Why modeling?

Good modeling of behavior goes beyond a description of the data (curve fitting). Good models help to break down perceptual and cognitive processes into stages, to characterize individual differences, to connect understanding across experiments or domains, or to give you variables to correlate neural data with.

Why Bayesian modeling?

Bayesian modeling is a powerful and arguably elegant framework for modeling behavioral data, especially but not exclusively in perception. People build Bayesian models for different reasons:

- Evolutionary/normative: Bayesian inference optimizes performance or minimizes cost. The brain might have optimized decision-making in common tasks.

- Empirical: in many tasks, people are close to Bayesian.

- Bill Geisler’s couch argument: it is harder to come up with a good model sitting on your couch than to simply work out the Bayesian model, which is recipe-like with little freedom.

- Sensible models can often be constructed by modifying the assumptions in the Bayesian model. Thus, the Bayesian model is a good starting point for model generation.
Would I be successful at Bayesian modeling?

Anybody can learn Bayesian modeling! The math might seem hard at first but after 10 to 50 hours of practice, depending on your background, it is more of the same. This tutorial will put you on the road.

For what kinds of experiments is Bayesian modeling useful?

Bayesian modeling is easiest if the stimuli have unambiguously defined features (not: natural scenes), you parametrically vary those features, and the observer has an unambiguously defined objective (typically simply maximizing task performance).

Overview of this tutorial

This tutorial is organized in a hands-on way: we will go through a set of case study of increasing sophistication, doing a bit of typical math for each. You can type your answers and comments in this document. Please participate actively by asking questions!

Bayesian modeling spans a wide range of topic areas, and we will introduce the principles using cases from different areas:

1. Unequal likelihoods: Gestalt laws
2. Competing likelihoods and priors: Motion sickness
3. Ambiguity from a nuisance parameter: Surface shade perception
4. Inference under measurement noise: Sound localization
5. Hierarchical inference: change point detection
6. Model fitting

Not to be confused

This tutorial is not about Bayesian statistics or Bayesian model comparison, it is about Bayesian observer models. Therefore, we will not discuss Bayesian versus frequentist statistics. Sorry!
1 Unequal likelihoods: Gestalt laws

You observe the five dots below all moving downward, as indicated by the arrows.

The traditional account is that the brain has a tendency to group the dots together because of their common motion, and perceive them as a single object. 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 true explanation of the percept and in some cases even make quantitative predictions.

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 $I_5$, each specifying the direction of movement of the corresponding dot’s image on the retina (i.e. “up” or “down”).

Step 1: 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 or hypotheses:

- **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.

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

a. The generative model diagram in Figure 1 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. Put the following variable names in the correct boxes: retinal images $I_1, I_2, I_3, I_4, I_5$, and motion directions $s$ (a single motion direction), $s_1, s_2, s_3, s_4$, and $s_5$. The same variable might appear more than once.
Step 2: Inference

In inference, the two scenarios become hypothesized scenarios. Inference involves likelihoods and priors. The **likelihood** of a scenario is the probability of the sensory observations under the scenario.

b. What is the likelihood of Scenario 1?
c. What is the likelihood of Scenario 2?
d. Do the likelihoods of the scenarios sum to 1? Explain.
e. What is wrong with the phrase “the likelihoods of the observations”?

Priors

Let’s say Scenario 1 occurs twice 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.

f. What are the prior probabilities of Scenarios 1 and 2?

Protoposteriors

The **protoposterior** of a scenario (a convenient, but not an official term) is its likelihood multiplied by its prior probability.

g. What are the protoposteriors of Scenarios 1 and 2?
h. Do protoposteriors of the scenarios sum to 1?

Posteriors

Posterior probabilities have to sum to 1.
i. To achieve that, divide each of the protoposteriors by their sum. Calculate the posterior probabilities of Scenarios 1 and 2. You have just applied Bayes rule.
Percept

The standard Bayesian observer model for discrete hypotheses states that the percept is the scenario with the highest posterior probability (maximum-a-posteriori or MAP estimation).

j. Would that be consistent with the law of common fate? Explain.

k. How does this Bayesian observer model complement - or go beyond - the traditional Gestalt account of this phenomenon?

Note that in this case, like very often, the action is in the likelihood. The prior is relatively unimportant.

The interpretation of the likelihood function is in terms of hypotheses. When an observer is faced with specific sensory observations, what is the probability of those observations when the world state (here Scenario 1 or 2) takes a certain value? Each possible value of the world state is a hypothesis, and the likelihood of that hypothesis is the observers belief that the observations would arise under that hypothesis. Priors and posteriors are similarly belief distributions, whose arguments are hypotheses.
2 Competing likelihoods and priors: Motion sickness

Michel Treisman has tried to explain motion sickness in the context of evolution. 
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 don’t 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. Sensory observations: Your brain has two sets of sensory observations: visual observations and vestibular observations. Hypotheses: Let’s say 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.

**Step 1: Generative model**

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

**Step 2: Inference**

b. In prehistory, people would of course move around in the world, but 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.

c. 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).

d. Draw a bar diagram that illustrates the posterior probabilities of the three scenarios.

---

1Michel Treisman (1977), *Motion sickness: an evolutionary hypothesis*, Science 197, 493-495.
e. Use numbers to illustrate the calculations in (b)-(d).

f. Explain using the posterior probabilities your “percept” — why you might vomit in this situation.
3 Ambiguity from a nuisance parameter: Surface shade perception

We switch domains once again and turn to the central problem of color vision, simplified to a problem for grayscale surfaces. We see a surface when there is a light source. The surface absorbs some proportion of the incident photons and reflects the rest. Some of the reflected photons reach our retina.

Step 1: Generative model

The diagram of the generative model is

The shade of a surface is the grayscale in which a surface has been painted. It is a form of color of a surface. Technically, shade is reflectance: the proportion of incident light that is reflected. Black paper might have a reflectance of 0.10 while white paper might have a reflectance of 0.90. The intensity of a light source (illuminant) is the amount of light it emits. Surface shade and light intensity are the external world state variables relevant to this problem.

The sensory observation is the amount of light measured by the retina, which we will also refer to as retinal intensity. The retinal intensity can be calculated as follows:

\[
\text{Retinal intensity} = \text{surface shade} \times \text{light intensity}
\] (1)

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.

Step 2: Inference

Let’s take each of these numbers 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 \times 0.2 = 0.1.

a. Suppose your retinal intensity is 0.2. Suppose further that you hypothesize the light intensity to be 1 (very bright light). Under that hypothesis, calculate what the surface shade must have been.

b. Suppose your retinal intensity is the same 0.2. Suppose further that you hypothesize the light intensity to be 0.4. Under that hypothesis, calculate what the surface shade must have been.
c. Explain why the retinal intensity provides ambiguous information about surface shade.

**Likelihood**

d. Suppose your retinal intensity is again 0.2. By going through a few more examples like the ones in (a) and (b), draw in the two-variable likelihood diagram in Figure 2 all combinations of hypothesized surface shade and hypothesized light intensity that could have produced your retinal intensity of 0.2. **Think of this plot as a 3D plot (surface plot)!**

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

![Figure 2](image)

**Prior**

f. Suppose you have a strong prior that light intensity was between 0.2 and 0.4, and definitely nothing else. In the two-variable prior diagram in Figure 2 (center), shade the area corresponding to this prior.

**Posterior**

g. In the two-variable posterior diagram in Figure 2 (right), indicate where the posterior probability is high.

h. What would you perceive according to the Bayesian theory?
4 Inference under measurement noise: Sound localization

The previous cases featured categorically distinct scenarios. We now consider a continuous estimation task, e.g. locating a sound on a horizontal line. This will allow us to introduce the concept of noise in the internal measurement of the stimulus. This case would be uninteresting without such noise.

Step 1: Generative model

In this case, the generative model is a description of the statistical structure of the task and the internal measurement. The stimulus is the location of the sound. The sensory observations generated by the sound location consist of a complex pattern of auditory neural activity, but for the purpose of our model, and reflecting common practice, we reduce the sensory observations to a single scalar, namely a noisy internal measurement \( x \). The measurement lives in the same space as the stimulus itself, in this case the real line. For example, if the true location \( s \) of the sound is \( 3^\circ \) to the right of straight ahead, then its measurement \( x \) could be \( 2.7^\circ \) or \( 3.1^\circ \).

Thus, the problem contains two variables: the stimulus (true sound location, \( s \)) and the observers internal measurement of the stimulus, \( x \). These two random variables appear in the generative model, which is depicted here. We call them “random variables” because they haven’t yet taken on a value - they are just abstract quantities with particular probability distributions. Thus, each node in the graph is associated with a probability distribution: the stimulus node with a stimulus distribution \( p(s) \), and the measurement node with a measurement distribution \( p(x|s) \) that depends on the value of the stimulus. When we actually consider a particular trial of the experiment, some of these variables will take on actual values and cease to be random, as we’ll see below. But first let’s consider the random variables and specify their distributions:

The stimulus distribution

The distribution associated with the stimulus \( s \) is denoted \( p(s) \). In our example, say that the experimenter has programmed \( p(s) \) to be Gaussian with a mean \( \mu \) and variance \( \sigma_s^2 \).

\[
p(s) = \frac{1}{\sqrt{2\pi \sigma_s^2}} e^{-\frac{(s-\mu)^2}{2\sigma_s^2}}, \tag{2}
\]

see Figure 3. This Gaussian shape with mean zero implies that the sound is most often presented straight ahead than at any other location. This is imposed by the experimenter, who decides how to sample the random stimulus locations from trial to trial.
Figure 3: The probability distributions that belong to the two variables in the generative model. (a) A Gaussian distribution over the stimulus, \( p(s) \), reflecting the frequency of occurrence of each stimulus value in the world. (b) Suppose we now fix a particular value of \( s \) (the dotted line). Then the measurements \( x \) will follow a Gaussian distribution around that \( s \). The diagram at the bottom shows a few samples of \( x \), which are evidently scattered around the true sound location \( s \) indicated by the arrow.

### The measurement

The *measurement distribution* is the distribution of the measurement \( x \) for a given stimulus value \( s \). We make the common assumption that the measurement distribution is Gaussian:

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

where \( \sigma \) is the standard deviation of the measurement noise, also called *measurement noise level* or *sensory noise level*. This Gaussian distribution is shown in Figure 3b. The higher \( \sigma \), the noisier the measurement and the wider its distribution.

### Step 2: Inference

#### The prior distribution

We introduced the stimulus distribution \( p(s) \), which reflects how often each stimulus value tends to occur in the experiment. Suppose that the observer has learned this distribution through training. Then, the observer will already have an *expectation* about the stimulus before it even appears. This expectation constitutes prior knowledge, and therefore, in the inference process, \( p(s) \) is referred to as the *prior distribution* (Figure 4a). The prior distribution is mathematically identical to the stimulus distribution \( p(s) \), but unlike it, the prior distribution reflects the observer’s *beliefs on a given trial*, rather than a fact about the external world.
The likelihood function

The likelihood function represents the observer’s belief about the stimulus based on the measurements only — absent any prior knowledge. Formally, the likelihood is the probability of the observed measurement under a hypothesized stimulus:

\[ L(s) \equiv p(x_{\text{obs}}|s). \] (4)

We note two novel notations in this equation:

- \( x_{\text{obs}} \) is now an actual measurement, with a particular value; it is not a random variable like \( s \), which is unknown to the observer

- It seems that we have just re-defined \( p(x|s) \) under a new name - why? The key point here is that the likelihood function is a function of \( s \), not of \( x \). The \( x \) variable is now fixed to the observed value \( x_{\text{obs}} \), and we want to consider the likelihood of this observed measurement under various possible values of the random variable \( s \). Thus, we treat the likelihood \( L(s) \) as a function of \( s \).

Under our assumption for the measurement distribution \( p(x|s) \), the likelihood function over the stimulus is

\[ L(s) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-x_{\text{obs}})^2}{2\sigma^2}}. \] (5)

(Although this particular likelihood is normalized over \( s \), that is not generally true. This is why the likelihood function is called a function and not a distribution.) The width of the likelihood function is interpreted as the observers level of uncertainty based on the measurements alone. A narrow likelihood means that the observer is certain based on the measurements alone, a wide likelihood that the observer is uncertain based on the measurements alone.
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 the stimulus variable \( s \) given an observed measurement \( x_{\text{obs}} \). Bayes’ rule states

\[
p(s|x_{\text{obs}}) = \frac{p(x_{\text{obs}}|s)p(s)}{p(x_{\text{obs}})}.
\]

Using the terminology introduced in Case 1, we can also write the posterior in terms of the protoposterior as

\[
p(s|x_{\text{obs}}) \propto p(x_{\text{obs}}|s)p(s) = L(s)p(s).
\]

a. Why can we get away with the proportionality sign?

All we have done so far is apply a general rule of probability calculus. What is important, however, is the interpretation of Eq. (7). As we discussed before for the prior distribution and likelihood function, this probability distribution reflects the observer’s beliefs about the value of the random variable \( s \). More concretely, Eq. (7) assigns a probability to each possible hypothesized value of the unknown stimulus \( s \). The observer can thus use this equation to consider all possible values of \( s \), and ask to what extent each of them is supported by the observed measurement \( x_{\text{obs}} \) and prior beliefs. The answer to this question is the posterior distribution, \( p(s|x_{\text{obs}}) \).

We will now compute the posterior under the assumptions we made in Step 1. Upon substituting the expressions for \( L(s) \) and \( p(s) \) into Eq. (7) we see that in order to compute the posterior, we need to compute the product of two Gaussian functions. An example is shown in Fig. 5.

b. Create a figure similar to Figure 5 through numerical computation of the posterior.

Numerically normalize prior, likelihood and posterior.

Beyond plotting the posterior, our assumptions in this case actually allow us to characterize the posterior mathematically (“analytically”).

c. Show that the posterior is a new Gaussian distribution

\[
p(s|x_{\text{obs}}) = \frac{1}{\sqrt{2\pi\sigma^2_{\text{posterior}}}} e^{-\frac{(s-\mu_{\text{posterior}})^2}{2\sigma^2_{\text{posterior}}}},
\]

with mean

\[
\mu_{\text{posterior}} = \frac{x_{\text{obs}}}{\sigma^2} + \frac{\mu}{\sigma^2},
\]

(9)
Figure 5: The posterior distribution is obtained by multiplying the prior with the likelihood function.

You may use the following auxiliary calculation:

\[- \frac{(s - \mu)^2}{2\sigma_s^2} - \frac{(s - x_{\text{obs}})^2}{2\sigma^2} = - \frac{s^2 - 2\mu s + \mu^2}{2\sigma_s^2} - \frac{s^2 - 2x_{\text{obs}}^2 s + x_{\text{obs}}^2 }{2\sigma^2} \]

\[= - \frac{s^2}{2\sigma_s^2} + \frac{2\mu s}{2\sigma_s^2} - \frac{\mu^2}{2\sigma_s^2} - \frac{s^2}{2\sigma^2} - \frac{2x_{\text{obs}}^2 s}{2\sigma^2} - \frac{x_{\text{obs}}^2}{2\sigma^2} \]

\[= - \left( \frac{1}{2\sigma_s^2} + \frac{1}{2\sigma^2} \right) s^2 + \left( \frac{\mu}{\sigma_s^2} + \frac{x_{\text{obs}}}{\sigma^2} \right) s + \text{junk} \]

\[= - \frac{1}{2} \left( \frac{1}{\sigma_s^2} + \frac{1}{\sigma^2} \right) \left( s^2 - \frac{2\mu}{\sigma_s^2} + \frac{x_{\text{obs}}}{\sigma^2} \right) + \text{junk} \]

\[= \frac{1}{2} \left( \frac{1}{\sigma_s^2} + \frac{1}{\sigma^2} \right) \left( s - \frac{\mu}{\sigma_s^2} + \frac{x_{\text{obs}}}{\sigma^2} \right)^2 + \text{junk} \]

where “junk” refers to terms that don’t depend on \(s\) (see question 4a above to understand why we can ignore these when calculating the new distribution).

The mean of the posterior, Eq. [9], is of the form \(ax_{\text{obs}} + b\mu\), in other words, a linear combination of \(x_{\text{obs}}\) and the mean of the prior, \(\mu\). The coefficients \(a\) and \(b\) in this linear combination are \(\frac{1}{\sigma_s^2 + \sigma^2}\) and \(\frac{1}{\sigma^2 + \sigma_s^2}\), respectively. These sum to 1, and therefore the linear combination is a \textit{weighted average}, where the coefficients act as weights. This weighted
average, $\mu_{\text{posterior}}$, will always lie somewhere in between $x_{\text{obs}}$ and $\mu$.

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

The intuition behind the weighted average is that the prior “pulls the posterior away” from the measurement $x_{\text{obs}}$ and towards its own mean $\mu$, 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 $\sigma$ is low — the posterior won’t budge much: it will be centered close to the mean of the likelihood function. This intuition is still valid if the likelihood function and the prior are not Gaussian but are roughly bell-shaped.

The variance of the posterior is given by Eq. (10). 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.

e. Show that the variance of the posterior can also be written as $\sigma^2_{\text{posterior}} = \frac{\sigma^2 \sigma^2_s}{\sigma^2 + \sigma^2_s}$.

f. Show that the variance of the posterior is smaller than both the variance of the likelihood function and the variance of the prior distribution. This shows that combining a measurement with prior knowledge makes an observer less uncertain about the stimulus.

g. What is the variance of the posterior in the special case that $\sigma = \sigma_s$?

The stimulus estimate

We now estimate $s$ on the trial under consideration. For a continuous variable, it is a good idea to simply take the mean of the posterior. We denote the estimate by $\hat{s}$. Thus, $\hat{s} = \mu_{\text{posterior}} = \frac{x_{\text{obs}}}{\sigma^2} + \frac{\mu}{\sigma^2_s}.$

This would be a Bayesian observer’s response in this localization task.

Step 3: Predicted response distribution

We’d now like to use this model to predict subjects’ behavior in this experiment. To do so, we’d like to compare our predicted responses, $\hat{s}$, to the subject’s actual responses. Looking at our above equation for $\hat{s}$ (Eq. (11), we note that to compute a predicted response on a given trial we need to know $x_{\text{obs}}$. But this is something we don’t know! $x_{\text{obs}}$ is the noisy measurement made by the observer’s sensory system - an internal variable that we have no access to as experimenters.

It is a common mistake in Bayesian modeling is to discuss the likelihood function (or the posterior distribution) as if it were a single, specific function on a given trial. This is
entirely incorrect. Both the likelihood and the posterior depend on the measurement $x_{\text{obs}}$, which itself is randomly generated on each trial, and therefore the likelihood and posterior will “wiggle around” from trial to trial (Fig. 6).

This variability propagates to the estimate: the estimate $\hat{s}$ also depends on the noisy measurement $x_{\text{obs}}$ via Eq. (11). Since $x_{\text{obs}}$ varies from trial to trial, so does the estimate: the stochasticity in the estimate is inherited from the stochasticity in the measurement $x_{\text{obs}}$. Hence, in response to repeated presentations of the same stimulus, the estimate will be a random variable with a probability distribution, which we will denote by $p(\hat{s}|s)$.

So rather than comparing our model’s predicted responses to subjects’ actual responses on individual trials, we’ll instead use our model to predict the distribution over subjects’ responses for a given value of the stimulus. The predicted distribution is precisely $p(\hat{s}|s)$. To compare our Bayesian model with an observer’s behavior, we thus need to calculate this distribution.

h. From Step 1, we know that when the true stimulus is $s$, $x_{\text{obs}}$ follows a Gaussian distribution with mean $s$ and variance $\sigma^2$. Show that when the true stimulus is $s$, the estimate’s distribution $p(\hat{s}|s)$ is a Gaussian distribution with mean $\frac{\hat{s}}{\sigma^2 + \frac{1}{\sigma^2}}$ and variance $\frac{1}{\sigma^2 + \frac{1}{\sigma^2}}$. In particular, it follows that the variance of the estimate is not equal to the variance of the posterior.

This completes the model: the distribution $p(\hat{s}|s)$ can now be compared to human behavior.
5 Hierarchical inference: change point detection

Our last case is change point detection, the task of inferring from a time series of noisy observations whether or when an underlying variable changed. There are two main reasons for considering this task. First, it is a common form of inference. For example, a new chef or owner might cause the quality of the food in a restaurant to suddenly improve (or drop). A neurologist may want to detect a seizure on an EEG in a comatose patient. Second, this case is representative of a type of inference that involves multiple layers of latent variables – in our case, not only the stimulus at each time point but also the “higher-level” variable of when the stimulus changed.

Figure 7: (a) Generative model for our change detection problem. (b) Example of a true stimulus sequence with \( t_{\text{change}} = 7 \). (c) Example sequence of measurements. When did the change from \(-1\) to \(1\) occur?

Step 1: Generative model

The generative model is shown in Figure 7. Time is discrete and goes from 1 to \( T \). A change occurs at exactly one time point \( t_{\text{change}} \), chosen with equal probability:

\[
p(t_{\text{change}}) = \frac{1}{T}. \tag{12}
\]

The stimulus \( s = (s_1, \ldots, s_T) \) is a sequence that starts with repetitions of the value \(-1\) and at some point changes to repetitions of the value \(1\). The change point is when the value \(-1\) changes to \(1\). Formally,

\[
s_t = \begin{cases} 
-1 & \text{if } t < t_{\text{change}} \\
1 & \text{if } t \geq t_{\text{change}} 
\end{cases} \tag{13}
\]

(According to this specification, if \( t_{\text{change}} = 1 \), then is never \(-1\) to begin with.) Finally, we assume that the observer makes independent, normally distributed measurements \( x = (x_1, \ldots, x_T) \), all with the same variance:

\[
p(x|s) = \prod_{t=1}^{T} p(x_t|s_t) \tag{14}
\]

\[
p(x_t|s_t) = \mathcal{N}(x_t; s_t, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_t-s_t)^2}{2\sigma^2}}. \tag{15}
\]
Step 2: Inference

The observer would like to infer when the change occurred. What makes this problem non-trivial is that the noise in the measurements can create “apparent changes” that are not due to a change in the underlying state $s_t$. The Bayesian observer solves this problem in an optimal manner.

The world state of interest is change point $t_{\text{change}}$. Therefore, the Bayesian observer computes the posterior over $t_{\text{change}}$ given a sequence of measurements, $x$ (we leave out the subscript “obs” for ease of notation). We first apply Bayes’ rule:

$$p(t_{\text{change}}|x) \propto p(t_{\text{change}})p(x|t_{\text{change}}). \quad (16)$$

Since the prior is constant, $p(t_{\text{change}}) = \frac{1}{T}$, this simplifies to

$$p(t_{\text{change}}|x) \propto p(x|t_{\text{change}}). \quad (17)$$

Thus, our challenge is to calculate the likelihood function

$$L(t_{\text{change}}) = p(x|t_{\text{change}}). \quad (18)$$

For each hypothesized value of $t_{\text{change}}$, this function tells us how expected the observations are under that hypothesis. The problem is that – unlike in Case 4 – the generative model does not give us this likelihood function right away; this is a consequence of the “stacked” or hierarchical nature of the generative model. We do have the distributions $p(x|s)$ and $p(s|t_{\text{change}})$. To make the link, we have to “average” over all possible values of $s$. This is called “marginalization”. Marginalization is extremely common in Bayesian models except for the very simplest ones – the reason being that there are almost always unknown states of the world that affect the observations but that the observer is not primarily interested in. The box describes the relevant probability calculus.

**Marginalization of probabilities.** If $A$ and $B$ are two discrete random variables, their joint distribution is $p(A, B)$. From the joint distribution, we can obtain the distribution of one of the variables by summing over the other one, for example:

$$p(A) = \sum_B p(A, B) \quad (19)$$

This is called the marginal distribution of $A$. Making use of the definition of conditional probability, $p(A|B) = \frac{p(A, B)}{p(B)}$, we further write

$$p(A) = \sum_B p(A|B)p(B). \quad (20)$$

We can obtain a variant of this equation by conditioning each probability on a third random variable $C$:

$$p(A|C) = \sum_B p(A|B, C)p(B|C). \quad (21)$$

This is the equation we use in Case 5.
We compute $L(t_{\text{change}})$ by marginalizing over $s$:

$$L(t_{\text{change}}) = \sum_s p(x|s)p(s|t_{\text{change}}), \quad (22)$$

a. Besides using Eq. (21), we used a property that is specific to our generative model. Which one?

b. For a given $t_{\text{change}}$, how many sequences $s$ are possible?

Now we understand why the sum reduces to a single term:

$$L(t_{\text{change}}) = p(\text{the one } s \text{ in which the change occurs at } t_{\text{change}}) \quad (23)$$

$$= \left( \prod_{t=1}^{t_{\text{change}}-1} p(x_t|s_t = -1) \right) \left( \prod_{t=t_{\text{change}}}^{T} p(x_t|s_t = 1) \right). \quad (24)$$

This looks complicated but we are not out of ideas!

c. Show that this can be written more simply as

$$L(t_{\text{change}}) \propto \prod_{t=t_{\text{change}}}^{T} \frac{p(x_t|s_t = 1)}{p(x_t|s_t = -1)}. \quad (25)$$

d. Now plug in $p(x_t|s_t) = \mathcal{N}(x_t; s_t, \sigma)$ to find

$$L(t_{\text{change}}) \propto e^{\frac{2}{\sigma^2} \sum_{t=t_{\text{change}}}^{T} x_t}. \quad (26)$$

e. Discuss with your neighbor whether this equation makes intuitive sense.

Combining Eqs. (17), (18), (26), we find for the posterior probability of a change point at $t_{\text{change}}$:

$$p(t_{\text{change}}|x) \propto e^{\frac{2}{\sigma^2} \sum_{t=t_{\text{change}}}^{T} x_t}. \quad (27)$$

To obtain the actual posterior probabilities, the right-hand-side has to be normalized (divided by the sum over all $t_{\text{change}}$).

f. The data in Figure 7 are $x = (-0.46, 0.83, -3.26, -0.14, -0.68, -2.31, 0.57, 1.34, 4.58, 3.77)$, with $\sigma = 1$. Plot the posterior distribution over change point.

However, if the goal is just to pick the most probable change point (MAP estimate), normalizing is not needed.

g. Why not?

h. If that is the goal, the decision rule becomes “Cumulatively add up the elements of $x$, going backwards from $t = T$ to $t = 1$. The time at which this cumulative sum peaks is the MAP estimate of $t_{\text{change}}$.” Explain.
Step 3: Predicted response distribution

In Case 4, we were able to obtain an equation for the predicted response distribution. Here however, and in many other cases, that is not possible. Nevertheless, we can still simulate the model’s responses to obtain an approximate prediction.

i. Assume $\sigma = 1$ and $T = 10$. Vary the true change point $t_{\text{change}}$ from 1 to $T$. For each value of $t_{\text{change}}$, we simulate 10,000 trials (or more if possible). On each simulated trial,

(a) Based on $t_{\text{change}}$, specify the stimulus sequence $s$.

(b) Simulate a measurement sequence $x$ from $s$.

(c) Apply the decision rule to each measurement sequence. The output is the simulated observer’s response, namely an estimate of $t_{\text{change}}$.

(d) Determine whether the response was correct.

Plot proportion correct as a function of $t_{\text{change}}$. Interpret the plot.

j. What is overall proportion correct (averaged across all $t_{\text{change}}$)?

k. Vary $\sigma = 1, 2, 3$ and $T$ from 2 to 16 in steps of 2. Plot overall proportion correct as a function of $T$ for the three values of $\sigma$ (color-coded). Interpret the plot.

l. How could our simple example be extended to cover more realistic cases of change point detection?

6 Model fitting

This section is not limited to Bayesian models.

Maximum-likelihood estimation

Each model has parameters. The most basic way to fit them is using maximum-likelihood estimation. The likelihood of a parameter combination is the probability of the data given that parameter combination (and a specific model):

\[ L(\text{parameters}) = p(\text{data}|\text{parameters}). \] (28)

This likelihood is different from the likelihood functions discussed in the Bayesian observer models, but analogous. To the observer, the state of the world is unknown. To the experimenter, the “state of the observer” is unknown.

As an example, consider the data from Case 4, which consists of an \((s, \hat{s})\) pair on each trial. Since \(\mu\) and \(\sigma_s\) are experimentally controlled, \(\sigma\) is the only parameter that the experimenter needs to estimate. Denoting the \(i^{th}\) trial by a subscript \(i\) and assuming that the trials are independent, the likelihood is

\[ L(\sigma) = p(\hat{s}_1|s_1, \sigma) \cdot p(\hat{s}_2|s_2, \sigma) \cdot p(\hat{s}_3|s_3, \sigma) \cdot \ldots \]

\[ = \prod_{i=1}^{n_{\text{trials}}} p(\hat{s}_i|s_i, \sigma). \] (29)

Maximizing the likelihood is equivalent to maximizing the log likelihood, which usually has a nicer form. The log likelihood is

\[ \log L(\sigma) = \sum_{i=1}^{n_{\text{trials}}} p(\hat{s}_i|s_i, \sigma). \] (30)

The expression for the log likelihood needs to be evaluated by using the model predictions (output of Step 3). For example, in Case 4 the probability \(p(\hat{s}_i|s_i, \sigma)\) is given in Exercise 4. Finally, we note that in practice, most model fitting algorithms used in Matlab or Python (e.g. fmincon, BADS, etc...see Section 6.2 below) minimize the negative log likelihood:

\[ \arg\min_{\sigma} -\log L(\sigma) = -\sum_{i=1}^{n_{\text{trials}}} p(\hat{s}_i|s_i, \sigma) \] (31)

which is equivalent to maximizing the log likelihood.

Algorithms for parameter estimation

Now that we have the log likelihood, we have to find the parameter combination to maximize it. What algorithm to choose?
• For one-dimensional problems (a single parameter), evaluating the log likelihood on a fine grid is simple and feasible. It has the advantage that the log likelihood can be plotted. However, grid search would quickly take too long for more parameters. Moreover, a grid always suffers from the drawback that it is impossible to find values in between the grid points.

• In general, we recommend fmincon in Matlab for most well-behaved problems. Well-behaved means that the log likelihood does not have many local maxima, and is not too noisy (when calculated through simulation).


Note that both fmincon and BADS are minimization rather than maximization algorithms; the function to be minimized is simply the negative log likelihood.

Whatever algorithm you use, issues to be aware of are local minima, ridges, and saddle points. An easy way to alleviate these problems is by running the algorithm multiple times, with a different starting point (initial parameter combination) each time.