6

Statistical Decision Theory and Biological Vision

LAURENCE T. MALONEY
Institut für Biophysik, Freiburg, Germany

> *I know of only one case in mathematics of a doctrine which has been accepted and developed by the most eminent men of their time . . . which at the same has appeared to a succession of sound writers to be fundamentally false and devoid of foundation. Yet this is quite exactly the position in respect of inverse probability [an estimation method based on Bayes' theorem].*
>
> R.A. Fisher (1930) *Inverse Probability*

Statistical Decision Theory (SDT) emerged in essentially its final form with the 1954 publication of Blackwell and Girshick's *Theory of Games and Statistical Decisions*. The elements out of which it developed antedate it, in some cases by centuries, and, as the title indicates, an immediate stimulus to its development was the publication of *Theory of Games and Economic Behavior* by von Neumann and Morgenstern (1944/1953). Like Game Theory, SDT is normative, a set of principles that tell us how to act so as to maximize gain and minimize loss.[1]

The basic metaphor of SDT is that of a game between an Observer and the World. The Observer has imperfect information about the state of the World, analogous to sensory information, and must choose an action from among a limited repertoire of possible actions. This action, together with the true state of the World, determines its gain or loss: whether it has stumbled off a cliff in the dark, avoided an unwelcome invitation to (be) lunch, or—most important of all—correctly responded in a psychophysical task. SDT prescribes how the Observer should choose among possible actions, given what information it has, so as to maximize its expected gain.

Bayesian Decision Theory (BDT) is a special case of SDT, but one of particular relevance to a vision scientist. Recently, a number of authors (see, in particular, Knill et al., 1996; Knill & Richards, 1996; Kersten & Schrater, this volume) have argued that BDT and related Bayesian-inspired techniques form a particularly congenial "language"

Perception and the Physical World: Psychological and Philosophical Issues in Perception.
Edited by Dieter Heyer and Rainer Mausfeld. © 2002 John Wiley & Sons, Ltd.

for modeling aspects of biological vision. We are, in effect, invited to believe that increased familiarity with this "*language*" (its concepts, terminology, and theory) will eventually lead to a deeper understanding of biological vision through better models, better hypotheses and better experiments. To evaluate a claim of this sort is very different from testing a specific hypothesis concerning visual processing. The prudent, critical, or eager among vision scientists need to master the language of SDT/BDT before evaluating, disparaging, or applying it as a framework for modeling biological vision.

Yet the presentation of SDT and BDT in research articles is typically brief. Standard texts concerning BDT and Bayesian methods are directed to statisticians and statistical problems. Consequently, it is difficult for the reader to separate important assumptions underlying applications of BDT to biological vision from the computational details; it is precisely these assumptions that need to be understood and tested experimentally. Accordingly, this chapter is intended as an introduction for those working in biological vision to the elements of SDT and to their intelligent application in the development of models of visual processing. It is divided into an introduction, four "sections", and a conclusion.

In the first of the four sections, I present the basic framework of SDT, including BDT. This framework is remarkably simple; I have chosen to present it in a way that emphasizes its visual or geometric aspects, although the equations are there as well. As the opening quote from Fisher hints, certain Bayesian practices remain controversial. The controversy centers on the representation of belief in human judgment and decision making, and the "updating" of belief in response to evidence. In the initial presentation of the elements of SDT and BDT in the next section, I will avoid controversy by considering only decisions made at a single instant of time ("*instantaneous BDT*"), where the observer has complete information.

SDT comprises a "mathematical toolbox" of techniques, and anyone using it to model decision making in biological vision must, of course, decide how to assemble the elements into a biologically-pertinent model: SDT itself is no more a model of visual processing than is the computer language Matlab®. The second section of the article contains a discussion of the elements of SDT, how they might be combined into biological models, and the difficulties likely to be encountered Shimojo and Nakayama (1992), among others, have argued that optimal Bayesian computations require more "data" about the world than any organism could possibly learn or store. Their argument seems conclusive. If organisms are to have accurate estimates of relevant probabilities in moderately complex visual tasks, then they must have the capability to assign probabilities to events they have never encountered, and to estimate gains for actions they have never taken. The implications of this claim are discussed.

The third section comprises two "challenges" to the Bayesian approach, the first concerning the status of the visual representation in BDT-derived models. To date, essentially all applications of BDT to biological vision have been attempts to model the process of arriving at internal estimates of depth, color, shape, etc., with little consideration of the real consequences of errors in estimation. A typical "default" goal is to minimize the least-square error of the estimate. But the consequences of errors in, for example, depth estimation depend on the specific visual task that the organism is engaged in—leaping a chasm, say, versus tossing a stone at a target. BDT is in essence a way to choose among actions given knowledge of their consequences: it is equally applicable to leaping chasms, and to tossing stones. What is not obvious is how BDT can be used to compute internal estimates when the real consequences of error are not known. This discussion is evidently relevant to issues concerning perception and action raised by Milner and Goodale (1996) and others.

The second challenge concerns vision across time and what I will call the *updating problem*. Instantaneous BDT assumes that, in each instant of time, the environment is essentially stochastic. Given full knowledge of the distributions of the possible outcomes, instantaneous BDT prescribes how to choose the optimal action. Across time, however, the distributional information may itself change, and change deterministically. The amount of light available outdoors in terrestrial environments varies stochastically from day to day but also cycles deterministically over every 24-hour period. I describe a class of *Augmented Bayes Observers* that can anticipate such patterned change and make use of it.

A recurring criticism of Bayesian biological vision is that is computationally implausible. Given that we know essentially nothing about the computational resources of the brain, this sort of criticism is premature. Nevertheless, it is instructive to consider possible implementations of BDT, and the fourth section of the article discusses what might be called "Bayesian computation" and its computational complexity.

Blackwell and Girshick's *Theory of Games and Statistical Decisions* appeared just 300 years after the 1654 correspondence of Pascal and Fermat in which they developed the modern concepts of expectation and decision making guided by expectation maximization (reported in Huygens, 1657; Arnauld 1662/1964; see Ramsey, 1931a). It appeared obvious to Pascal, Fermat, Arnauld and their successors that any reasonable and reasonably intelligent person would act so as to maximize gain. It is a peculiar fact that all of the ideas underlying SDT and BDT (probabilistic representation of evidence, expectation maximization, etc.) were originally intended to serve as both normative *and* descriptive models of human judgment and decision making. Many advocates of a "Bayesian framework" for biological vision find it equally evident that perceptual processing can be construed as maximizing an expected gain (Knill et al., 1986; Kersten & Schrater, this volume).

It is therefore important to recognize that, as a model of *conscious* human judgment and decision making, BDT has proven to be fundamentally wrong (Green & Swets, 1966/1974; Edwards, 1968; Tversky & Kahneman, 1971, 1973, 1970; Kahneman & Tversky, 1972; see also Kahneman & Slovic, 1982; Nisbett & Ross, 1982). People's use of probabilities and information concerning possible gains deviates in many respects from normative use as prescribed by SDT/BDT and the axioms of probability theory. The observed deviations are large and patterned, suggesting that, in making decisions consciously, human observers are following rules other than those prescribed by SDT/BDT.

Therefore, those who argue that the Bayesian approach is a "necessary", "obvious" or "natural" framework for perceptual processing (Knill et al., 1996; Kersten & Schrater, this volume) should perhaps explain why the same framework fails as a model of human conscious judgment, for which it was developed. It would be interesting to systematically compare "cognitive" failures in reasoning about probability, gain, and expectation to performance in analogous visually-guided tasks. I will return to this point in the final discussion.

A companion article in this volume (Kersten & Schrater, this volume) contains a review of recent work in Bayesian biological vision, and a second companion article (von der Twer, Heyer & Mausfeld, this volume) contains a spirited critique. Knill and Richards (1996) is a good starting point for the reader interested in past work. Williams (1954) is still a delightful introduction to Game Theory, a component of SDT. Ferguson (1967) is an advanced mathematical presentation of SDT and BDT, while Berger (1985) and O'Hagan (1994) are excellent, modern presentations with emphasis on statistical issues.

AN OUTLINE OF STATISTICAL DECISION THEORY

> *. . . to judge what one ought to do to obtain a good or avoid an evil, one must not only consider the good and evil in itself, but also the probability that it will or will not happen; and view geometrically*[2] *the proportion that all these things have together. . . .*
>
> *A. Arnauld (1662/1964) Logic, or the Art of Thinking*

ELEMENTS

As mentioned above, Statistical Decision Theory (Blackwell & Girshick, 1954) developed out of Game Theory (von Neumann & Morgenstern, 1944/1953), and the basic ideas underlying it are still most easily explained in the context of a game with two players, to whom I'll refer as the *Observer* and the *World.*

In any particular application of Statistical Decision Theory (SDT) in biological vision, the Observer and the World take on specific identities. The possible states of the World may comprise a list of distances to surfaces in all directions away from the Observer, while the Observer is a depth estimation algorithm. Alternatively, the World may have only two possible states (SIGNAL and NO-SIGNAL) and the Observer judges the state of the World. As these examples suggest, the same organism may employ different choices of "Observer", "World", and the other elements of BDT in carrying out different visual tasks via different visual "modules".

In both of these examples, the Observer's task is to *estimate* the state of the World. SDT and the subset of it known as Bayesian Decision Theory (BDT) are typically used to model estimation tasks within biological vision: "the World is in an unknown state; estimate the unknown state". Recent textbooks tend to emphasize estimation, and vision scientists do tend to view early visual processing as fundamentally an estimation task (Marr, 1982; Wandell, 1995; Knill & Richards, 1996).

Yet SDT itself has potentially broader applications: earlier presentations (Blackwell & Girshick, 1954; Ferguson, 1967) emphasized that SDT is fundamentally a theory of preferable actions, with estimation regarded as only one particular kind of action. Rather than estimate the distance to a nearby object, the Observer can decide whether it is desirable to throw something at it, or to run away, or both, or neither. And, rather than assess whether a SIGNAL is or is not present, the Observer may concentrate on what to tell the experimenter in a signal detection task, so as to maximize his reward. In both cases the emphasis is on the consequences of the Observer's actions, and the Observer's "accuracy" in estimating the state of the World is of only secondary concern, if it is of any concern at all.

What is constant in all applications of SDT is that (1) the Observer has imperfect information about the World through a process analogous to sensation, that (2) the Observer acts upon the World, and that (3) the Observer is rewarded as a function of the state of the World and its own actions. I'll begin by describing the elements of SDT (and eventually BDT) at *a single instant of time.* We are not yet concerned with past or future but only with selecting the best action at one point in time that I'll refer to as a *turn.*

On each turn, the World is in one of several possible states,

$$\Theta = \{\theta_1, \theta_2, \ldots, \theta_m\}, \tag{6.1}$$

and the current state of the World is denoted θ_*. Each of the states of the World can be a vector (a list of numbers) and need not be just a single number. In any modeling application, the World need only include the information needed to determine the consequences of the Observer's possible actions.

On each turn, the Observer's selects one of its possible *actions,*

$$A = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}. \tag{6.2}$$

Each action can be a vector (a list of actions). An action might, for example, specify a sequence of motor commands to be issued. The chosen action is denoted α_*. The current state of the World and the Observer's choice of action together determine the Observer's *gain.* The *gain function,* $G(\alpha, \theta)$, is simply a tabulation of the gain corresponding to any combination of World state and action.

If the current state of the world, θ_*, were known, it would be a very simple matter to find an action α_* that maximized $G(\alpha_*, \theta_*)$, the gain to the Observer (there may be several actions that each maximize gain). We will assume that the Observer does not have direct knowledge of the current state of the World and must select an action without knowing precisely what gain will result. The framework developed so far is that of Game Theory, and any text on Game Theory contains descriptions of strategies to employ when we have no information about the current state of the World (for example, Williams, 1954).

Within the framework of Statistical Decision Theory, the Observer has additional, imperfect information about the current state of the World, in the form of a random variable, X, whose distribution depends upon it. The random variable, X, which serves to model sensory input, can only take on values[3] in a set of *sensory states,*

$$X = \{\chi_1, \chi_2, \ldots, \chi_p\}. \tag{6.3}$$

Again, each of the sensory states can be a vector. For example, the current sensory state could comprise the instantaneous excitations of all of the retinal photoreceptors of an organism. The probability of taking on any particular value χ_* during the current turn depends, at least formally, on the current state of the world, θ_*. The *likelihood function*,

$$\lambda(\theta_i, \chi_j) = P[X = \chi_j \mid \theta_* = \theta_i], \tag{6.4}$$

serves as a summary of these conditional probabilities. I will refer to the current value of X on the current turn as the Observer's *current sensory state*, denoted χ_*.

On a single turn, the only useful information available to the Observer about the current state of the World is the sensory state, the value of the random variable X. I will assume, for now, that his choice of action, α_*, is completely determined by this current sensory state, χ_*. We can then write,

$$\alpha_* = \delta(\chi_*), \tag{6.5}$$

a *(deterministic) decision rule*.[4] Since χ_* is a random variable, so is α_*, and so is the gain, $G(\delta(\chi_*), \theta_*)$.

There are n^p possible, distinct decision rules (each of the p sensory states can be mapped to any one of n possible actions). Because there are only a finite number of possible World states, possible Sensory states, possible Actions, and possible decision rules, I will be able to present SDT in a very straightforward and intuitive manner. In the section titled "*The continuous case*" (p. 158), I'll describe how the basic mathematical results of the theory change once we abandon the assumption of finiteness.

For any given choice of rule, we can compute the *expected gain* (EG)[5] for any particular state of the World, θ,

$$EG(\delta, \theta) = \sum_{j=1,2,\ldots,p} \lambda(\theta, \chi_j) G(\delta(\chi_j), \theta). \tag{6.6}$$

The expected gain depends upon both the decision rule, δ, and the unknown state of the World, θ. SDT assumes that all preferences among rules are determined by the expected gains. Statistical Decision Theory is the study of how to choose a good decision rule, δ. Its elements are summarized in Figure 6.1. We next consider criteria that help us decide which rules are "better" than others.

DOMINANCE AND ADMISSIBILITY

For now, let's assume that there are only two possible World states, θ_1 and θ_2. Each of the points in Figure 6.2 corresponds to a decision rule. For each rule, δ, the expected gain $EG(\delta, \theta_1)$ in World state θ_1 is plotted on the horizontal axis versus the expected gain $EG(\delta, \theta_2)$ in World state θ_2. I'll refer to this point as the *gains point* corresponding to the rule. Of course, two rules may share a single gains point if they result in identical expected gain in each World state. I'll sometimes refer to "the rules corresponding to a particular gains point" or "a rule corresponding to a particular gains point". If there were more than two World states we would add dimensions to this *gains plot*, but each decision rule would still map to a single point in this higher dimensional space.

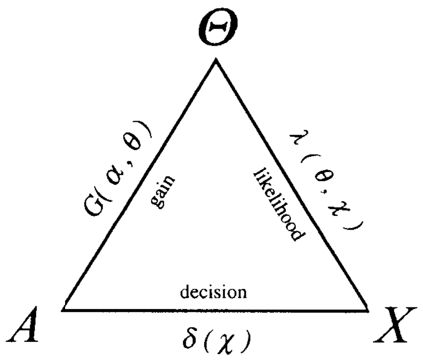

Figure 6.1 *The elements of Statistical Decision Theory.* The three sets at the vertices are Θ, the possible *states of the World*, X, the possible *sensory states*, and A, the available *actions*. The three edges correspond to the gain function, G, the likelihood function, λ, and the decision rule, δ.

Next, consider the expected gains obtainable from each of the three rules labeled δ_1, δ_2, and δ_3. Expected gain increases as we go up or to the right in the graph in Figure 6.2. Examining the rules, it is clear that some of them have higher gain than others, independent of the state of the World. Rule δ_3, in particular, is a sad creature. No matter what the state of the World, rule δ_1 has a higher expected gain than rule δ_3. Rule δ_1 is said to *dominate* rule δ_3 and it is evident that rule δ_3 should never be employed if rule δ_1 is available. The exact definition of dominance is slightly more complicated: one rule is said to dominate a second if its expected gain is never less than that of the second rule for any state of the World, and is strictly greater for at

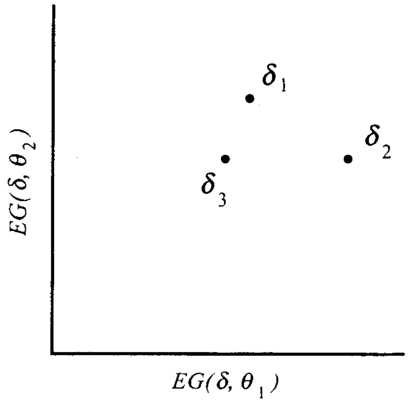

Figure 6.2 *A gains plot.* For any decision rule δ, we plot its expected gain in the first World state, $EG(\delta, \theta_1)$, on the first axis, its expected gain in the second World state, $EG(\delta, \theta_2)$, on the second, etc. The resulting *gains points* for three rules are shown, labeled δ_1, δ_2, and δ_3. The plot shown is two-dimensional and, consequently, can only correspond to a World with two states. If there are m World states the gains plot will be m-dimensional.

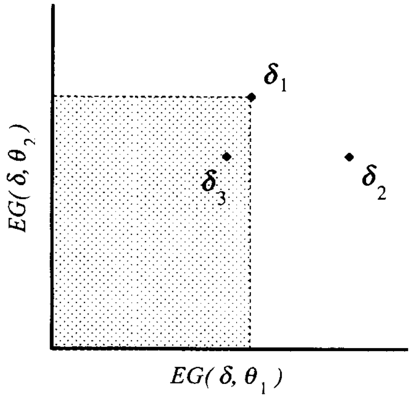

Figure 6.3 *Dominance.* Examples of a *dominated* rule, δ_3, and two *admissible* rules δ_1 and δ_2 in a gains plot are shown. The rule δ_3 is dominated by the rule δ_1 since the latter has a higher expected gain in both World states. It is also dominated by the rule δ_2 since the latter has the same expected gain in one World state and a strictly higher expected gain in the other. The *dominance shadow* of δ_1, enclosing δ_3, is shown. A rule whose gains points fall in this region (including the edges but *not* the Northeast vertex) is dominated by rule δ_1.

least one state of the World. By this definition, rule δ_2 dominates rule δ_3, even though the two rules have the same expected gain in World state 2. The dotted lines in the gains plot in Figure 6.3 sketch out the "dominance shadows" of one of the rules. Any rule falling in the dominance shadow of a second is dominated by it.

Rule δ_2 does not dominate rule δ_1 in Figure 6.2, nor does rule δ_1 dominate rule δ_2. Rules that are dominated by no other rule are *admissible rules* (Ferguson, 1967). The wise decision maker, in choosing a rule, confines his attention to the admissible rules.

MIXTURE RULES

Given two rules, δ_1 and δ_2, we can create a new rule d by *mixing* them as follows: "Given the current sensory state χ_*, take action $\delta_1(\chi_*)$ with probability q, otherwise take action $\delta_2(\chi_*)$." The new rule is an example of a *randomized decision rule* or *mixture rule.* We will use the letter d to denote such rules. From now on I'll refer to the original, non-randomized decision rules as *deterministic rules.* While there are only finitely many deterministic rules, the mixture of any two of them results in infinitely many randomized decision rules, one for each value of q. The application of a mixture rule to the current sensory state is, accordingly, denoted $d(\chi_*)$.

When the mixture probability q is 1, the resulting mixture rule is identical to the deterministic rule $\delta_1(\chi_*)$, so we can regard all of the deterministic rules as mixture rules as well. The expected gain of a mixture rule is easily computed,

$$EG(d, \theta_*) = qEG(\delta_1, \theta_*) + (1-q)EG(\delta_2, \theta_*) \qquad (6.7)$$

that is, one may expect to receive the expected gain associated with rule δ_1 with probability q, and otherwise, with probability 1-q, the expected gain associated with rule δ_2. Further, it is permissible to mix mixture rules to get new mixture rules.

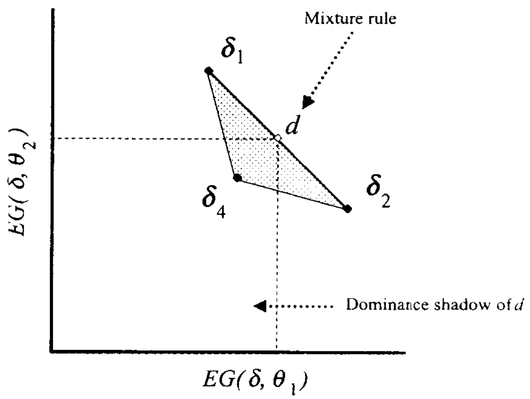

Figure 6.4 *Mixture rules.* The upper-right edge of the shaded triangle contains the gains points for all the randomized decision rules resulting from probabilistic mixtures of the deterministic rules, δ_1 and δ_2. The triangle contains the gains points of all randomized rules resulting from probabilistic mixtures of δ_1, δ_2, and δ_4. The rule δ_4 is dominated by several of the rules resulting from mixtures of δ_1 and δ_2, and one dominance shadow containing δ_4 is shown. Note that δ_4 is dominated by a mixture of δ_1 and δ_2 but not by either δ_1 or δ_2 alone.

The graphical representation of mixture rules is very simple: as q is varied between 1 and 0, the gain points corresponding to the new mixture rules fall on the line segment joining the points corresponding to δ_1 and δ_2 (see Figure 6.4). If we mix the new mixture rules corresponding to points along this line segment with the point labeled δ_4, the resulting points fill a triangle with vertices labeled δ_1, δ_2, and δ_4. These are precisely the expected gains in the two World states that can be achieved, given the three deterministic rules and all their mixtures. Note that δ_4 is not dominated by either of the deterministic rules δ_1 or δ_2 but is dominated by a mixture of the two. The dominance shadow of one of the mixtures that dominates δ_4 is shown in Figure 6.4.

The shaded area in any gains plot (the *region of achievable gains*) will always be convex,[6] and the admissible rules will correspond to points along its upper-right frontier. The admissible rules in the gains plot in Figure 6.4 are precisely the mixtures of δ_1 and δ_2, including, of course, the two deterministic rules themselves.

From now on, the term "rule" will be used to refer to both mixture rules and deterministic rules considered as special cases of mixture rules. The letter used to denote a rule will typically be d.

EXAMPLE: THE THEORY OF SIGNAL DETECTABILITY

Consider a very simple perceptual task: The World is in one of two states,

$$\Theta = \{\text{SIGNAL, NO-SIGNAL}\} \qquad (6.8)$$

and the Observer has two possible actions,

$$A = \{\text{SAY-YES, SAY-NO}\} \qquad (6.9)$$

The sensory information X available to the Observer has different distributions, depending on the state of the World, known in the Theory of Signal Detectability (TSD)[7] as the *signal + noise distribution* and the *noise distribution*. These two distributions, taken together, determine the likelihood function introduced above. Much work in TSD theory begins with an explicit assumption concerning the parametric form of the signal + noise and noise distributions (Green & Swets 1966/1974; Egan, 1975) but the particular choice of distributions is not relevant to this example.

TSD can be treated as an application of SDT (Statistical Decision Theory) to the simple problem just outlined (Green & Swets, 1966/1974; Egan, 1975).[8] We can define the gain function as follows: one unit of gain results if the state of the World is SIGNAL and the action selected is SAY-YES or if the state of the world is NO-SIGNAL and the action selected is SAY-NO. Otherwise the gain is 0. The expected gain is easily computed: when the state of the World is SIGNAL, it is the probability of SAY-YES, when the state of the World is NO-SIGNAL, it is the probability of SAY-NO. In the standard terminology of TSD, these two probabilities are referred to as the HIT rate, denoted H, and the CORRECT-REJECTION rate, denoted CR. Figure 6.5 is the gains plot for this version of TSD. The convex shaded area corresponds to the gains achievable by any possible rule. The admissible rules fall on the darkened edge facing up and to the right, as shown.

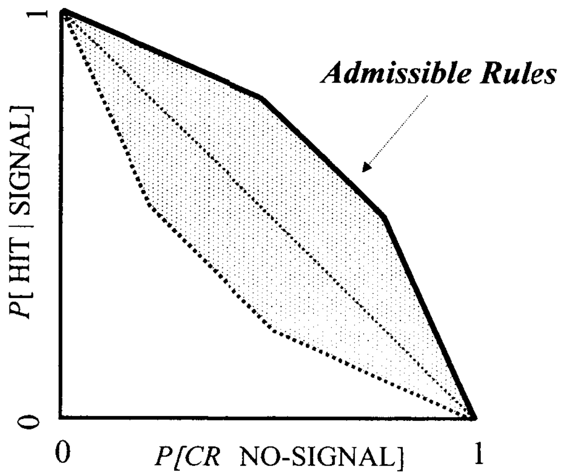

Figure 6.5 *A gains plot for a version of the Theory of Signal Detectability.* The two possible World states are SIGNAL and NO-SIGNAL, and the expected gains are the probability of saying YES when the World state is SIGNAL (a "Hit" in the terminology of TSD) and the probability of saying NO when it is NO-SIGNAL (a "Correct Rejection"). The gains points corresponding to admissible rules (bold solid contour) form the "ROC curve" of TSD, reflected around the vertical axis (in TSD it is customary to plot the False Alarm rate on the horizontal axis, not the Correct Rejection rate). The shape of the ROC curve depends on the choice of the underlying distributions and it may be smooth or polygonal as shown here (Egan, 1975).

Of course, the set of admissible rules is precisely the Receiver Operating Characteristic (ROC) curve[9] of TSD, slightly disguised. We plotted CR as the measure of gain along the horizontal axis where the World-state is NO-SIGNAL. In TSD, it is customary to use the FALSE-ALARM rate, denoted FA, which is just 1-CR. The net effect of this is simply to flip the normal TSD plot around the vertical axis. Viewed in a mirror, the locus of admissible rules takes on the appearance of the familiar ROC curve.

The shaded area represents all the possible observable performances for the Observer. Even if the Observer attempts to do as badly as possible in the task, for example, by replying YES when NO is dictated by an admissible rule, his performance will just fall on the mirror of the ROC curve, the locus of optimally-perverse performance. Even if he switches from rule to rule at random, his averaged performance will fall somewhere within the shaded area.

ORDERING THE RULES

The reader familiar with TSD may have remarked that we neglected to include some of the familiar components of TSD, notably the *prior probability* that a signal will occur. We will introduce such prior distributions in the next section, remedying the omission. However, it is important to realize that Statistical Decision Theory (SDT) is not limited to the case where we know the prior probability that the World is in any one of its states. It is applicable even when the World state cannot reasonably be modeled as a random variable, as, for example, when the World is another creature capable of anticipating any strategy we develop and dedicated to defeating us. It is important to understand what is gained through knowing this prior distribution and what is lost by acting as if it were known when in fact, it is not.

Note that SDT, as developed so far, cannot, in general, tell us which of two rules to choose. Only in the special case where one rule dominates the other, is it clear that the dominated rule can only lead to reduced expected gain. Although SDT cannot tell us which rule is the best, we can assume that the best rule will be an admissible rule.

We seek an ordering criterion that allows us to order the rules unambiguously, and to select the best among them. The Bayes criterion, presented in the next section, is such a criterion.

[*An aside*: The literature concerning Bayesian approaches to biological vision is almost entirely concerned with rules judged to be optimal by the Bayes criterion. The Bayes criterion can also be used to order rules (and visual systems) that are distinctly sub-optimal. We'll return to this point in later sections.]

There are plausible criteria for ordering the rules other than Bayes. The remainder of this section concerns a second ordering criterion, the Maximin criterion. The Maximin criterion of Game Theory (von Neumann & Morgenstern, 1944/1953) assigns to each rule its "worst case" gain: for any rule δ, the *Maximin gain* is,

$$\mathrm{Mm}(\delta) = \min_{\theta\in\Theta} EG(\delta, \theta). \qquad (6.10)$$

The expected gain for rule δ can be no less than Mm(δ), no matter what the state

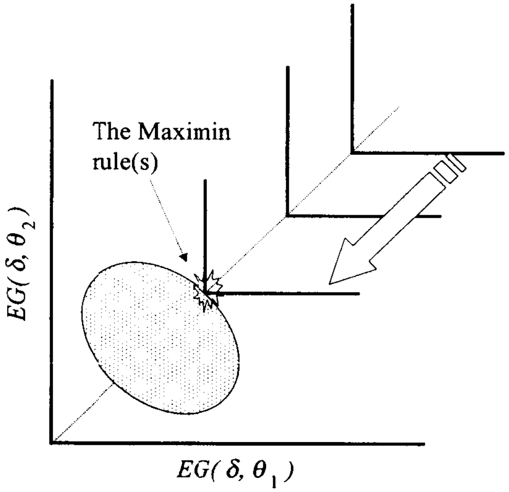

Figure 6.6 *Graphical computation of the gains point corresponding to the Maximin rule.* The "wedge" slides along the 45-degree line from upper-right to lower-left until it just touches the convex area at the point surrounded by a "blast". Any rule with this gains point is a Maximin rule.

of the World. The rules are now ordered by their Maximin gains, and the rule with largest Maximin gain is a *Maximin rule* (whose Maximin gain is the *max*imum of the *min*ima of the gains of all the rules).

The gains plot in Figure 6.6 serves to illustrate how a Maximin rule can be defined graphically. The right-angled wedge "slides down" the 45 degree line until it first touches the convex set of gains. Any rule corresponding to this point (there may be several) is a Maximin rule. If you compare the gains for a Maximin rule to that for any of the other possible rules, you will find that in at least one World state, the second rule would do worse.

The Maximin criterion is particularly appropriate when the Observer faces an implacable, omniscient World, capable of anticipating the strategic options of the Observer and taking advantage of any error. The Maximin Observer is guaranteed the Maximin gain no matter what the World chooses to do. Should the World prove to be a bit dim, indifferent, or even benevolent, the Maximin Observer can only do better.

The Maximin criterion can, of course, be used to order any two rules, admissible or not. Graphically defined, the better rule is the one whose gain point in the gains plot (Figure 6.6) first[10] touches the sliding wedge as it goes from upper-right to lower-left.

Savage (1954) criticizes the use of the Maximin criterion, notably its excessive pessimism. In particular, the Maximin Observer makes no use of any non-sensory information he may have about the state of the World. Of course, this in itself is not unreasonable. If, for example, the World is an intelligent opponent, there is reason for him to act "improbably" precisely so as to gain an advantage. Little Red Riding Hood had an accurate prior belief that wolves were not often present in Grandmother's house, and certainly not in Grandmother's bed. The Wolf took advantage of her prior belief.

PRIOR DISTRIBUTION AND BAYES' GAIN

Like Maximin Theory, Bayesian Decision Theory (BDT) provides a criterion for imposing a complete ordering on all rules, specifying when two rules are *Bayes-equivalent*, and otherwise which of the two is the better. It can also tell us which, of all the rules, is the best. In making the transition to Bayesian theory, we must first assume that the current state of the World is drawn at random from among the possible states of the World, Θ: the Intelligent, Malevolent World of Game Theory has been reduced to a set of dice. The probability that state θ is picked is denoted $\pi(\theta)$ and the probability mass function $\pi(\theta)$ is referred to as the *prior distribution* or simply the *prior* on the states of the World.

The ordering principle inherent in BDT is based on the *expected Bayes' gain* of each rule d, defined as,

$$EBG(\delta) = \sum_{\theta \in \Theta} \pi(\theta) EG(\delta, \theta), \tag{6.11}$$

The expected Bayes' gain is the "Expected Expected Gain", averaging across the states of the World. According to the Bayes criterion, one rule is better than a second if its *EBG* is greater. Any rule with the maximum *EBG* is a *Bayes' rule*.

The graphical definition of the Bayes' rule is particularly pleasing. Consider, in Figure 6.7, the solid line that passes through the points (0, 0) and $(\pi(\theta_1), (\pi(\theta_2))$, the *prior line*. The dashed lines are perpendicular to the prior, and the points on each of these dashed lines satisfy,

$$EBG(\delta) = \pi(\theta_1)EG(\delta, \theta_1) + \pi(\theta_2)EG(\delta, \theta_2) = \text{constant} \tag{6.12}$$

These are the *lines of constant (expected) Bayes' gain*[11]; any two rules that fall on a single line of constant Bayes' gain have the same Bayes' gain: they are *Bayes-equivalent.* Bayes' gain increases as we travel up or to the right, and the optimal rules, according to the Bayes criterion, correspond to the point or points lying on the dashed line that just "touches" the upper-right frontier of the convex set of possible gains (marked in Figure 6.7). The ordering of the rules is the ordering of the lines of constant Bayes' gain.

If one accepts the assumptions of SDT and the additional assumption that there is a known prior on the (randomly-selected) states of the World, *and* if one seeks to maximize Bayes' gain, any Bayes' rule is the optimal decision rule. Please re-read the previous sentence.

The consequences of having the wrong prior distribution are very easy to visualize. In Figure 6.8 the dashed lines of constant Bayes' gain correspond to an incorrect choice of prior distribution. The solid line corresponds to the correct one The incorrect and correct maximum gains points are highlighted and a double-arrow line shows the amount of Bayes gain lost by choosing the incorrect prior.

It is also interesting to consider the relation between the ordering of rules induced by the Maximin criterion and the ordering of rules induced by the Bayes criterion for a given prior. Is there a prior distribution such that the gains point corresponding to

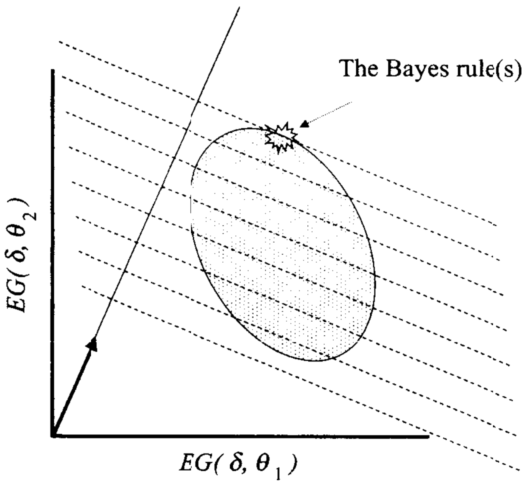

Figure 6.7 *A graphical computation of the gains point corresponding to a Bayes rule for a given prior.* (a) The bold vector has, as coordinates, the prior probabilities of the World states: $(\pi(\theta_1), \pi(\theta_2))'$. The dashed lines are all perpendicular to the prior vector: they are the *lines of equivalent Bayes' gain.* All rules whose gains points fall on the same line of equivalent Bayes' gain have the same Bayes' gain. The Bayes' gain for these lines increases as the line is further to the North and East. The gain point with the highest Bayes' gain is marked with a "blast". The rules corresponding to this gain point are the Bayes rules for this prior. (b) The same plot, but for a different choice of prior. Note that the gain point of the Bayes rules has changed.

the Maximin rule is among the gains points corresponding to the Bayes' rule? The answer is yes: there is always[12] a choice of prior such that the gains point for any admissible rule is among the gains points of the Bayes' rule for that prior (Ferguson, 1967). As the Maximin rule is admissible, there must be a choice of prior that results in a Bayes point that has the same gains point as the Maximin rule.

This prior is sometimes, but not always, the *maximally-uninformative or uniform prior* that assigns equal probability to every World state, but it need not be. Figure 6.9 contains three diagrams, the first illustrating a case where it is, and two where it is not.

[*A caution:* When there are more than two World states, the geometric version of the Bayesian approach remains valid. The *prior line* remains a line, but the perpendicular lines of constant Bayes' gain become planes or hyper-planes. The ordering of these planes along the prior line induces the ordering of the rules.]

THE CONTINUOUS CASE

I've presented SDT and BDT in the special case where the number of possible World states, possible Sensory states, and possible actions are all finite. So soon as this finiteness assumption is abandoned, both the derivation and presentation of the basic results of the theory become difficult. Remarkably, the basic geometric intuitions remain more or less intact, even when the gains plot becomes infinite-dimensional. The corresponding proofs become difficult and non-intuitive, and center on issues of existence. Is there always one or more admissible rules, a Maximin rule, a Bayes' rule for every prior? ("No", "No", "No".) Even when a Bayes' rule does not exist, we can typically find rules that, although they are not admissible, come as close as we like to the performance of the non-existent Bayes' rule. Ferguson (1967) presents this difficult material very clearly with constant reliance on geometric intuition accumulated in consideration of the finite case.

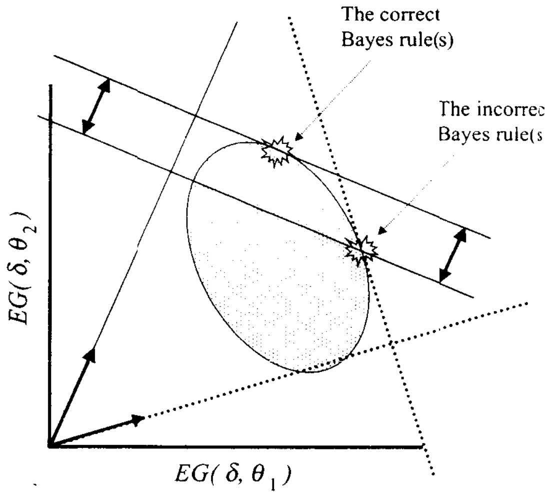

Figure 6.8 *Consequences of selecting an incorrect prior.* The dotted lines mark the location of the gain point of the Bayes rules for a prior distribution that is incorrect. The lines of equivalent Bayes' gain for the true prior are shown as solid line. Note that the Bayes rules for the incorrect prior share a Bayes' gain strictly inferior to the Bayes' gain for the *true* Bayes' rules. The double-headed arrow marks the cost of having incorrect prior information.

To translate from the finite case to what I will refer to as the continuous case, we need only change the notation above slightly. Recall that θ, χ, and α were potentially vectors above, something we made no use of (and will make no use of). The sets Θ, *A*, and *X* are subsets of real vector spaces of possibly different dimensions, the gain function *G* (α, θ) is defined as before, and the likelihood function λ(θ, χ) is a probability density function on χ for any choice of the world state θ (it is a parametric family of probability density functions with parameter θ). The summation signs in the finite case are replaced by integrals. Expected gain (Equation 6.6) becomes (replacing the notation for a deterministic decision rule δ by that for a randomized rule *d*),

$$EG(d, \theta) = \int \lambda(\theta, \chi) G(d(\chi), \theta) d\chi, \tag{6.13}$$

a multidimensional integral if *X* is a subset of a multidimensional real space.

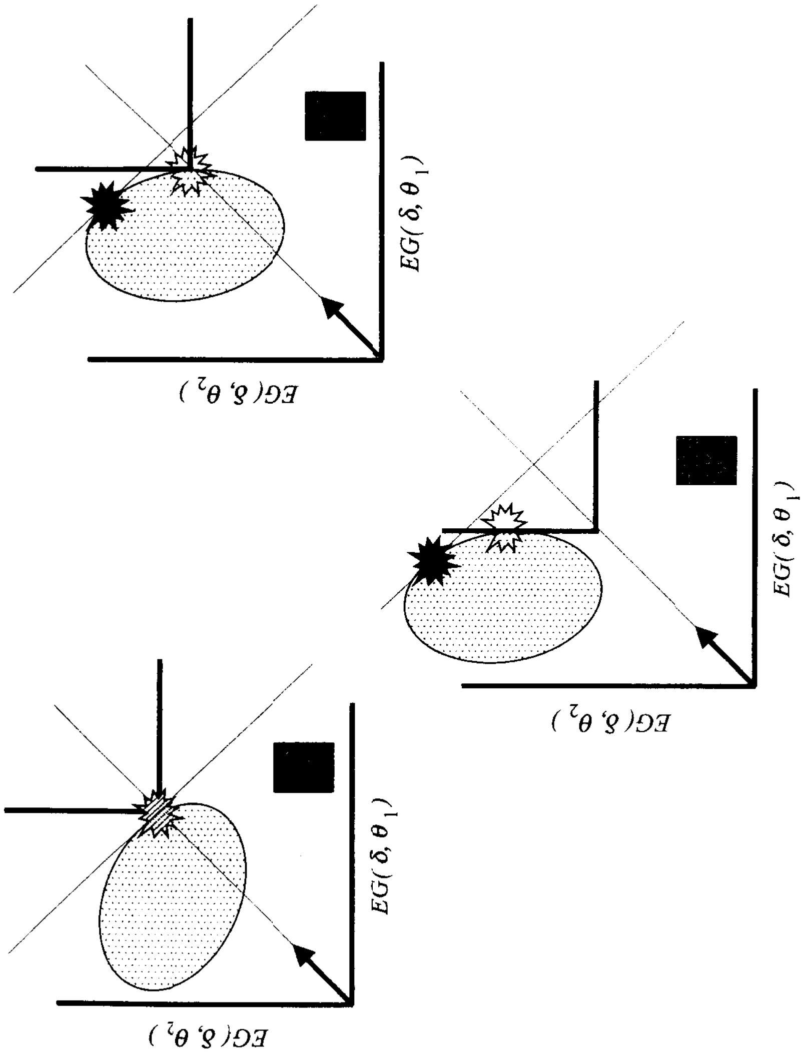

Figure 6.9 *Maximin and Bayes.* The Maximin rule has the same gains point (marked with a starburst) as the Bayes' rule with a uniform (*maximally-uninformative*) prior. (b) and (c). The prior of the Bayes' rule corresponding to the Maximin rule is not uniform. The gains point corresponding to the Maximin rule is marked with a white starburst, the gains point corresponding to the Bayes' rule with uniform prior is marked with a shaded starburst.

Expected Bayes' gain (Equation (6.11)) becomes,

$$EBG(\delta) = \int \pi(\theta)EG(d, \theta)d\theta, \qquad (6.14)$$

which is a multidimensional integral if θ is multidimensional. Substituting Equation (6.13) into Equation (6.14), we find that

$$EBG(\delta) = \iint \pi(\theta)\lambda(\theta, \chi)G(d(\chi), \theta)\, d\chi d\theta. \qquad (6.15)$$

If, for a given prior, there is a rule whose expected Bayes' gain, computed by the previous equation, is greater than or equal to that of all other rules, then it is a Bayes' rule.

BAYES' THEOREM AND THE POSTERIOR DISTRIBUTION

A vision scientist familiar with the usual presentation of BDT in the vision literature may feel that something is missing. We have developed Bayesian Decision Theory without mentioning Bayes' Theorem! A version of Bayes' Theorem[13] is often the first equation to appear in a vision article concerned with Bayesian approaches. In fact, Bayes' Theorem plays a very minor role in BDT, serving only to help us develop a clever way to compute optimal rules based on Equation (6.11) or (6.14). Bayes' Theorem lets us develop a simple method for computing the rule d that maximizes,

$$EBG(\delta) = \iint \pi(\theta)\lambda(\theta, \chi)G(d(\chi), \theta)d\chi d\theta. \qquad (6.16)$$

In this section, I'll first describe how Bayes' Theorem allows us to simplify Equation (6.16). Of course, were we ignorant of Bayes' Theorem, we could still maximize Equation (6.16) numerically by choice of d (see O'Hagan, 1994, Ch. 8).

First note that the likelihood function $\lambda(\theta, \chi)$ is, within the framework of BDT, a conditional distribution $f(\chi \mid \theta)$ of the random variable χ on the random variable θ and, by a variant of Bayes' Theorem,[14] we can find probability density functions g and h such that,

$$f(\chi \mid \theta)\pi(\theta) = g(\theta \mid \chi)h(\chi) \qquad (6.17)$$

Substituting the right-hand side of Equation (6.17) into Equation (6.16), and reversing the order of integration by Fubini's Theorem (Buck, 1978), we have,

$$EBG(\delta) = \int \left[\int g(\theta \mid \chi)G(d(\chi), \theta)d\theta\right] h(\chi)d\chi \qquad (6.18)$$

The probability density function $h(\chi)$ is non-negative: to maximize the outer integral, it suffices to maximize the inner integral separately for each choice of χ, plausibly a simpler computation than the maximization of the original integral.

This method of computation, made possible by an application of Bayes' Theorem, has a straightforward interpretation. Once the Observer, following a Bayes' rule, has learned the current Sensory state χ_*, he effectively forgets that there were ever alternative outcomes for the Sensory state and chooses his action α so to maximize,

$$\int g(\theta \mid \chi_*) G(\alpha, \theta) d\theta \tag{6.19}$$

the expected gain with respect to the *posterior distribution* $g(\theta \mid \chi_*)$ on θ. At this point, the current sensory state (or rather, its realization) χ_* is known, non-stochastic. We can interpret the posterior distribution as an updated prior distribution and, arguably, the Observer should use it, rather than the prior on a subsequent turn, all else being equal. This use of the posterior as the new prior is a controversial aspect of Bayesian theory, and I'll return to it in the third section.

SDT AS A MODEL OF BIOLOGICAL VISUAL PROCESSING

> *This model will be a simplification and an idealization, and consequently, a falsification. It is to be hoped that the fectures retained for discussion are those of greatest importance in the present state of knowledge.*
>
> A.M. Turing (1952) *The Chemical Basis of Morphogenesis*

Let us distinguish two possible applications of instantaneous BDT to biological vision. We could, first of all, use SDT/BDT to model the instantaneous visual environment of an Observer, making no claims about how the Observer processes visual information. Most psychophysical experiments are instantiations of instantaneous *Bayesian environments* designed by an Experimenter: there is a well-defined and typically small set of possible world states with specific prior probabilities and a limited set of actions available to the Observer, etc. The Experimenter takes care that the state of the World on any trial is a random variable, independent of the state of the World on other trials. We could apply the results of the previous section to compute the expected Bayes gain of an Ideal Bayesian Observer in such an experiment and compare ideal performance to the Observer's performance. This sort of application of BDT is important (Geisler, 1989; Wandell, 1995) but neither new nor controversial.

Alternatively, we could develop a model of human visual processing as a *Bayesian Observer* which,[15] given sensory information, employs BDT to maximize its expected Bayes' gain. If this Bayesian Observer had perfect information concerning the likelihood function, gain function, prior, and so forth, then it is simply the Ideal Observer just discussed. When human performance in a visual task falls short of ideal, as is typical (Geisler, 1989), the Ideal Observer is evidently an inappropriate model for human visual processing.

In this section, I consider, as candidate models of human visual processing, Bayesian Observers that have less than perfect information concerning the gain function and prior probabilities. (We could also consider Bayesian Observers that have less than perfect information concerning the other elements of SDT/BDT such as the likelihood function, but we will not do so here.)

In the previous section we saw that the Bayes criterion not only allows us to determine which rules are optimal (the Bayes rules) but also how to order all rules, optimal or not. In Figure 6.8 we saw how an incorrect choice of prior affects the expected Bayes' gain of an otherwise optimal Bayesian Observer, and we can similarly evaluate the consequences of choosing an incorrect gain function. In brief, within the framework described in the previous section, we can analyze and compare the performances of all Bayesian Observers from bad to (nearly) ideal. Such *Non-ideal Bayesian Observers* would seem to be the obvious candidate models of biological visual processing that could properly be called "Bayesian". This section comprises an analysis of the components of *Bayesian Observers*, ideal and non-ideal.

NON-IDEAL BAYESIAN OBSERVERS

> *The drop over the edge at the top is fatal but the views are splendid.*
>
> R. Maqsood (1996) *Petra: A Traveler's Guide*

Hesitating a few steps away from Maqsood's cliff, bordering the High Place of Sacrifice,[16] straining to see over the edge, even the sworn Bayesian may be allowed to doubt whether he has correct estimates of the instantaneous gain function and the prior distribution on friction coefficients of sandstone. He may also feel that, however much he trusts the prior distribution and gain function that allow him to navigate the streets of a city, the current situation requires something else.

Bayesian Observers choose actions by Bayesian methods, specifically by maximizing expected Bayes' gain given information about the environment encoded as priors, gain functions, etc. The Observer's environment is assumed to be a Bayesian environment with a well-defined set of World states, possible actions, and so forth. The prior distribution and the gains function, in particular, are objective, measurable parts of this environment just as much as the intensity of illumination. In this section, as in the previous section, I'll consider only a single instant of time and the action to be chosen at that instant of time.

The Ideal Bayesian Observer is assumed to have the correct values of all of the elements of SDT in Figure 6.1 and, in addition, the correct prior. In this section, we consider Bayesian Observers whose information about the prior distribution and the gains function may not be the true prior or gain function of the Environment. To raise this issue, requires a small change in notation. In addition to the objectively correct components of SDT (Figure 6.1) and the prior of BDT, which accurately describe the Bayesian environment, we have the corresponding elements available to the *Bayesian Observer*: a gain function $\tilde{G}(\theta, \alpha)$, and a prior distribution, $\tilde{\pi}(\theta)$. The tilde over each symbol indicates that the element belongs to the Observer and need not be the same as the corresponding Environmental element.

THE PRIOR FUNCTION AS "PROBABILITY ENGINE"

A recurring criticism of Bayesian approaches to modeling biological vision is that, in even very simple visual tasks, the number of possible states of the world is large, and

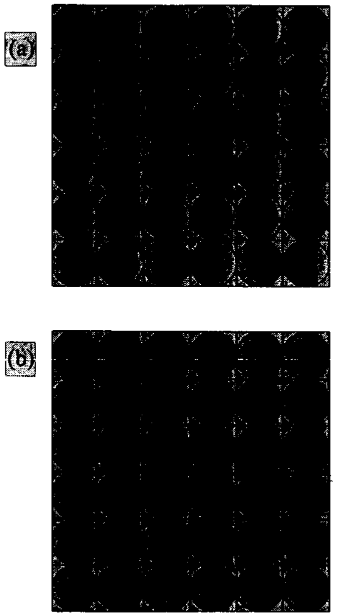

Figure 6.10 *Patterns and probabilities.* A Bayesian Observer, designed to model pattern vision, must assign probabilities to very large numbers of patterns including the "checkerboard patterns" illustrated here. The sheer number of such patterns guarantees that almost all of them have never been seen by a human observer before. The demands of the Bayesian formalism insist that a non-zero probability be assigned to such a pattern if it is to be seen at all.

it is difficult to imagine how a biological organism comes to associate the correct prior probability with each state. The states of the world in a visual task might correspond to all possible arrangements of surfaces in a scene, and is difficult to see how a visual system could acquire or encode all of these prior probabilities (Shimojo & Nakayama, 1992). Of course, we are considering non-ideal as well as ideal Bayesian Observers and we need not demand that the organism arrive at exactly the correct prior probability for each state.

Even for relatively simple visual tasks, the number of possible World states can be very large. Consider, for example, stimuli like the one in Figure 6.10(a), $N \times N$ checkerboards, with a single "item" placed in each square and where each "item" has only one of two possible states. There are 2^{N^2} possible states in this very simple World and it would be easy to envision using such stimuli in psychophysical experiments. Yet the number of "states" in this simple world (with $N = 6$) is about 69 billion. For comparison, the number of seconds in the nominal 70-year life span is about 2.2 billion. It would take about two millennia, viewing these patterns at the rate of one per second, to encounter them all just once. It is implausible that a prior probability distribution on such patterns could be based on frequency counts of the occurrences of such patterns.

Yet these patterns form a very small proportion of the patterns that a prior distribution for a "pattern vision" Bayesian Observer should encompass. The checkerboards have an evident interpretation as "shape-from-shading" stimuli, and thus all of these stimuli fall within the domain of a Bayesian Observer model devoted to "shape from shading." Yet how could these probabilities even be stored in a nervous system? If the checkerboard is expanded to 8×8, there are more than 10^{19} patterns possible, a number larger by far than either the number of neurons or the number of synapses in the human brain. We might also wish to save a bit of brain for something besides storage of checkerboard pattern priors.

So long as we continue to think in terms of explicit storage of learned estimates of prior probabilities, the objection of Shimojo and Nakayama is unanswerable. There are too many things we *might* see, and every evaluation of expected Bayes' gain (Equation (6.11) or Equation (6.15)) involves the prior probability of every one of them. Even if we somehow decided to ignore most of the patterns in evaluating Equation (6.11) (or its continuous version, Equation (6.15)), we certainly must include the prior probability for the pattern we in fact see in Figure 6.10(a). Yet that pattern could have been any one of the 2^{N^2} possible patterns.

In the fourth section, I address the apparently overwhelming computational demands of Equations (6.11) and (6.15) and suggest that they are illusory. Yet it seems inescapable that a Bayesian Observer, even one that is specialized to be a model of just pattern vision, must be able to assign probabilities to very large numbers of possible visual outcomes, almost all of which it has never seen and almost certainly will never see.

It seems an inescapable implication of the Bayesian approach that the visual system assign probabilities to large numbers of possible scenes (or components of scenes) and that these probabilities affect visual processing. The mechanism of assignment of probabilities to scenes I'll refer to as a *Probability Engine*, for concreteness. The Probability Engine of a Bayesian Observer corresponds to $\tilde{\pi}(\theta)$ in the mathematical formulation.

Consideration of the analogous problem in human judgment and decision making is instructive. Given a sentence that describes a state of affairs in the world such as (A) "Boris Yeltsin is roller-blading in Red Square", can you assign a probability to it? You may feel that, although you have a consistent assignment of probabilities to events including the one just described, you cannot come up with a number that you could write down or say out loud. You may be capable of reasoning with such probabilities, but are unable to turn them into numerical estimates on demand. Even so, there are several alternative ways for me to test whether you can coherently assign probabilities to events.

Suppose that you agree that you can order events according to their probabilities: given any two events, you can tell me which of the two is more probable. Given two sentences, the one above, and the alternative, (B) "Boris Yeltsin is asleep", you can order them by probability. Given only your ordering responses, I cannot reconstruct the probabilities you assign to these events, but I can test whether you are assigning probabilities in a way that is consistent with probability theory. Consider a third event, (C) "Boris Yeltsin is secretly married to Madonna". You assert that B is more probable than C and that C is more probable than A. Next I ask you to compare B and A. If you respond that A is more probable than B, then your pattern of responses is inconsistent. If P[B] > P[C] and P[C] > P[A], then it is not possible that P[B] > P[A]. I have tested, and rejected, the hypothesis that your orderings are consistent with any pattern of underlying probabilities assigned to events.

The probabilities assigned by the Probability Engine of any Bayesian Observer must also conform to the axioms of probability theory. This constraint can provide the basis for the sort of empirical test just outlined. If we can design experiments that plausibly allow us to infer which of any pair of patterns drawn from the class of patterns illustrated in Figure 6.10(a) is assigned a higher probability by the visual system, then we can test the transitivity property just discussed (For any three events, A, B, and C, P[A] > P[B] and P[B] > P[C] implies that P[A] > P[C]).

We can test other implications of probability theory. To return to the case of human conscious judgment and decision making, the sentence "Boris Yeltsin is roller-blading somewhere" must *not* be ranked as less probable than the sentence "Boris Yeltsin is roller-blading in Red Square". The former event, includes the latter and, by elementary properties of probability, cannot be less probable. Yet previous research suggests that, for at least some pairs of events, human judges fail precisely this kind of test (Tversky & Kahneman, 1980; see also Nisbett & Ross, 1982). It is certainly of interest, given an experimental situation where perceptual prior probabilities can be ordered, to determine whether this essential Bayesian assumption holds up.

If we can develop experimental methods that allow us to estimate not only the ordering but also the difference or ratio between pairs of events, then we can develop correspondingly more powerful tests of the claim that visual modules combine evidence according to the axioms of probability theory (Edwards, 1968; Krantz et al., 1971).

A "pattern vision" Bayesian Observer, then, must assign coherent probabilities to Figure 6.10(a) and also to the highly-regular Figure 6.10(b). If the Bayesian approach to biological vision is taken seriously, then it becomes of some importance to understand how these probabilities are generated, and it is plausible that the presence or absence of subjective patterns may influence the assignment of probabilities.

Research concerning human conscious judgment of the probabilities of patterns is perhaps relevant. In reasoning about sequences arising from independent tosses of a "fair coin" ($P[H] = P[T] = 1/2$), human judges consistently judge the sequence HHHHHH to be less probable than the sequence HHTHTH (Kahneman & Tversky, 1972; Nisbett & Ross, 1982). Of course, for a fair coin, any sequence of six tosses is as likely as any other and the human judges have gotten it wrong once again. It is plausible that the judges are responding to patterns (or the absence of patterns) in the coin toss sequences, assigning lower probability to patterned outcomes. If this were so, then it suggests that a mechanism for assigning probabilities to visual patterns is not completely unreasonable.

It would certainly be of interest to determine whether the prior probabilities assigned by a pattern vision Bayesian Observer to the patterns in Figure 6.10(a) and (b) and other patterns of this sort and try to understand how a Probability Engine assigns probabilities to never-before-encountered stimuli.

The ability to reason and judge the possible sequences resulting from successive, independent tosses of a "fair coin" itself presupposes something like a Probability Engine in cognition. It is unlikely that you have ever encountered a "fair coin": ". . . whenever refined statistical methods have been used to check on actual coin tossing, the result has been invariably that head and tail are not equally likely" (Feller, 1968, p. 19). Feller argues that a "fair coin" is a model, an idealization: ". . . we preserve the model not merely for its logical simplicity, but essentially for its usefulness and applicability. In many applications it is sufficiently accurate to describe reality." (Feller, 1968, p. 19). Just as the mathematical idealization called a "fair coin" can assign probabilities to never-before encountered coin-toss sequences, so a Probability Engine assigns probabilities to scenes. They need not be precisely correct, only useful.

THE LIKELIHOOD FUNCTION AND THE LIKELIHOOD PRINCIPLE

The likelihood function serves two roles in SDT and BDT. First of all it summarizes what we need to know about the operating characteristics of the sensors that provide information about the state of the World. Second of all, once the current sensory state χ_* is known, the likelihood function $\lambda(\theta, \chi_*)$, as a function of θ, is precisely what the Bayesian Observer knows about the state of the World. At first glance, it might seem that we would be better off retaining the actual sensory data χ_* rather than running the risk of losing information by discarding it and retaining only the likelihood function. Or perhaps it would be better to supplement the likelihood function with additional measures derived from the data.

It turns out that we lose *no* information about the state of the World when we replace the raw sensory data by the likelihood function, a remarkable result known in statistics as the *Likelihood Principle*: "All of the information about θ obtainable from an experiment is contained in the likelihood function for θ given X" (Berger & Wolpert, 1988, p. 19).

Probably every psychophysicist who collects psychometric data by an adaptive psychophysical procedure such as a "staircase" method has wondered whether it could really be correct to fit the resulting data exactly the same as if it had been collected by method of constant stimuli. After all, using an adaptive procedure, the specific intensities presented to the observer depended on the observer's performance on previous trials. Yet the fitting procedure is exactly the same as if the experiment had chosen precisely those intensities before the start of the experiment and presented them to the Observer in some other, randomized order. The justification for computing the same maximum-likelihood estimate of a psychometric function in both cases is the *Likelihood Principle*.

The likelihood function is an example of a *sufficient statistic*, a transformation of the data that retains all of the information concerning the parameters that gave rise to the data (the World state, θ, for our purposes). Any additional information in the data that is lost, is not relevant to θ.

Suppose, for example, that we have a sample, $X_1, X_2, \ldots, X_N$, of size N from a Gaussian distribution with unknown mean μ and unknown variance σ^2. We compute the maximum likelihood estimates[17] of the unknown parameters:

$$\overline{X} = \sum_{i=1}^{N} X_i/N \quad \text{and} \quad S^2 = \sum_{i=1}^{N}(X_i - \overline{X})^2/N,$$

and consider the joint statistic $(\overline{X}, S^2)$. Given $(\overline{X}, S^2)$, how much *additional* information about μ and σ^2 is contained in the raw data, $X_1, X_2, \ldots, X_N$? The answer is: *none*. The joint statistic $(\overline{X}, S^2)$ is an example of a sufficient statistic that captures all of the information relevant to estimating the unknown parameters. Put another way, the N numbers in the original data set have been compressed to only 2 without loss of information concerning the unknown parameters. Note that we have lost information. Given $(\overline{X}, S^2)$, we cannot reconstruct the raw data when $N > 2$. We cannot even determine the order in which the data occurred. Permuting the data does not affect $(\overline{X}, S^2)$ at all and consequently no order information is preserved. What is the case is that the conditional probability distribution of $X_1, X_2, \ldots, X_N$ *given* $(\overline{X}, S^2)$ does not depend on μ or σ^2: this property is essentially the definition of a jointly sufficient statistic. Further discussions of likelihood and sufficiency can be found in Edwards (1972) and Berger and Wolpert (1988).

A visual system which retains the likelihood function, then, can do no better. Helmholtz, and especially Barlow, emphasized that neural processing concerns the representation and processing of likelihood (von Helmholtz, 1909; Barlow, 1972, 1995), a viewpoint buttressed by the Likelihood Principle.

If sensory data from multiple sources are independent, likelihood information can be readily combined across the sources. In our terminology, if the sensory data $\chi = (\chi^1, \ldots, \chi^p)$ is itself a vector, representing sensory information from p independent sensors, then the overall likelihood function is just the product of the likelihood functions based on the individual sensory data:

$$\lambda(\theta, \chi) = \prod_{k=1}^{p} \lambda(\theta, \chi^k). \tag{6.20}$$

Barlow has suggested that one of the organizing principles of neural processing is to transform sensory data so that the resulting encoding comprises many independent channels, each signaling likelihood (Barlow 1972, 1995).

GAIN FUNCTION

The choice of a gain function is, of course, important, and a non-ideal Bayesian Observer may have less than perfect information concerning the true gain function in the environment it inhabits. Freeman and Brainard (1995; Brainard and Freeman, 1997) analyze different candidate gain functions, comparing them against one another. The intent of their research is laudable, but there is a fundamental incoherence in their approach. An evident criterion for choice of a gain function is whether it reflects the true gains to the Observer: and comparing different formal gains functions to one another to see which one produces the "best" performance, judged intuitively, cannot be correct.

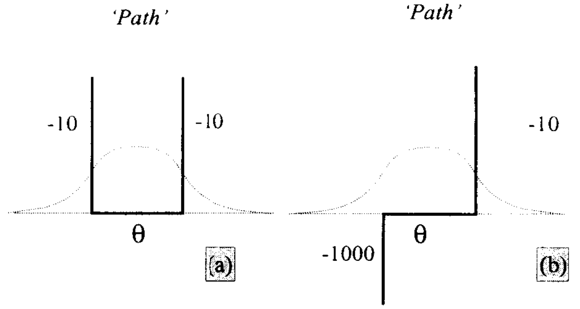

Figure 6.11 *Biases introduced by gain functions*. The figures correspond to two versions of the same visual task. The visual information is the same in both cases: a single Gaussian variable X drawn from a Gaussian distribution with mean θ. The distribution is sketched for both cases. (a) In the first the Observer must choose where to step in a path given imperfect visual information. The gain associated with going to far to right or left ("bumping into a wall") is symmetric and, across many trials the Observer's choice of step point will be symmetric about the midpoint of the path. (b) In the second version, the cost of deviations toward the left ("a sheer drop") is much greater than the cost of deviations to the right ("bumping into a wall"). The Observer's choice of step point is (correctly) biased to the right.

One possible approach would be to develop psychophysical methods that allow us to estimate the gain function of a human Observer in a particular task (just as we might estimate a contrast sensitivity function). Consideration of such empirical gain functions would give us some insight into the rewards and penalties embodied in visual processing.

The possible effect of the gain function on performance can be illustrated by a simple thought example (Figure 6.11). The visual task is to choose a location to place one's foot on a rather narrow path. There is considerable visual uncertainty concerning the location of the center of the path (perhaps it is night time) but the width of the path is known: 20 cm. The sensory data is a single random variable X, drawn from a Gaussian distribution whose mean is the center of the path, θ, and whose standard deviation is half the width of the path: 10 cm. The likelihood function is, as a consequence, a Gaussian of the same width and mean X. The maximum likelihood estimate of θ is just X, but the task is not to estimate θ: the task is to decide where to place one's foot. To make the example easier to follow, let's assume that there is no "motor" uncertainty. The foot will land wherever it is aimed. The only uncertainty in the thought example is due to the visual uncertainty surrounding the location of the center of the path.

To decide where to place the foot, we must next consider the gain function. In Figure 6.11(a), there are symmetric penalties involved in running into the two walls beside the path (-10 for running into either wall). If the likelihood function is unimodal, symmetric about its center, then a simple argument from symmetry suggests that the

Observer will place his foot in the middle of the current likelihood function, i.e. he will step on the point marked by X.

In Figure 6.11(b), however, the gain function is highly asymmetric: there is a cliff face to the right (-10 for collision) and a sheer drop (-1000) to the left. The Bayesian choice of foot placement, taking into account this asymmetry in gain, will be skewed to the right, away from the sheer drop, toward the cliff face. Again the visual information is symmetric and the bias in the response is solely due to the asymmetric gain function. I'll return to this discussion in the next section.

EXPERIMENTAL APPROACHES

The gain function $\tilde{G}(\theta, \alpha)$ and the prior $\tilde{\pi}(\theta)$ of the non-ideal Bayesian Observer are estimable psychophysical parameters, no different in kind than spectral sensitivities or contrast sensitivity. Unfortunately, we do not yet know how to design experiments so that it is possible to obtain estimates of $\tilde{G}(\theta, \alpha)$ and $\tilde{\pi}(\theta)$, directly from the data. Ramsey (1931b), von Neumann and Morgenstern (1944/1953), and Savage (1954) all developed methods that permitted estimation of subjective probability and/or subjective utility based on human performance in preference tasks. In the simple case of the Theory of Signal Delectability, prior odds and gains could, in part, be estimated from the Observer's performance once the experimenter assumed specific parametric forms for the noise and signal + noise distributions (Green & Swets, 1966/1974). The conclusions drawn were hostage to the parametric assumptions made, but it was in principle possible to separately test and verify the distributional assumptions, e.g. by consideration of the precise shape of the ROC curve (Green & Swets, 1966/1974; Egan, 1975).

Mamassian and Landy (1996; see also Mamassian et al., in press), for example, consider simple shape-from-shading stimuli where prior distributions on both the direction of illumination and on contour cues are varied independently. They are able to estimate both distributions from Observers' data with parametric assumptions on the possible priors. This sort of estimation of the components of Bayesian Decision Theory from the data would seem to be a very promising and important result of the use of BDT as a modeling framework. Of course this is only possible with strong assumptions on the possible distributions, functions, etc., that must also be independently tested.

CHALLENGES

> *Unfortunately, Bayes's rule has been somewhat discredited by metaphysical applications . . .* William Feller (1968) *An Introduction to Probability Theory*

The first part of this section contains a discussion of the status of the visual representation in an instantaneous Bayesian Observer. Simply put, most work in Bayesian vision is directed to modeling the estimation of internal representations of visual information: depth, shape, and so on. Yet SDT and BDT are theories of *preferred action*, not theories of representation, and the basis for preferring one action to another are the consequences of the actions. It is not obvious what consequences an internal visual estimate of depth can have. In this first section, I will discuss possible links between the actions of a Bayesian observer and claims about its internal visual representation, and describe how observed inconsistencies in representation inferred from different kinds of actions (Milner & Goodale, 1996) may be illusory, a simple consequence of the form of the gain function.

In the second part of this section, I will discuss the problems encountered in modeling optimal (or even "good") performance in environments where the true prior distributions and gain functions change deterministically across time, as they do in almost any environment outside the psychophysicist's laboratory.

ACTION AND REPRESENTATION

> *"How many fingers, Winston?"*
> *"Four! Four! What else can I say? Four!"* . . . .
> *"How many fingers, Winston?"*
> *"Four. I suppose there are four. I would see five if I could. I am trying to see five."*
> *"What do you wish: to persuade me that you see five, or really to see them?"*
> *"Really to see them."* George Orwell (1983) *1984*

SDT and BDT are theories of preferable actions, not representations. A moment's consideration of Winston's horrific situation, at the mercy of O'Brien in the Ministry of Love, serves to emphasize the distinction. O'Brien's goal is that Winston see five fingers when O'Brien holds up four, and his means of persuasion are most painful for Winston. It is not enough that Winston "see" four and respond five. He must "see" five, something he claims he really wants to do. The gain function here is clear, as are the possible actions, etc. If Winston were a good Bayesian Observer he would "see" five rather quickly, with a minimum of pain, something that, in the course of the novel he apparently manages to do, but only at great personal cost.

Applications of BDT to date are overwhelmingly models of internal estimates of depth, color, etc. To focus on a particular problem, let's suppose we want to estimate the distance to the edge of Maqsood's cliff. The possible answer are a range of (I would hope non-negative) real numbers, the possible states of our World. The available sensory data is the posterior distribution including the prior. We combine this posterior with a fixed gain function (I'll refer to this combination of posterior distribution and gain function as *Bayesian resolution*). Perhaps we choose a gains function that results in a least-squares estimate.

We are about to take a step and we know that errors that lead us beyond the edge of the cliff lead to very different consequences than the same magnitude of error that falls short of the edge of the cliff. The gain function we would like to combine with the posterior distribution should be highly asymmetric. Unfortunately, if we have access only to the distance representation, we have lost the posterior distribution and it is too late to compute the correct stepping distance using all of the available sensory information. The sequence of computations involved is summarized in Figure 6.12(a). The Bayesian resolution occurs before the representation, and makes use of a fixed gain function that is inappropriate to the current situation.

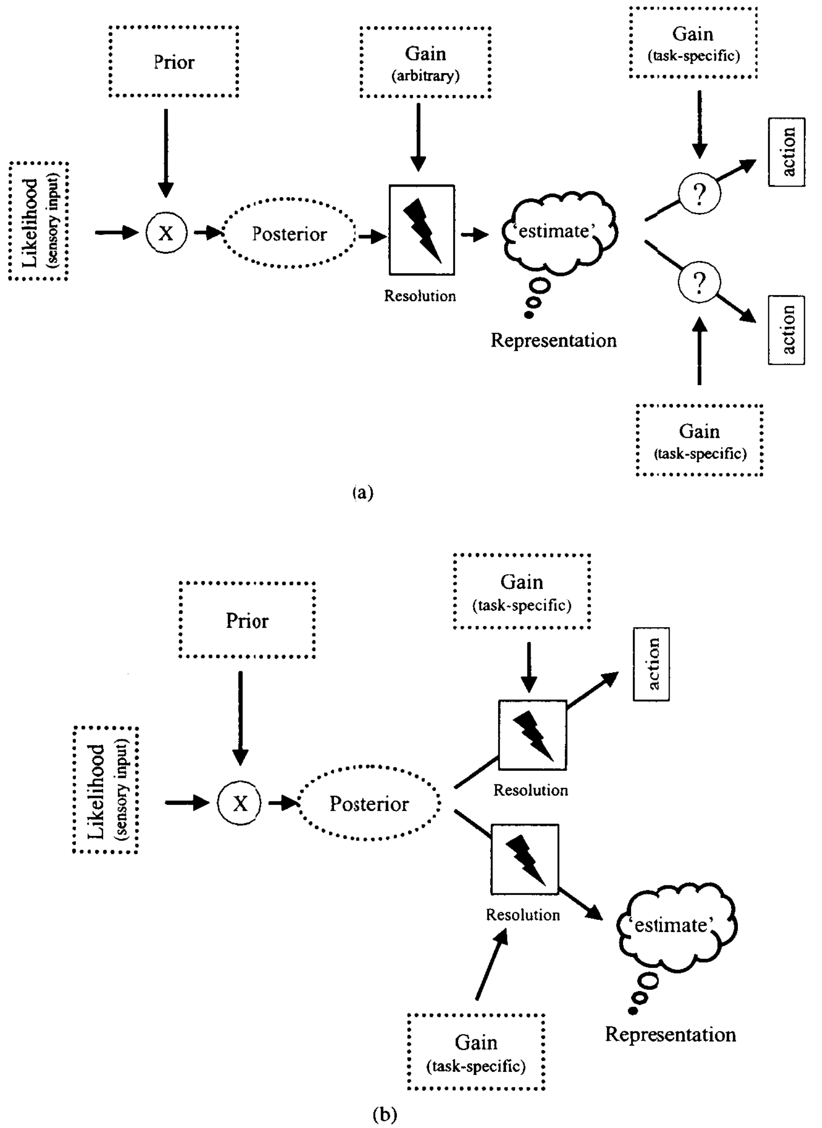

Figure 6.12 *Bayesian resolution and visual represent.* (a) Bayesian resolution occurs before the visual representation employing a fixed nominal gain function. The resulting point estimates of visual quantities such as depth must then be combined with realistic gains functions, appropriate to the current situation. The rule of combination cannot be Bayesian as the distributional information (the posterior distribution) is no longer available. (b) The visual representation is viewed as one among many different visual tasks, each of which may have a distinct gain function. (c) The visual representation is identified with the posterior distribution.

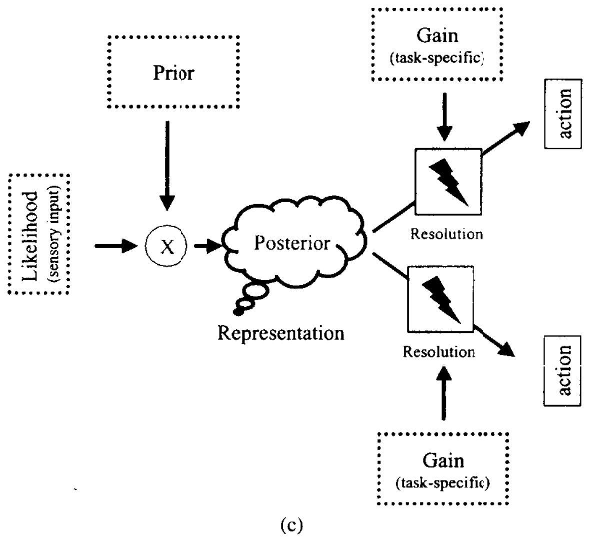

Figure 6.12 (*continued*)

A second possibility (Figure 6.12(b)) is that we combine a realistic gain function with the sensory information. If we use different realistic gain functions for determining the visual representation and for deciding how close to the edge to step, then we may exhibit an apparent dissociation between vision and action. The edge, judged psychophysically and by a stepping task, are estimated to be in different locations. It is not at all obvious what a "realistic gain function" for the representation would be: representation has no direct consequences.

Figure 6.12(c) suggests a third alternative, that what we should think of as the representation is itself the posterior distribution, or at least includes the posterior distribution. The advantage is obvious: we postpone the Bayesian resolution until after the representation and we can now use a more flexible gain function, possibly one that represents gain near the edge of a cliff or in the hands of O'Brien. There is however a marked disadvantage. What does it mean to perceive a distribution rather than a point estimate of depth?

One possible solution is to assume that conscious access to the sensory representation is itself a kind of act, one that we can model as a Bayesian resolution, but with a fixed choice of gain function specific to that sort of act (Figure 6.12(b)). Given another class of actions, such as stepping toward the cliff, we, in effect, use a different gain function.[18] If we are told to point to the edge of the cliff, or to say how far the edge of the cliff is from our feet, or to step to the edge of the cliff, the associated gain functions may lead to actions that can be interpreted as inconsistent estimates of the location of the edge.

Milner and Goodale (1996) describe the apparent discrepancies in inferred location and shape of objects based on different tasks and classes of actions. I wish to point out here that this is to be expected under BDT given Figure 6.12(c) and the discussion of bias introduced by the choice of gain function in the previous section (Figure 6.11). Recall that we examined how the structure of the gain function affected where a Bayesian Observer might choose to step on a difficult-to-see path. When the gain function is asymmetric, the Bayesian Observer's aim point is biased away from the greater source of danger (the sheer drop to the left) and toward the wall. If we were to change the class of action from step-point selection to throwing a projectile down the path, then the gain function in Figure 6.11(b) is plausibly no longer asymmetric. It makes little difference (I'll assume) whether the projectile vanishes off the path into the chasm to the left, or hits the wall to the right and stops. The Bayesian Observer for stepping and for projectile throwing would "aim" for different places on the path, given exactly the same visual information. It is no great leap to the idea that different biases are associated with different classes of actions.

Again, if it were possible to measure experimentally the gains function and prior for different tasks in the same scene, we could determine whether a verbal estimate, or pointing has the same gain function as an estimate based on a different task.

PATTERNS ACROSS TIME

We may imagine Chance and Design to be, as it were, in Competition with each other, for the production of some sort of Events....
Abraham de Moivre (1718/1967), *The Doctrine of Chances*

Updating Priors, Updating Gain Functions

The discussion in the previous section touched on the obvious idea that both objective priors on the states of the world may change, as may objective gain functions.[19] The idea of a *probability engine* was introduced to make it clear that even very simple Bayesian Observers, specialized to depth or pattern vision, say, must be able to assign probabilities to states of the World never before encountered and that perhaps will never be encountered. In this section I consider a closely related problem. It is plausible (and will prove to be true) that an Observer who correctly updates his own subjective prior distribution and gains function can outperform an Observer with a fixed prior and a fixed gain function, across many successive turns. For this to be possible, information from the past must be preserved and used in selecting the new prior.

[*An aside:* BDT provides an obvious candidate for an updating rule, known as *Bayesian updating.* We encountered it before: at the end of each turn, we need only substitute the current posterior distribution (incorporating the sensory information) for the previous prior. The use of Bayesian updating in judgment and decision making is one of the more controversial aspects of the theory in human judgment and decision making. Even confirmed Bayesians such as Jeffrey argue that Bayesian updating is inappropriate (Jeffrey, 1983). These criticisms and potential problems are not directly relevant to the point made here, which is simply the following: Bayesian updating is a method for updating when the prior is stationary (not changing across time). It is tempting to consider using it on slowly-changing or even rapidly-changing priors in

ope that it will "track" the prior across time. The following example illustrates Bayesian updating is not very suitable for tracking changing priors.]

Night and Day

example concerns Bayesian updating when the objective prior does in fact ge. There are two states of the World, *Day* and *Night. Day* endures for 40 turns ved by *Night* for 40 turns. The sensory states of the Observer are *Light* and *Dark.* nitial prior on the state of the World is uniform: the probability of *Day* and of are both 0.5. The likelihood of *Light* during the *Day* is 0.75, of *Dark,* is 0.25. ig the *Night,* the likelihood of *Light* is 0.25, that of *Dark* is 0.75. We will follow bserver's prior through a full *Day* and *Night.* Of course, the prior distribution is ribution and we need only follow *P[Day],* as *P[Night]* is 1 − *P[Day].*

llowing "dawn", the Bayesian updating procedure rapidly moves the prior proby that it is *Day* toward 1.[20] After 10 turns, the prior is 0.9959, that it is *Day.* roblem arises at "dusk", where the state of the World turns deterministically to : the Observer's prior only slowly migrates away from 1. Indeed, three-quarters way through the night, the prior probability that it is *Day* is still 0.9959 despite idence to the contrary in the most recent 30 trials. Even at the end of the night, robability that it is *Night* is only 0.5: the 40 turns of *Day* and of *Night* have led out, returning the prior to indifference just as "dawn" breaks.

e defect exhibited here is not superficial. The essential problem is that Bayesian ing gives exactly the same weight to evidence from recent turns and from turns ccurred much earlier. It has no mechanism for discarding a prior that is in serious reement with (recent) sensory data. As a consequence, it will tend to be very sitive to changes in the prior distribution of states of the World, both sudden and al. Note that this example is in no way a criticism of Bayesian updating, but a criticism of attempts to use it to track a prior that changes over time.

e Bayesian Observer in *Night-and-Day* needs an alarm clock that tells it when card the current posterior and adapt a better estimate of the prior. Or perhaps it a small program that detects temporal edges in the *Light/Dark* sensory states, esets its prior, something an instantaneous Bayesian Observer cannot do.

Augmented Bayes Observers

ose that we address the updating problem directly. Let's first of all concentrate vironments where the prior distribution on the states of the World, $\pi(\theta)$, changes ninistically, according to a specific algorithm. We can imagine that, on the τth a *Prior Demon* selects a new prior, $\pi_\tau(\theta)$, for the World. In the *Night-and Day* ple, the Prior Demon need do little more than count to 40, switch from the *Night* to the *Day* prior, or vice versa, reset the counter to 0 and start over again.

e true priors on (*Day*, *Night*) on any turn are, of course, degenerate: π(*Day*) nd π(*Night*) is 0 during the first 40 of each group of 80 successive turns; *y*) is 0 and π(*Night*) is 1 during the second 40. The choice of the uniform prior ½) for $\tilde{\pi}(\theta)$ is a compromise imposed upon the instantaneous Bayesian Observer, tially due to its ignorance of the deterministic pattern in the successive choices of state of the World, its inability to make use of this pattern. If the state of the World were in fact drawn at random on every trial with the uniform prior, the instantaneous Bayesian Observer is, of course, optimal. An instantaneous Bayesian Observer, then, is poorly equipped to act in a World where the prior distribution on states of the World changes algorithmically. The lack of a mechanism for updating $\tilde{\pi}(\theta)$ to "follow" a deterministically-changing $\pi(\theta)$ is its essential weakness.

Several initiatives in Bayesian vision can be viewed as attempts to augment the instantaneous Bayesian Observer to permit it to deduce from sensory input a plausible, instantaneous choice of $\tilde{\pi}(\theta)$. For example, the competitive priors of Yuille and Bülthoff (1996) provide a simple method for "prior switching" in response to sensory data. Kersten and Schrater (this volume) describe various alternative approaches to tracking changing priors across time.

The challenge proposed here can be treated as a search for an algorithm that, given sensory data across multiple turns, $X_1, X_2, \ldots, X_t$, can provide estimates of a prior that evolves in time taking on the value $\pi_t(\theta)$ at time t. Formally, we seek operators $T_t(X_1, X_2, \ldots, X_t)$ that estimate $\pi_t(\theta)$ given the sensory data so far available. This estimate at each point of time serves as the prior of the instantaneous Bayesian Observer. The resulting *Augmented Bayes Observer* has the potential to outperform an instantaneous Bayes Observer with a fixed prior when the environmental prior changes deterministically[21] and it is possible to estimate the current prior given only sensory data.

Cat-and-Mouse

Suppose that, as you are reading this section in your comfortable office with the doors and windows closed. You suddenly see a mouse scurry across the floor. Your instantaneous prior on mice in your surroundings $T_1(X_1)$ given this sensory event, X_1, is likely going to change. If you aren't the owner of a pet mouse, it has likely increased from the small value T_0 which was your prior before you saw the mouse.

The mouse has vanished but it is reasonable to assume it is still in your (sealed) office. How does your prior estimate $T_t(X_1, X_2, \ldots, X_t)$ evolve across time, in the absence of further sightings of the mouse? What will it be a day later, after you've left the office and returned (perhaps the janitor let the mouse escape)? How does the temporally evolving prior affect your perception of any sudden motion in the periphery? When does it return to its initial value (if ever)?

Your intuition concerning the time-evolution of the prior in this thought example likely reflect considerable knowledge about deterministic aspects of the environment. Mice, like most objects, do not vanish, and do not leave rooms unless there is an exit of some sort. They may be able to gnaw an exit, but that would take some time. If alive, they will likely be seen from time to time. Of course, they may die in your bookcase, changing the modality of the problem from visual to olfactory. The issue then, is whether the visual system makes use of earlier sensory data from a few minutes ago or even days ago in selecting its current prior.

Designing intelligent Augmented Bayes Observers, then, is the challenge proposed here. As noted at the beginning of this section, human observers are typically evaluated in experiments that perversely mimic instantaneous Bayesian environments. To understand how human observers update prior and gain function across time (if, in fact, they do), new kinds of experiments are needed.

BAYESIAN COMPUTATION AND COMPUTATIONAL COMPLEXITY

If you, dear reader, are weary with this tiresome method of computation, have pity on me, who had to go through it at least seventy times, with an immense expenditure of time.
Johannes Kepler (1609) *Astronomica Nova*

Bayesian Computation

Let's consider the computational demands implicit in Bayesian Decision Theory in the evaluation of Equations (6.11) or (6.15). If we could dissect a Bayesian Observer, what characteristic computational resources would we find? Alternatively, if we were to dissect an arbitrary Observer, what about its computational resources would convince us it was, in fact, a Bayesian Observer? SDT and BDT do not prescribe any particular kind of processing despite the formulas included. We have already seen that the "difficult" computation implicit in Equation (6.15) is reduced to a "simpler" computation in Equation (6.19) by an application of Bayes' Theorem. A particular Bayesian Observer is characterized by its decision rule, d, a mapping from sensory states to actions. BDT imposes an ordering on all decision rules but does not require that any particular decision rule be computed in any particular fashion. In this section I will explore possible "Bayesian architectures" and develop computational algebras that allow use to replace operations on full distributions by induced operations on a finite number of parameters. Anyone who has translated a multiplication (hard) into an addition (easy) by means of logarithms has done something similar.

This section is somewhat more mathematical than the remainder of the chapter, and the reader uninterested in the details can skip to the *Discussion* where the main points of this section are summarized.

Multiplication–Normalization

Only two computational operations are really needed in evaluating Equation (6.15). The first is the multiplication of two distributions[22] $f(\theta)$ and $g(\theta)$. The second computational operation needed is the maximization of Equation (6.19) by choice of action for any particular gain function and posterior distribution. In this section, I'll confine attention to the first operation. At any point prior to the choice of an action, this is the single operation presupposed by BDT.

The product of the two distributions is typically scaled so that it is also a distribution. The computation of the posterior distribution from the prior and the likelihood function is an example of this operation. When the likelihood function is the product of likelihood functions for independent sources of sensory information, it can also be computed by the same multiplication–normalization. If we let $\otimes$ denote this

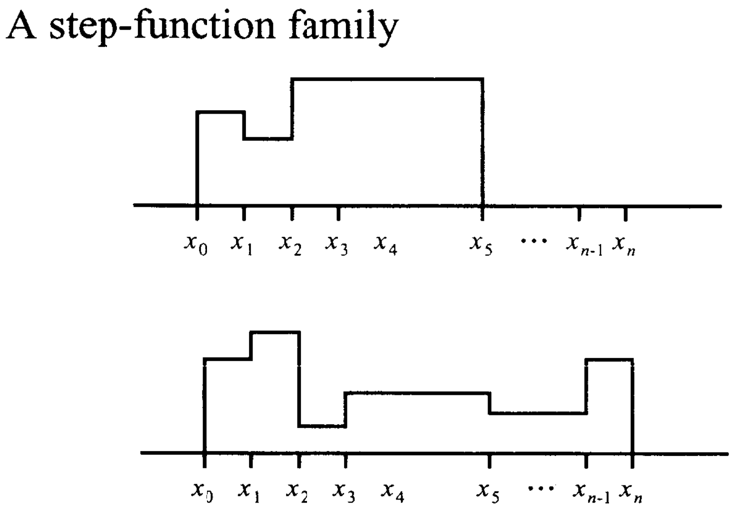
A step-function family
x_0 x_1 x_2 x_3 x_4 x_5 ... x_{n-1} x_n
x_0 x_1 x_2 x_3 x_4 x_5 ... x_{n-1} x_n

Figure 6.13 *A family of probability density functions that are step-functions.* The step regions, defined by the $n+1$ points $x_0, x_1, \ldots, x_n$, are fixed, part of the definition of the family. The members of the family differ in the non-negative values $(v_1, v_2, \ldots, v_n)$, each has on the n intervals delimited by $x_0, x_1, \ldots, x_n$. Each step function is 0 outside these intervals and the area under each step function must be 1. Multiplication of two members of the same step-function family $v = (v_1, v_2, \ldots, v_n)$ and $v' = (v'_1, v'_2, \ldots, v'_n)$ is equivalent to component-wise multiplication of the entries of the two vectors followed by a scaling of all the entries.

multiplication–normalization operation, then it is defined as,

$$(f \otimes g)(\theta) = \frac{f(\theta)g(\theta)}{\int f(\theta)g(\theta)\mathrm{d}\theta} \tag{6.21}$$

when the denominator is non-zero. The denominator can only be zero if the product of the two distributions is the zero-function. It will prove to be convenient to introduce the zero-function $0(\theta) = 0$, as an "honorary" distribution. With this convention, I define $f \otimes g$ to be $0(\theta)$ when the denominator in Equation (6.21) is 0.

Suppose that the choice of possible distributions is restricted to a parametric family Ξ indexed by a finite number of parameters $\xi = (\xi^1, \ldots, \xi^\nu)$. The reader is very likely familiar with a number of parametric families with a finite number of parameters: the Gaussian, the Exponential, etc. A less familiar example of such a parametric family is constructed as follows. First, we assume that θ is a real number, not a vector of real numbers, and we select ν intervals on the real line. The values $\xi = (\xi^1, \ldots, \xi^\nu$ are interpreted as the values of a step-function (Figure 6.13), which has constant value ξ^i on the ith interval and is otherwise 0. In order to be a distribution each of the values ξ^i must be non-negative, and the area under the step function must be 1; we assume that these conditions are met. This finite-parameter "Step Function" family is one we might employ in approximating Equation 6.15 numerically.

Closed Parametric Families

Let us confine attention to parametric families Ξ that are *closed* under the multiplication–normalization operation: whenever f and g are in Ξ, then $f \otimes g$ is also in Ξ. If a likelihood function and a prior distribution are both members of a closed family, then so is the resulting posterior distribution. Put another way, if likelihood and prior share a common finite-parameter representation in a closed family, then the posterior can be expressed in terms of the same parameters. The same can be said of a likelihood function produced as a product of likelihood functions: if a series of likelihood functions are all members of a closed family, then so is their product which then has the same parametric representation as its factors.

What are examples of closed parametric families? The step-function family of Figure 6.13 is almost 1. The product of any two step functions with the same interval boundaries is also a step function with those interval boundaries. The problem is that the resulting step function may be uniformly 0: $\xi = (0, 0, \ldots, 0)$; we need only add the zero function to the family to solve the problem.

The Gaussian Family

A second example of a closed parametric family is the Gaussian,

$$f(\theta; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-[(\theta-\mu)^2/2\sigma^2]} \tag{6.22}$$

with parameters $\xi = (\mu, \sigma)$. Of course, any one-to-one transformation of the parameters can equally well serve as a parameterization. If we use the parameterization, $\xi = (\mu, r)$, where $r = 1/\sigma^2$, then Equation (6.22) becomes

$$f(\theta; \mu, r) = \sqrt{\frac{r}{2\pi}} \mathrm{e}^{-[r(\theta-\mu)^2/2]} \tag{6.22'}$$

With this new parameterization, we can compute the outcome of a multiplication–normalization very easily. If the two Gaussian distributions have parameterizations $\xi_1 = (\mu_1, r_1)$ and $\xi_2 = (\mu_2, r_2)$, then the result of multiplication–normalization is a Gaussian distribution with parameters $(\bar{r}_1\mu_1 + \bar{r}_2\mu_2, r_1 + r_2)$, where $\bar{r}_i = r_i/(r_1 + r_2)$, $i = 1, 2$.

The multiplication–normalization operation of Equation (6.21), restricted to a closed parametric family, induces an operation on the parameters themselves. We can unambiguously write, for the Gaussian case,

$$(\bar{r}_1\mu_1 + \bar{r}_2\mu_2, r_1 + r_2) = (\mu_1, r_1) \otimes (\mu_2, r_2) \tag{6.23}$$

knowing that this operation on the parameters mirrors the operation on the distributions defined by Equation (6.21). To give a formal definition, if one distribution has parameters ξ_a, and a second has parameters ξ_b, then the parameters of the distribution resulting from the multiplication–normalization of the two distributions are, by definition, $\xi_a \otimes \xi_b$. Of course, we are now using the symbol $\otimes$ in two distinct ways, as an operator on distributions and as an operator on their parameters, but this should lead to no confusion.

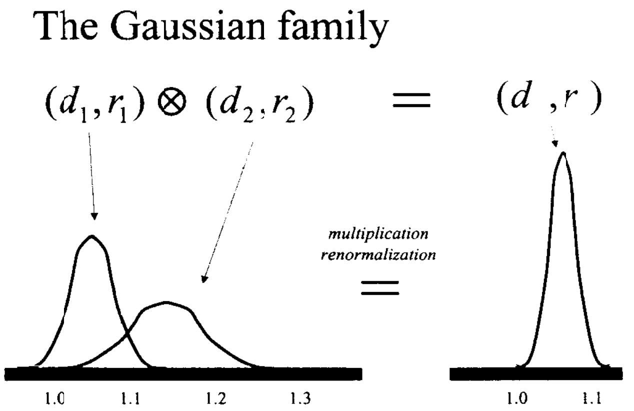

Figure 6.14 *Operations on parameters induced by multiplication–normalization of probability density functions.* Multiplication–normalization of members of the Gaussian family induce operations on the parameters of the family.

Figure 6.14 may help to clarify this induced operation. We represent a distribution by its parameters and represent operations on distributions by induced operations on the parameters, and vice versa.[23]

Looking back at the simple Gaussian example, we see that a simple weighted average and an addition (Equation (6.23)) is equivalent to a multiplication-renormalization operation on Gaussian distributions: visual processing confined to humble weighted averages and sums, is, in fact, equivalent to Bayesian resolution on certain corresponding distributions.

The Uniform Family

A third, and final, example of a parametric family closed under multiplication–normalization is the family of all uniform distributions (on open intervals). Figure 6.15 illustrates a few members. The product of any two of them is either the zero distribution or, after normalization, another member of the family. Accordingly, we once again include the zero function as a member of the family. If we parameterize each such distribution by its endpoints, (a, b), then the multiplication–normalization of (a_1, b_1) and (a_2, b_2) induces the following operation on the parameters,

$$(\max\{a_1, a_2\}, \ \min\{b_1, b_2\}) = (a_2, b_2) \otimes (a_2, b_2) \qquad (6.24)$$

with the convention that (a, b) with $a > b$ denotes the 0-function.

Combining Families

We can take any two parametric families Ξ and Ξ' and construct a new one, denoted[24] $\Xi \otimes \Xi'$ as follows: the distributions in the new family are the normalized products of pairs of distributions, one from the first family and one from the second:

$$\Xi'' = \Xi \otimes \Xi' = \{f \otimes g \mid f \in \Xi \text{ and } g \in \Xi'\} \qquad (6.25)$$

The parameter list for each $f \otimes g$ parameter list is the concatenation of the lists for f and the list for g. There is a natural choice of a parameter list for the product of two families: it is the concatenation of the parameters lists of the two families. For the Gaussian–uniform family the new parameter list is a 4-tuple,

$$(\mu, r, a, b) = (\mu, r) \oplus (a, b), \qquad (6.26)$$

where the symbol $\oplus$ denotes concatenation. The operation induced by multiplication–normalization on the product family is definable in terms of the induced operations on the two original families,

$$(\mu_1, r_1, a_1, b_1) \otimes (\mu_2, r_2, a_2, b_2) = [(\mu_1, r_1) \otimes (\mu_2, r_2)] \oplus [(a_1, b_1) \otimes (a_2, b_2)]\,. \qquad (6.27)$$

There are infinitely many closed parametric families, and, as we have just seen, we can construct new ones from old. Further, we can re-parameterize the parameters of a family by any one-to-one transformation as we did for the Gaussian, replacing (μ, σ) by (μ, r). There is a well-defined operation induced on the new parameters as well.

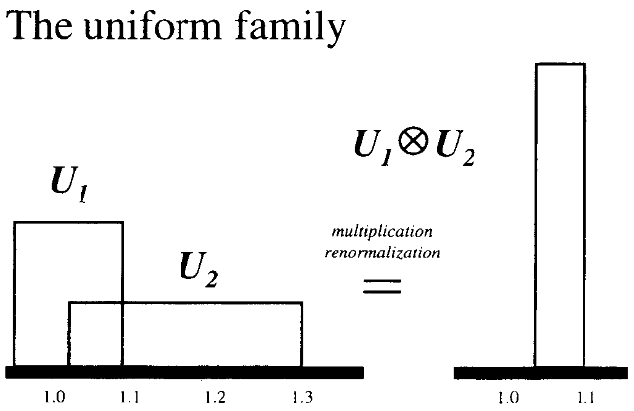

Figure 6.15 *Some members of the Uniform family.* Multiplication–normalization of members of the Uniform family induces operations on the parameters of the family.

Equivalent Data Principle

Once we choose some specific finite-parameter family, the corresponding parameter lists $\xi = (\xi^1, \ldots, \xi^\nu)$ represent *evidence.* Sensory data and prior distribution are represented in the same format and, consequently, the prior distribution at any instant is precisely equivalent to a piece of sensory data that never occurred. Suppose, for example, that there is one source of sensory information and that, for convenience, the Gaussian family with the $\xi = (\mu, r)$ is appropriate. Let $\xi_d = (\mu_d, r_d)$ be the sensory data and $\xi_p = (\mu_p, r_p)$ the instantaneous prior distribution. The first quantity μ in each 2-tuple is the estimate of the quantity of interest, the second, an estimate of its reliability as described above. The resulting estimate will have a bias in the direction of μ_p, but the magnitude of the bias depends on the relative reliability of the prior and the sensory data. If r_d is 0 (the data are worthless), the resulting estimate will be the prior estimate μ_p. If $r_d \gg r_p$, then the prior "data" will be (almost) ignored.

Edward's Challenge

A standard criticism of the Bayesian approach (Edwards, 1972) is the following: If μ_p is not very different from μ_d, the effect of including the prior is not very great. If μ_p *is* very different from μ_d, why would you want to contaminate the sensory data with a prior that almost contradicts it? The preceding example discloses a new role for a prior, as a default mechanism, that ceases to enter into visual processing so soon as reasonable sensory data is available, and it would suggest in agreement with Edwards argument, that the effect of the prior is only noticeable when the sensory data is of poor quality or ambiguous in some respect, that prior information is, in effect, only used when little or no sensory information is available.

Implications

The implications for the complexity issues is straightforward: *there is no characteristic form of Bayesian computation:* two observers with very different computational resources and sequences of computations may be equivalent as Bayesian Observers. The only invariant across these finite-parameter Observers is the number of independent parameters. Further the computational demands of Bayesian resolution need not be very great if Bayesian Observers are constrained to closed, finite-parameter distributional families and carry out computations by induced operations on parameters.

CONCLUSION

> *"Yes, I suppose what I am saying does sound very general," said Malta Kano. "But after all, Mr. Okada, when one is speaking of the essence of things, it often happens that one can only speak in generalities.'*
>
> H. Murakami (1998) *The Wind-up Bird Chronicles*

The first section of this chapter included an introduction to the elements and basic results concerning Statistical Decision Theory (SDT) and Bayesian Decision Theory (BDT). In the second section I introduced a family of finite-parameter Bayesian Observers that all share the same descriptions of the states of the World, possible actions, possible sensory states, and likelihood functions. They differ in their assumed gain functions and prior distributions. The Bayesian Observer whose gain function and prior distribution matches the true gain function and prior distribution of the environment is the Ideal Bayesian Observer, but the other, *Non-ideal Bayesian Observers* were more plausible choices as BDT-derived models of biological visual processing.

A simple counting argument was used to motivate the claim that prior distributions and gain functions could not be learned by repeated exposure to all possible states of the World. A BDT-derived model of biological visual processing must have the ability to compute the prior probability of World states never before encountered and the likely consequences of particular actions never before taken.

Consideration of the *updating problem* and *vision across time* led to a similar conclusion: the Bayesian Observer's prior distribution and gain function need to change as the true prior distribution and gain function change. The change can be deterministic or nearly so, and BDT is an awkward language to describe deterministic change. An instantaneous BDT Observer can be woefully inferior to a hybrid Augmented Bayes Observer that incorporates a small amount of additional computational capacity.

On page 77ff we considered the computational demands of BDT. These need not be great, at least if we can confine our modeling to some closed parametric family of distributions and compute by means of induced operations on the parameters of the family. Whether this is possible is simply a statement about the form of the priors that occur in environment, the choice of visual sensors, and the choice of early transformations in the visual system to enhance computability (Barlow, 1972, 1995).

The discussion of Bayesian computation leads from a different starting point to the conclusion that non-optimal Bayesian Observers are the Bayesian Observers of interest in modeling biological vision. Yet, if all distributions employed by a Bayesian Observer are constrained to a specific closed, finite-parameter family of distributions, it is implausible that the true environmental prior would happen to be a member of the family. If the Bayesian Observer cannot represent the true prior, it will be sub-optimal. A similar discussion of the representation of gain functions (not considered) leads to the same conclusion: the Bayesian Observers of interest to biological vision are not the ideal Bayesian Observers.

Last of all, estimation models of visual representation were discussed and, as it turns out, they are somewhat difficult to justify within the Bayesian framework. In a World of changing gains functions, it is not clear why one would reduce the available sensory information to a representation using a fixed gain function before deciding on the task at hand and the particular dangers and opportunities available in the current scene. In short, there is some confusion in current applications of BDT to biological vision as to the point in visual processing where prior and gain are combined to select actions.

How could we decide that the Bayesian approach to modeling biological vision is worthwhile? What sorts of experimental results would suggest that it is in serious trouble? As I noted at the beginning, the Bayesian approach is not a specific falsifiable hypothesis but rather a (mathematical) language that allows us to describe the structure of the environment and the flow of visual processing. It is a powerful language and therein lies a difficulty. After the data are collected it is not very difficult to develop a Bayesian model that accounts for it. Indeed, almost all of the applications of Bayesian tools to vision are post-hoc fitting exercises.[25] If Bayesian models are to be judged useful, they must also permit prediction of experimental outcomes, quantitatively as well as qualitatively.

The prior distributions of a Bayesian observer are readily interpreted as claims about the environment. In the discussion of the sub-optimal Bayesian observer, I argued that it was reasonable to expect the prior embodied in a biological observer to be discrepant from the true objective prior and consequently, an observed discrepancy between the prior on X estimated from experimental data and the true prior on X in the world is not conclusive evidence against the Bayesian approach. However, if we find ourselves estimating the same prior on X in two different experiments, and find that the two estimates are discrepant, then there are serious grounds to question the entire Bayesian enterprise.

Am I a Bayesian? Not yet, though the temptation is there. The concepts underlying Bayesian Decision Theory are both evident and profound. I do think that a careful program of experimentation devoted to evaluating the Bayesian approach will lead to a much deeper understanding of how the visual system represents and combines evidence—whether or not the Bayesian approach survives the program.

To conclude, it is interesting to compare the current status of Bayesian models in cognition and in perception.

Bayesian models of (cognitive) decision making are controversial and inconsistent with experimental results. The controversies concern *belief*: whether our beliefs can be represented as probability distributions on the states of the World, whether our beliefs follow the axioms of probability, or even whether they should. The sharpest attacks on the Bayesian position concentrate on alternative representations of belief (Fisher, 1936; Edwards, 1972; Shafer, 1976). The Bayesian counterattack (Ramsey, 1931b; Savage, 1954; see Berger, 1985) is equally vigorous.

There is considerable evidence suggesting that our beliefs are not consistent with probability theory (Edwards, 1968; Green & Swets, 1966/1974; Kahneman & Tversky, 1972; Tversky & Kahneman, 1971, 1973; Kahneman & Slovic, 1982; Nisbett & Ross, 1982). Consider, for example, the following problem (adapted from Edwards (1968), illustrated in Figure 6.16):

> There are two urns. The "Black Urn" contains two-thirds black balls and one-third white. The "White Urn" contains two-thirds white balls and one-third black. One of the two urns is selected at random by tossing a fair coin. You'd most likely accept that the probability that the "Unknown Urn" is the "Black Urn" is one-half, that it is the "White Urn" is one-half. Your prior distribution on the two Urns is (0.5, 0.5). Now let's take a sample from the "Unknown Urn". We sample (with replacement) from the Unknown Urn 17 times. There are 11 black balls and 6 white. You'll likely consider that to be evidence in favor of the claim that the Unknown Urn is in fact the "Black Urn". But what exactly is the probability now, after seeing the data, that the "Unknown Urn" is the "Black Urn"?

Please look at Figure 6.16 and make an estimate before reading further.

Most people, given the problem above, estimate that the posterior probability that the "Unknown Urn" is the "Black Urn" to be about 0.75 (Edwards, 1968). The posterior probability distribution on the two Urns is then (0.75, 0.25), about 3:1 odds in favor of the "Black Urn". The correct answer is (0.97, 0.03), or odds of 32:1, in favor of the "Black Urn". The discrepancy between the intuitive estimate of human observers and the correct odds is an example of *Conservatism*, a pervasive error in human reasoning with probabilities: Humans estimate odds that are roughly the cube root of the correct odds (3:1 instead of 32:1).

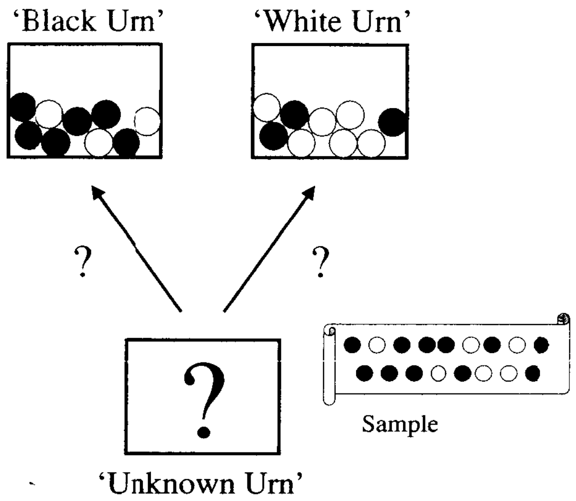

Figure 6.16 *"Conservatism"*. The "Unknown Urn" is either the "Black Urn" or the "White Urn" and the prior odds that it is the one or the other is (½, ½). A sample of size 17 is drawn from the "Unknown Urn" and the results are shown. What is the probability, now that you have seen the sample, that the "Unknown Urn" is the "Black Urn"? The correct answer is given in the text.

Conservatism is observed not only in word problems such as the problem above but also in human performance in the Theory of Signal Detectability: whatever the prior odds of SIGNAL + NOISE and NOISE, human observers respond as if (roughly) the cube root of the prior odds were, in fact, the true prior odds (Green & Swets, 1966/1974). Conservatism, the "cognitive illusions" of Kahneman and Tversky (Kahneman & Slovic, 1982) and other documented failures of human probabilistic reasoning are difficult to explain away as minor deviations from probability theory.

What if advocates of a Bayesian approach to biological vision turn out to be correct? What if BDT-derived models do not break down as completely as their cognitive counterparts, exhibiting analogous failures? What if a consensus develops that such models mirror visual processing in important respects? In sum, what if Bayesian Decision Theory turns out to be the natural "language" for developing accurate models of visual processing? If that all came to pass, we could certainly draw solace from the idea that, although we don't seem able to judge or reason very well, at least something in our skull, our visual system, can.

ACKNOWLEDGMENTS

Preparation of this chapter was supported by grant EY08266 from the National Institute of Health, National Eye Institute, and a research grant from the Human Frontiers Science Program. The author is also grateful to the Institute for Biophysics at the University of Freiburg for support as a guest professor while preparing this chapter. He is particularly grateful to Dr. Lothar Spillmann for his encouragement and help.

Several people were kind enough to comment on earlier drafts or to sit through presentations and discussions of the topics discussed here: Bart Anderson, Donald Hoffman, Dan Kersten, Michael Landy, Paul Schrater, and the complement of the Rochester Symposium on Environmental Statistical Constraints, University of Rochester, Rochester, New York, June, 1998. I am especially grateful to Pascal Mamassian for his willingness to discuss Bayesian ideas while waiting for the stimuli to dry.

NOTES

1. I will use the terms *gain*, *expected gain*, etc., throughout and avoid the terms loss, expected loss (= risk), etc. Any loss can, of course, be described as a negative gain. This translation can produce occasional odd constructions as when we seek to "maximize negative least-squares'. You win some, you negative-win some.
2. The phrase "view geometrically the proportion" describes what we would now call "compute the expected value."
3. A reader of an earlier version of this chapter wondered whether the term "random variable", most often encountered in phrases such as "Gaussian random variable" or "uniform random variable", is applicable to a process where there are only finitely many possible outcomes. It is. The set of possible values of random variables can be finite and can even contain non-numeric values such as "HEADS" or "TAILS".
4. The term *strategy* can also be used. We will encounter *randomized decision rules* in a later section.
5. Not to be confused with *Expected Bayes' Gain*, defined further on. *Expected Gain* depends on the state of the World, *Expected Bayes' Gain* does not.
6. A set of points is convex if the line segment joining any two points in the set is also in the set. An "hourglass" is an example of a non-convex set.
7. The Theory of Signal Detectability (TSD) is better known as Signal Detection Theory, whose abbreviation (SDT) is identical to that of Statistical Decision Theory. To avoid confusion, I will use TSD throughout in referring to the Theory of Signal Detectability / Signal Detection Theory.
8. TSD also takes into account rewards and penalties associated with different kinds of errors. The current discussion illustrates only one way to model TSD within SDT.
9. The reader may be surprised that the "ROC curve" in the figure consists of a series of line segments instead of the usual smooth curve see in text books. The region of achievable gains is always a convex polygon if the set of sensory states and the set of possible actions are both finite, as we are currently assuming they are. The particular shape of the ROC curve is of no importance to the example.
10. If more than one gain point touches the sliding wedge at the same time, then the Maximin rules correspond to the gain point that is furthest up or to the right among the simultaneously touching points
11. For the reader familiar with vector notation: Equations (6.11) and (6.12) are inner products of the prior vector with gains points and Equation (6.12) is just the usual formula for the lines perpendicular to a given vector.
12. I emphasize: in the finite-dimensional case.
13. Bayes' Theorem is (Equation 6.17) below. It can be found in almost any probability or statistics text (e.g. O'Hagan, 1994, Ch. 1).

14. Both sides of Equation (6.17) are equal to the joint probability density function $l(\theta, x)$ of the random state of the world θ and the random sensory state χ. The two sides are just the two possible ways to define conditional probabilities of θ on χ and vice versa. Bayes's signal contribution was to correctly define conditional probability. His "theorem" is an obvious consequence of his definition of conditional probability.
15. I emphasize that a Bayesian Observer is a piece of mathematics intended to describe some component of visual processing. While the language of probability and gain may prove useful in describing this component, there is no assumption that the human observer is consciously aware of these probabilities or gains or that his own beliefs concerning probability or gain influence visual processing.
16. On the peak of Jabal al-Najar in Petra, Jordan.
17. Note that the maximum likelihood estimate of the variance has N, not $N - 1$, in the denominator.
18. I will speak of different gain functions for different actions but, of course, only one gain function is needed, one that we partition according to the different kinds of actions.
19. Of course, all the elements of SDT and BDT may change from moment to moment, but we will be mainly concerned with priors and gain functions here.
20. The posterior distributions (the successive prior distributions) are themselves random variables that depend on the exact sequence of *Light* and *Dark*. For simplicity, in this example, I have compute the priors that would result from the mean number of *Lights* and *Darks* after a given number of turns. That is, after 12 turns of *Day*, I assume that *Light* has occurred exactly 9 times (0.75 × 12). The probabilities reported, then, are not the probabilities to be expected in any single "run" of a simulation of the prior updating process nor need it be the mean of the probabilities across many runs. The essential points, that Bayesian updating responds slowly to change and that it gives the same weight to recent and long-past information are, however, correct.
21. The argument is readily extended to the case where the choice of prior on each turn is partly deterministic and partly stochastic. Many of these considerations could as readily be applied to the gains function as to the prior.
22. For a continuous random variable, the term "probability density function" should be used here. For a discrete random variable, the term "probability mass function" is appropriate. I'll refer to both by the term "distribution" in this section.
23. The relation between the closed family of distributions and the parameters is an example of an isomorphism.
24. We are implicitly assigning a third meaning to the symbol $\otimes$ in defining this product of families.
25. Also known as "death by a thousand parameters."

REFERENCES

Arnauld, A. (1662/1964). *Logic, or the art of thinking ("The Port-Royal Logic")*. Bobbs-Merrill.

Barlow, H.B. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception*, **1**, 371–394.

Barlow, H.B. (1995). The neuron doctrine in perception. In M. Gazzaniga (Ed.) *The cognitive neurosciences* (Ch. 26, pp. 415–435). Cambridge, MA: MIT Press.

Berger, J.O. (1985). *Statistical decision theory and Bayesian analysis*. New York: Springer.

Berger, J.O. & Wolpert, R.L. (1988). *The Likelihood Principle: A review, generalizations, and statistical implications* (2nd edn.). Lecture Notes—Monograph Series, Vol. 6. Hayward, CA: Institute of Mathematical Statistics.

Blackwell, D. & Girshick, M.A (1954). *Theory of games and statistical decisions*. New York: Wiley.

Brainard, D.H. & Freeman, W.T. (1997). Bayesian color constancy. *Journal of the Optical Society of America, A*, **14**, 1393–1411.

Buck, R.C. (1978). *Advanced calculus*, McGraw-Hill.

de Moivre, A. (1718/1967). *The doctrine of chances*. Reprint: New York: Chelsea (reprint).

Edwards, A.W.F. (1972). *Likelihood*. Cambridge University Press.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.) *Formal representation of human judgment*. New York: Wiley.

Egan, J.P. (1975). *Signal Detection Theory and ROC Analysis*. Academic Press.

Feller, W. (1968). *An introduction to probability Theory and its applications* (Vol. I, 3rd edn.). New York: Wiley.

Ferguson, T. (1967). *Mathematical statistics: A decision theoretic approach*. Academic Press.

Fisher, R.A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society*, **26**, 528–535.

Fisher, R.A. (1936). Uncertain inference. *Proceedings of the American Academy of Arts and Sciences*, **71**, 245–258.

Freeman, W.T. & Brainard, D.H. (1995). Bayesian decision theory, the maximum local mass estimate, and color constancy. *Proceedings of the Fifth International Conference on Computer Vision* (pp. 210–217).

Geisler, W. (1989), Sequential ideal-observer analysis of visual discrimination. *Psychological Review*, **96**, 267–314.

Green, D.M. & Swets, J.A. (1966/1974). *Signal Detection Theory and Psychophysics*. New York: Wiley. Reprinted 1974, New York: Krieger.

von Helmholtz, H. (1909). *Handbuch der physiologischen Optik*. Hamburg: Voss.

Huygens, C. (1657). *De Rationicii in Aleae Ludo (On calculating in games of luck)*. Reprinted in Huygens, C. (1920) *Oeuvres Completes*. The Hague: Martinus Nijhoff.

Jeffrey, R. (1983). Bayesianism with a human face. In J. Earma (Ed.) *Testing scientific theories*. University of Minnesota Press.

Kahneman, D. & Slovic, P. (1982). *Judgment under uncertainty*. Cambridge, UK: Cambridge University Press.

Kahneman, D. & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, **3**, 430–454.

Kepler, J. (1609). *Astronomica Nova*. Translated as Kepler, J. (1992) *New Astronomy*. Donahue, W.H. (trans.), Cambridge, UK: Cambridge University Press.

Knill, D.C., Kersten, D. & Yuille, A.L. (1996). Introduction. In D.C. Knill & W. Richards (Eds.) *Perception as Bayesian inference* (pp. 1–221). Cambridge University Press.

Knill, D.C. & Richards, W. (Eds.) (1996). *Perception as Bayesian inference*. Cambridge University Press.

Krantz, D.H., Luce, R.D., Suppes, P. & Tversky, A. (1971). *Foundations of measurement* (Vol. 1): *Additive and Polynomial Representation*. Academic Press.

Maqsood, R. (1996). *Petra: A Traveler's Guide, New Edition*. Garnet.

Mamassian, P. & Landy, M.S. (1996). Cooperation of priors for the perception of shaded line drawings. *Perception*, **25**, Suppl., 21.

Mamassian, P., Landy, M.S. & Maloney, L.T. (in press). A primer of Bayesian modeling for visual psychophysics. In R. Rao, B. Olshausen & M. Lewicki (Eds.) *Statistical theories of the brain*. Cambridge, MA: MIT Press.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: Freeman.

Milner, A.D. & Goodale, M.A. (1996). *The visual brain in action*. Oxford: Oxford University Press.

Murakami, H. (1998), *The wind-up bird chronicles*. J. Rubin (Translator), New York: Vintage.

von Neumann, J. & Morgenstern, O. (1944/1953). *Theory of games and economic behavior* (3rd edn.). Princeton University Press.

Nisbett, R.E. & Ross, L. (1982). *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, NJ: Prentice Hall.

O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics; Vol. 2B; Bayesian Inference*. New York: Halsted Press (Wiley).

Orwell, G. (1983). *1984*. New York: Vintage.

Ramsey, F.P. (1931a). Truth and probability. In *The foundations of mathematics and other logical essays*. London: Routledge & Kegan Paul.

Ramsey, F.P. (1931b). *The foundations of mathematics and other logical essays* London: Routledge & Kegan Paul.

Savage, L.J. (1954). *The foundations of statistics*. New York: Wiley.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton, NJ: Princeton.

Shimojo, S. & Nakayama, K. (1992). Experiencing and perceiving visual surfaces. *Science*, **257**, 1357–1363.

Turing, A.M. (1952). The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society*, **B237**, 37–72.

Tversky, A. & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, **2**, 105–110.

Tversky, A. & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, **4**, 207–232.

Tversky, A. & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman & P. Slovic (Eds.) *Judgment under uncertainty* (pp. 84–98). Cambridge, UK: Cambridge University Press,

Wandell, B.A. (1995). *Foundations of vision*. Sinauer.

Williams, J.D. (1954). *The compleat strategyst*. New York: McGraw-Hill.

Yuille, A.L. & Bülthoff, H.H. (1996) Bayesian decision theory and psychophysics. In D.C. Knill & W. Richards (Eds.) *Perception as Bayesian inference* (pp. 123–162). Cambridge: Cambridge University Press.