraction: a Bayesian low-speed prior captures the cutaneous rabbit and related illusions”. In: *Frontiers in psychology* 4 (2013), page 221 (cited on pages 82, 237).

[Gop+04] Alison Gopnik et al. “A theory of causal learning in children: causal maps and Bayes nets.” In: *Psychological review* 111.1 (2004), page 3 (cited on pages 17, 138, 283).

[GT06] Thomas L Griffiths and Joshua B Tenenbaum. “Optimal predictions in everyday cognition”. In: *Psychological science* 17.9 (2006), pages 767–773 (cited on pages 54, 283).

[GT09] Thomas L Griffiths and Joshua B Tenenbaum. “Theory-based causal induction.” In: *Psychological review* 116.4 (2009), page 661 (cited on page 174).

[HMC21] Michael J Hautus, Neil A Macmillan, and C Douglas Creelman. “Detection Theory: A User’s Guide”. In: (2021) (cited on page 156).

[HS25] HV Helmholtz and JPC Southall. “Treatise on physiological optics. III. The perceptions of vision.” In: (1925) (cited on page 37).

[Jac99] Robert A Jacobs. “Optimal integration of texture and motion cues to depth”. In: *Vision research* 39.21 (1999), pages 3621–3629 (cited on page 118).

[JM06] Mehrdad Jazayeri and J Anthony Movshon. “Optimal representation of sensory information by neural populations”. In: *Nature neuroscience* 9.5 (2006), pages 690–696 (cited on page 277).

[JS10] Mehrdad Jazayeri and Michael N Shadlen. “Temporal context calibrates interval timing”. In: *Nature neuroscience* 13.8 (2010), pages 1020–1026 (cited on page 82).

[JL11] Matt Jones and Bradley C Love. “Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition”. In: *Behavioral and brain sciences* 34.4 (2011), page 169 (cited on pages 282, 287).

[KPT07] Charles Kemp, Andrew Perfors, and Joshua B Tenenbaum. “Learning overhypotheses with hierarchical Bayesian models”. In: *Developmental science* 10.3 (2007), pages 307–321 (cited on pages 138, 288).

[KMY04] Daniel Kersten, Pascal Mamassian, and Alan Yuille. “Object perception as Bayesian inference”. In: *Annu. Rev. Psychol.* 55 (2004), pages 271–304 (cited on pages 174, 189).

[Kni03] David C Knill. “Mixture models and the probabilistic structure of depth cues”. In: *Vision research* 43.7 (2003), pages 831–854 (cited on page 189).

[KS03] David C Knill and Jeffrey A Saunders. “Do humans optimally integrate stereo and texture information for judgments of surface slant?” In: *Vision research* 43.24 (2003), pages 2539–2558 (cited on page 118).

[KTS07] Konrad P Kording, Joshua B Tenenbaum, and Reza Shadmehr. “The dynamics of memory as a consequence of optimal adaptation to a changing body”. In: *Nature neuroscience* 10.6 (2007), pages 779–786 (cited on page 237).

[KW04a] Konrad P Körding and Daniel M Wolpert. “Bayesian integration in sensorimotor learning”. In: *Nature* 427.6971 (2004), pages 244–247 (cited on page 82).

[Kör+04] Konrad P Körding et al. “A neuroeconomics approach to inferring utility functions in sensorimotor control”. In: *PLoS biology* 2.10 (2004), e330 (cited on page 255).

[Kör+07] Konrad P Körding et al. “Causal inference in multisensory perception”. In: *PLoS one* 2.9 (2007), e943 (cited on page 210).

[KW04b] Konrad Paul Körding and Daniel M Wolpert. “The loss function of sensorimotor learning”. In: *Proceedings of the National Academy of Sciences* 101.26 (2004), pages 9839–9842 (cited on page 255).

[LHC15] Rosa Lafer-Sousa, Katherine L Hermann, and Bevil R Conway. “Striking individual differences in color perception uncovered by ‘the dress’ photograph”. In: *Current Biology* 25.13 (2015), R545–R546 (cited on page 189).

[Lap86] Pierre Simon Laplace. “Memoir on the probability of the causes of events”. In: *Statistical science* 1.3 (1986), pages 364–378 (cited on page 138).

[LG20] Falk Lieder and Thomas L Griffiths. “Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources”. In: *Behavioral and Brain Sciences* 43 (2020) (cited on page 288).

[LKK95] Zili Liu, David C Knill, and Daniel Kersten. “Object classification for human and ideal observers”. In: *Vision research* 35.4 (1995), pages 549–568 (cited on page 174).

[Ma12] Wei Ji Ma. “Organizing probabilistic models of perception”. In: *Trends in cognitive sciences* 16.10 (2012), pages 511–518 (cited on page 288).

[MJ14] Wei Ji Ma and Mehrdad Jazayeri. “Neural coding of uncertainty and probability”. In: *Annual review of neuroscience* 37 (2014), pages 205–220 (cited on pages 285, 288).

[Ma+06] Wei Ji Ma et al. “Bayesian inference with probabilistic population codes”. In: *Nature neuroscience* 9.11 (2006), pages 1432–1438 (cited on page 277).

[Ma+09] Wei Ji Ma et al. “Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space”. In: *PloS one* 4.3 (2009), e4638 (cited on page 118).

[Ma+11] Wei Ji Ma et al. “Behavior and neural basis of near-optimal visual search”. In: *Nature neuroscience* 14.6 (2011), pages 783–790 (cited on page 224).

[Mad+16] Tamas J Madarasz et al. “Evaluation of ambiguous associations in the amygdala by learning the structure of the environment”. In: *Nature neuroscience* 19.7 (2016), pages 965–972 (cited on page 138).

[MM09] Laurence T Maloney and Pascal Mamassian. “Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer”. In: *Visual neuroscience* 26.1 (2009), pages 147–155 (cited on pages 285, 288).

[MTL07] Laurence T Maloney, Julia Trommershäuser, and Michael S Landy. "Questions without words: A comparison between decision making under risk and movement planning under risk." In: *Integrated models of cognitive systems* 29 (2007), pages 297–313 (cited on page 255).

[MVM12] Helga Mazyar, Ronald Van den Berg, and Wei Ji Ma. "Does precision decrease with set size?" In: *Journal of vision* 12.6 (2012), pages 10–10 (cited on pages 223, 224).

[MM76] Harry McGurk and John MacDonald. "Hearing lips and seeing voices". In: *Nature* 264.5588 (1976), pages 746–748 (cited on page 118).

[MHK01] Barbara Mellers, Ralph Hertwig, and Daniel Kahneman. "Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration". In: *Psychological Science* 12.4 (2001), pages 269–275 (cited on page 288).

[MW08] Shane T Mueller and Christoph T Weidemann. "Decision noise: An explanation for observed violations of signal detection theory". In: *Psychonomic bulletin & review* 15.3 (2008), pages 465–494 (cited on page 101).

[MCR12] Gregory L Murphy, Stephanie Y Chen, and Brian H Ross. "Reasoning with uncertain categories". In: *Thinking & Reasoning* 18.1 (2012), pages 81–117 (cited on page 175).

[Myu03] In Jae Myung. "Tutorial on maximum likelihood estimation". In: *Journal of mathematical Psychology* 47.1 (2003), pages 90–100 (cited on page 335).

[NG05] Jiri Najemnik and Wilson S Geisler. "Optimal eye movement strategies in visual search". In: *Nature* 434.7031 (2005), pages 387–391 (cited on page 224).

[NP10] Daniel J Navarro and Amy F Perfors. "Similarity, feature discovery, and the size principle". In: *Acta Psychologica* 133.3 (2010), pages 256–268 (cited on pages 283, 288).

[NPG08] Mohamed AF Noor, Robin S Parnell, and Bruce S Grant. "A reversible color polyphenism in American peppered moth (Biston betularia cognataria) caterpillars". In: *PloS one* 3.9 (2008), e3142 (cited on page 36).

[Nor+19] Elyse H Norton et al. "Human online adaptation to changes in prior probability". In: *PLoS computational biology* 15.7 (2019), e1006681 (cited on page 237).

[OM17] A Emin Orhan and Wei Ji Ma. "Efficient probabilistic inference in generic neural networks trained with non-probabilistic feedback". In: *Nature communications* 8.1 (2017), pages 1–14 (cited on page 277).

[PVP00] John Palmer, Preeti Verghese, and Misha Pavel. "The psychophysics of visual search". In: *Vision research* 40.10-12 (2000), pages 1227–1268 (cited on pages 223, 224).

[Qam+13] Ahmad T Qamar et al. "Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization". In: *Proceedings of the National Academy of Sciences* 110.50 (2013), pages 20332–20337 (cited on page 175).

[RD18] Dobromir Rahnev and Rachel N Denison. "Suboptimality in perceptual decision making". In: *Behavioral and Brain Sciences* 41 (2018) (cited on page 288).

[Rig+14] Lionel Rigoux et al. "Bayesian model selection for group studies—revisited". In: *Neuroimage* 84 (2014), pages 971–985 (cited on page 335).

[Ros01] Ruth Rosenholtz. "Visual search for orientation among heterogeneous distractors: Experimental results and implications for signal-detection theory models of search." In: *Journal of Experimental Psychology: Human Perception and Performance* 27.4 (2001), page 985 (cited on page 224).

[SAN96] Jenny R Saffran, Richard N Aslin, and Elissa L Newport. “Statistical learning by 8-month-old infants”. In: *Science* 274.5294 (1996), pages 1926–1928 (cited on page 138).

[San96] Terence David Sanger. “Probability density estimation for the interpretation of neural population codes”. In: *Journal of neurophysiology* 76.4 (1996), pages 2790–2793 (cited on page 277).

[STA07] Yoshiyuki Sato, Taro Toyoizumi, and Kazuyuki Aihara. “Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli”. In: *Neural computation* 19.12 (2007), pages 3335–3355 (cited on page 210).

[Sim15] Chris R Sims. “The cost of misremembering: Inferring the loss function in visual working memory”. In: *Journal of vision* 15.3 (2015), pages 2–2 (cited on page 256).

[Ten99] Joshua B Tenenbaum. “Bayesian modeling of human concept learning”. In: *Advances in neural information processing systems* (1999), pages 59–68 (cited on page 138).

[Tod04] Emanuel Todorov. “Optimality principles in sensorimotor control”. In: *Nature neuroscience* 7.9 (2004), pages 907–915 (cited on page 256).

[TML03] Julia Trommershäuser, Laurence T Maloney, and Michael S Landy. “Statistical decision theory and the selection of rapid, goal-directed movements”. In: *JOSA A* 20.7 (2003), pages 1419–1433 (cited on pages 256, 283).

[TML08] Julia Trommershäuser, Laurence T Maloney, and Michael S Landy. “Decision making, movement planning and statistical decision theory”. In: *Trends in cognitive sciences* 12.8 (2008), pages 291–297 (cited on page 288).

[Tro+05] Julia Trommershäuser et al. “Optimal compensation for changes in task-relevant movement variability”. In: *Journal of Neuroscience* 25.31 (2005), pages 7169–7178 (cited on page 256).

[Van+15] Ruben S Van Bergen et al. “Sensory uncertainty decoded from visual cortex predicts behavior”. In: *Nature neuroscience* 18.12 (2015), pages 1728–1730 (cited on page 277).

[Van+12] Ronald Van den Berg et al. “Optimal inference of sameness”. In: *Proceedings of the National Academy of Sciences* 109.8 (2012), pages 3178–3183 (cited on page 210).

[Vri12] Scott I Vrieze. “Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC).” In: *Psychological methods* 17.2 (2012), page 228 (cited on page 335).

[Vul+14] Edward Vul et al. “One and done? Optimal decisions from very few samples”. In: *Cognitive science* 38.4 (2014), pages 599–637 (cited on page 287).

[Wal+20] Edgar Y Walker et al. “A neural basis of probabilistic computation in visual cortex”. In: *Nature Neuroscience* 23.1 (2020), pages 122–129 (cited on page 277).

[Wal17] Pascal Wallisch. “Illumination assumptions account for individual differences in the perceptual interpretation of a profoundly ambiguous stimulus in the color domain:“The dress””. In: *Journal of Vision* 17.4 (2017), pages 5–5 (cited on page 189).

[WK10] Kunlin Wei and Konrad Körding. “Uncertainty of feedback and state estimation determines the speed of motor adaptation”. In: *Frontiers in computational neuroscience* 4 (2010), page 11 (cited on page 237).

[WS15] Xue-Xin Wei and Alan A Stocker. “A Bayesian observer model constrained by efficient coding can explain’anti-Bayesian’percepts”. In: *Nature neuroscience* 18.10 (2015), pages 1509–1517 (cited on page 82).

[WSA02] Yair Weiss, Eero P Simoncelli, and Edward H Adelson. “Motion illusions as optimal percepts”. In: *Nature neuroscience* 5.6 (2002), pages 598–604 (cited on page 82).

[WS08] Louise Whiteley and Maneesh Sahani. “Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes”. In: *Journal of vision* 8.3 (2008), pages 2–2 (cited on pages 256, 285, 288).

[WH01] Felix A Wichmann and N Jeremy Hill. “The psychometric function: I. Fitting, sampling, and goodness of fit”. In: *Perception & psychophysics* 63.8 (2001), pages 1293–1313 (cited on page 101).

[WC19] Robert C Wilson and Anne GE Collins. “Ten simple rules for the computational modeling of behavioral data”. In: *Elife* 8 (2019), e49547 (cited on page 335).

[WNG10] Robert C Wilson, Matthew R Nassar, and Joshua I Gold. “Bayesian online learning of the hazard rate in change-point problems”. In: *Neural computation* 22.9 (2010), pages 2452–2476 (cited on page 237).

[Wol97] Daniel M Wolpert. “Computational approaches to motor control”. In: *Trends in cognitive sciences* 1.6 (1997), pages 209–216 (cited on page 237).

[XLM13] Ting Xiang, Terry Lohrenz, and P Read Montague. “Computational substrates of norms and their violations during social exchange”. In: *Journal of Neuroscience* 33.3 (2013), pages 1099–1108 (cited on page 138).

[XT07] Fei Xu and Joshua B Tenenbaum. “Word learning as Bayesian inference.” In: *Psychological review* 114.2 (2007), page 245 (cited on pages 17, 139, 283, 288).

[YLW16] Scott Cheng-Hsin Yang, Mate Lengyel, and Daniel M Wolpert. “Active sensing in the categorization of visual patterns”. In: *Elife* 5 (2016), e12215 (cited on page 224).

[ZDP98] Richard S Zemel, Peter Dayan, and Alexandre Pouget. “Probabilistic interpretation of population codes”. In: *Neural computation* 10.2 (1998), pages 403–430 (cited on page 277).

[ZAM20] Yanli Zhou, Luigi Acerbi, and Wei Ji Ma. “The role of sensory uncertainty in simple contour integration”. In: *PLoS computational biology* 16.11 (2020), e1006308 (cited on page 210).