

# Bayesian perception: an introduction

Wei Ji Ma, Konrad Kording, Daniel Goldreich

## Introduction

*Probabilistic, or Bayesian, inference* is the process of drawing conclusions based on uncertain evidence. This book explains how perception can be mathematically modeled as a form of Bayesian inference. It is about how the human brain behaves like a smart and sophisticated data scientist (or crime scene investigator, or diagnosing physician, ...) when dealing with noisy and ambiguous sensory data. In recent decades, this approach to perception has been increasingly popular and increasingly widely tested, and this book will provide a gentle introduction to this vast body of literature.

Inference plays a particularly central role in perception: we typically perceive based on partially uncertain information, and we usually only become aware of this when things go wrong. When recognizing sound waves as spoken words or as the chirps of birds, when interpreting a visual image as the face of a friend or as a clock on the wall, when navigating our environment on foot or in a car, we are guided by our inferences. Our inferences are usually correct, but sometimes the lack of full knowledge becomes apparent. The central tenet of this book is that perception is a form of probabilistic inference: from incomplete and imperfect sensory observations, the brain strives to figure out the state of the world.

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

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

Clearly, performing probabilistic inference under such circumstances can never be error-free. However, among all possible strategies that can be used for solving a perceptual inference task, there is always a *best possible* one. This strategy is called optimal probabilistic inference or optimal Bayesian inference. It consists of computing the probability of each possible

interpretation of the observations, and then acting in a manner that has the greatest expected benefit. For instance, if the goal is to make as few errors as possible, the best strategy is for the organism to perceive the interpretation that has the greatest probability of being correct. Optimality does not mean making no mistakes. It means making the best out of the information available to the observer. (Suboptimal Bayesian inference also exists, as we will discuss.)

This book shows step-by-step how to create Bayesian models of perception, and how to compare them with human behavior. It turns out that on many tasks human behavior conforms closely to the predictions of Bayesian models. Carried by this success, the Bayesian perceptual framework has grown rapidly in popularity in recent decades, but a lot of exciting unexplored terrain remains. In the final chapter of the book, we consider open questions and future directions in Bayesian modeling work.

Our approach to modeling perception is “computational” in the sense of David Marr, who described three levels of building a theory about the brain. The top level is computational. At this level, the researcher asks why the brain performs a certain function: what is its goal, and what are the computational principles that can describe and explain this function? The middle level is algorithmic: what are the specific representations or codes used? The bottom level is implementational: how is the function concretely executed by physical processes in the neural hardware? The computational approach highlighted in this book can be considered a top-down approach to neuroscience. We start at the top level by asking “what computations *should* the brain carry out to perceive optimally?” – this is called a *normative* question. By having a successful mathematical model of optimal performance in a particular perceptual task, we then hope to constrain our understanding of the underlying neural processes. This contrasts with a bottom-up approach, in which one might start by modeling small circuits of neurons in biophysical detail, and then attempt to build up the models by combining multiple small circuits. These two approaches are complementary, and both contribute to the understanding of brain function.

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

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

elementary introduction. The book will be suitable to students in a wide range of disciplines, including neuroscience, psychology, cognitive science, computer science, physics, statistics, mathematics, and engineering.

No previous knowledge of Bayesian inference is expected or required of the reader. Because it is about modeling, however, this book necessarily involves mathematics. The elegant and powerful language of mathematics avoids ambiguity and enables quantitative, testable, predictions. The reader should not be intimidated by the math in this book. We firmly believe that every equation in a good model has an intuitive explanation, and we have tried our best to provide those explanations throughout. Readers with a basic understanding of calculus will find the book accessible; those who are uncomfortable with calculus will still be able to understand the majority of the content.

We recommend readers to take sufficient time to work through the within-chapter exercises, end-of-chapter problems, and computer-based lab problems. For building understanding, there is no substitute for struggling with concepts, equations and computer implementations. Some of the exercises and problems require more advanced mathematics or longer calculations; we have marked these with an asterisk (\*). At the end of each chapter, in addition to analytic problems, the reader will see a section of lab problems based around computer simulations. These computer-based problems will be accessible even to readers who are less comfortable with the analytic math in the book.

In Bayesian models of brain function, an observer tries to infer the state of the world from sensory observations. This contrasts with Bayesian statistical data analysis, in which an experimenter tries to infer the value of a model parameter from collected data. The mathematical formalism is the same, but in this book we focus on how the brain perceives, not on how statistical data can be analyzed. That being said, our Bayesian perceptual models also have parameters whose values need to be inferred, and we advocate using Bayesian statistical methods for that purpose. For this reason, we include an Appendix on Bayesian model fitting and model comparison.

We would like to thank Zach Mainen for hosting the graduate course at the Gulbenkian Institute in Portugal where the authors taught and first started thinking about writing a book together. We thank a large number of our colleagues for providing deep and useful feedback on the content and exposition.

We hope you enjoy the book, and we welcome your feedback. Extra material, including solutions to problems, interactive demonstrations, and further reading, can be found on the accompanying website ([BayesianPerception.com](http://BayesianPerception.com)).

## **Dedication**

We want to devote this book to the memory of David Knill (1961-2014). All three of us have learned a good part of what we know about Bayesian modeling of perception from him. He helped us understand its implications, and he helped us see the links between the different models. Lastly, he made studying this topic a lot more enjoyable for all of us. The field of

Bayesian modeling of perception would not be where it is without him and this book would probably never have been written.

## Overview of chapters

*The following chapter organization and numbering is in flux and may be different later.*

- Chapter 1: Probability is everywhere**
- Chapter 2: Building Bayesian models**
- Chapter 3: Understanding Bayesian models**
- Chapter 4: Cue combination**
- Chapter 5: Binary decisions**
- Chapter 6: Structure perception and causal inference**
- Chapter 7: Time and learning**
- Chapter 8: Cost and reward**
- Chapter 9: Basics of neural coding**
- Chapter 10: From neural activity to probability**
- Chapter 11: Neural computation with probabilities**
- Chapter 12: Outlook**

Chapters can be mixed and matched to form courses of various lengths and themes. A course focused on modeling psychophysics experiments should aim to cover Chapters 1 through 6 and 8, with 7, 9 and 10 as potential extras. Chapters 1 is a general introduction to the centrality of uncertainty and inference to perception and action. Chapter 2 discusses a very simple, but complete and common example of a Bayesian model of perception: inferring a real-valued variable by combining a single cue with a Gaussian prior. Chapter 3 interprets the concepts that occur in this model and points out common mistakes. Chapter 4 moves to the case of composite likelihoods, which are encountered in cue combination tasks.

In later chapters, we will examine Bayesian models of tasks that in various ways go beyond the simple case of this chapter. In Chapter 4, we discuss the problem of combining multiple observations of the same stimulus. In Chapter 5, we discuss inference of binary variables, which will allows us to explain the connection between Bayesian modeling and signal detection theory. In Chapter 6, we examine tasks with more complex statistical structures, in particular ones in which other world state variables besides the one of interest play a role. Furthermore, we note that in the present example the world state is a simple stimulus feature, but this is not always the case. There, we will encounter cases in which the world state of interest is a more abstract, relational quantity than a physical stimulus.

In decision-making in a wide range of contexts, from social to political to economic, the optimal strategy involves keeping track of all possible interpretations and continuously updating their probabilities as data come in. Beliefs held in the absence of supporting evidence or prior experience are suboptimal. In most such situations though, the ultimate decision is not only a matter of reporting the most probable value of a state-of-the-world variable, but of acting in a way that maximizes some form of expected utility. In Chapter 8, we discuss how Bayesian inference is combined with utility (cost and reward).

The description of our Bayesian models in terms of an abstract measurement may be puzzling to readers with a background in neuroscience: after all, neurons communicate using spikes rather than with abstract scalar variables. In Chapters 9 to 11, we will make a direct link between Bayesian models and neurons.

### **Routes through the book**

- A course focused on probabilistic models of behavior only could consist of Chapters 1 through 8.
- A course focused on neural probabilistic models could consist of Chapters 1 through 4 and 11 through 13.

1	Chapter 1: Probability is everywhere.....	1-1
1.1	Perception as Probabilistic Inference .....	1-2
1.2	Conceptualizing inference through an example: the baggage claim.....	1-6
1.2.1	The likelihood function.....	1-7
1.2.2	The prior probability distribution.....	1-8
1.2.3	The posterior probability distribution .....	1-9
1.3	Bayesian inference: a closer look.....	1-12
1.3.1	Derivation of Bayes' rule.....	1-12
1.3.2	Factors affecting the likelihood function .....	1-13
1.3.3	Factors affecting the prior .....	1-15
1.3.4	Historic background: perception as unconscious inference.....	1-16
1.4	Bayesian inference in visual perception.....	1-17
1.4.1	Recognizing a friend .....	1-17
1.4.2	Slippery when wet.....	1-19
1.4.3	Camouflage .....	1-20
1.5	Bayesian inference in auditory perception .....	1-22
1.5.1	Birds on a wire.....	1-23
1.5.2	Mondegreens.....	1-25
1.6	Concluding remarks .....	1-29
1.7	References .....	1-30
1.8	Further reading .....	1-31
1.9	Problems.....	1-32

## 1 Chapter 1: Probability is everywhere

Whenever humans perceive, make a prediction, or deliberate over a decision, we are reasoning with probabilities, even if we do not realize it. Specifically, we are using the information we have at hand to infer something else that interests us. The information we have is usually only partially informative, so our inference is not certain. For instance, the fact that a floor is shiny (the information we have) only *suggests* that it may be slippery (the focus of our interest). Using the available sensory information, the brain must determine the probability of each interpretation

(slippery or not). How can the brain make such judgments in the best possible manner? This book describes the optimal method for performing inference; this optimal method is a form of *Bayesian* or *probabilistic* inference.

Plan of the chapter: We outline the perceptual inference process, emphasizing the uncertainty that is inherent in perception. Using a simple example, we introduce the probabilities involved in perceptual inference – the likelihood, the prior, and the posterior – and how they are related through Bayes’ rule. We then illustrate the ubiquity of perceptual inference in daily life with a series of examples involving visual and auditory perception. Our main goal is to provide an intuitive understanding of the perceptual inference process, which will serve as a foundation for the more rigorous mathematical treatments in the following chapters.

## 1.1 Perception as Probabilistic Inference

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

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

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

The transition from sensation to perception requires *conditional probabilities*. A conditional probability is a probability of one event given another: for example, the probability that you are in a good mood given that it is raining outside. . We denote conditional probabilities as  $p(B | A)$ , read “the probability of  $B$  given  $A$ .”

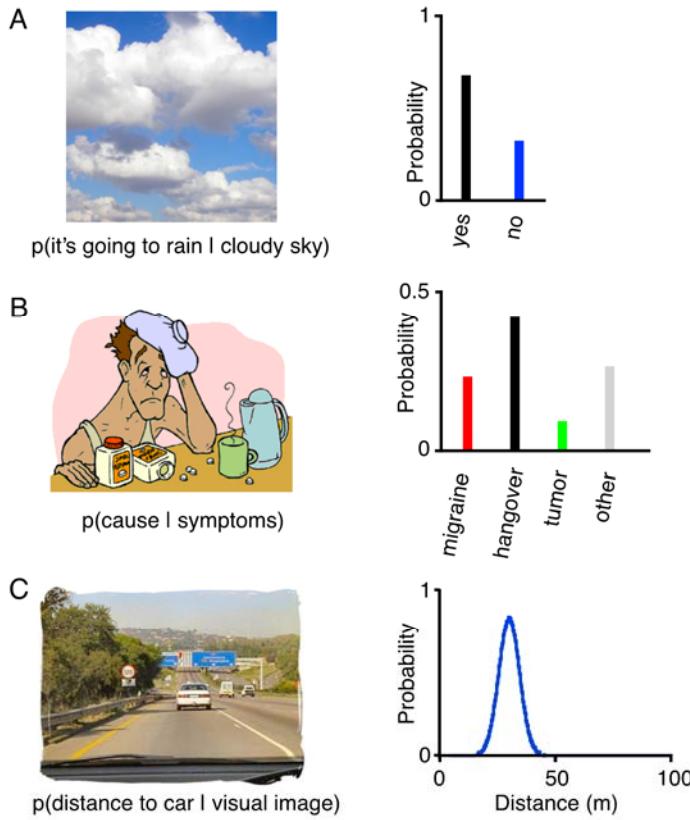
In perception, what is given ( $A$ ) is the sensory input or sensory observations, for example the activation of the photoreceptors in our retina. Given the sensory observations, the observer is interested in a state of the world ( $B$ ). For example, the observer might want to know how probable it is that a floor is slippery, given that the floor is shiny. Since the observer does not know the true state of the world,  $B$  is a hypothesis that the observer is entertaining, and we refer to  $B$  as the hypothesized world state. The conditional probability of interest to the observer is  $p(B|A)$ , the probability of a hypothesized world state given the sensory observations. Whether people are aware of it or not, we make conditional probability judgments very frequently in daily life (Fig. 1).



**Figure 1.** Probability judgments. The notation  $p(B | A)$  is read “the probability of event  $B$  given event  $A$ .”

Depending on the situation, the observer may be concerned with evaluating the conditional probabilities of just two hypothesized world states ( $B_1$ : the floor is slippery, and  $B_2$ : the floor is not slippery), multiple distinct world states, or even a continuum (infinite number) of world states. Ultimately, we would like to express the results of our inference by calculating the

probability of each world state, given the observation (Fig. 2). This would allow us to make an informed decision about the world.

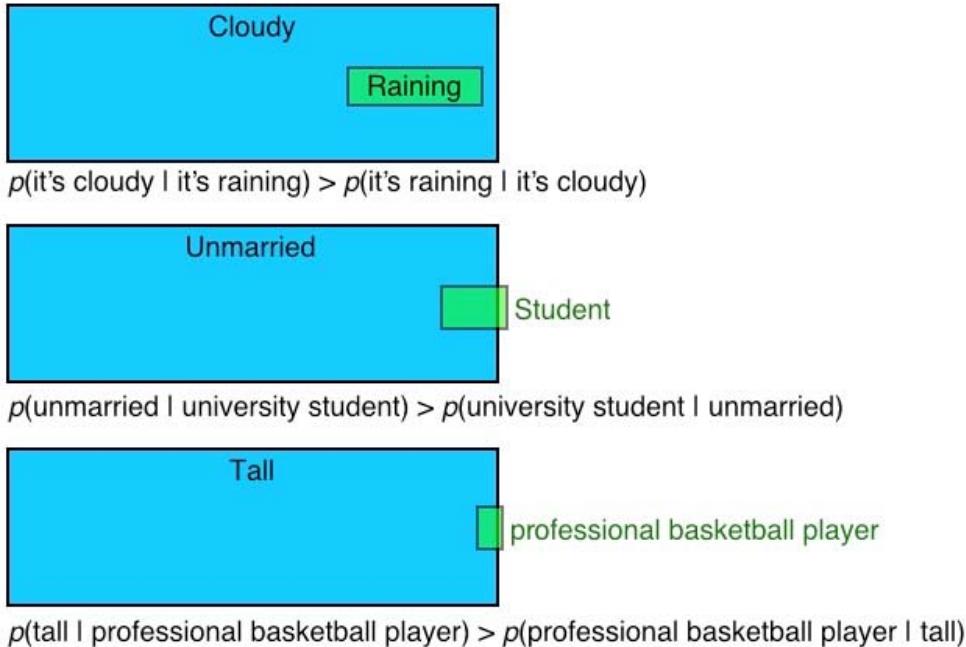


**Figure 2.** Probability distributions. **A.** Two-hypothesis reasoning. **B.** Multiple hypotheses. **C.** Continuous (infinitely many) hypotheses.

Conditional probabilities have a direction. It is important to understand that  $p(\text{world state} \mid \text{observation})$ , the probability the observer most wants to know, is not the same as  $p(\text{observation} \mid \text{world state})$ . Indeed, in general  $p(A \mid B) \neq p(B \mid A)$ . One way to envision probabilities is with areas of rectangles (Fig 3). The area of rectangle A is proportional to the probability of event A, denoted  $p(A)$  and the area of rectangle B is proportional to the probability of event B, denoted  $p(B)$ . The area of overlap between two rectangles is the probability of both events occurring together, denoted  $p(A,B)$ . Lastly, the ratio of overlap area to rectangle area is the conditional probability:  $p(A|B) = p(A,B)/p(B)$  and  $p(B|A) = p(A,B)/p(A)$ .

For instance, suppose  $B = \text{cloudy}$  and  $A = \text{raining}$ . There are more cloudy days than rainy days, so the area of the blue rectangle is greater than the area of the green rectangle (Fig. 3, top panel). The overlap area is the same as the green rectangle area, showing that  $p(B|A)$  is 1: the

probability of clouds, given rain, is 100%. However, the overlap area is much less than the blue rectangle area, showing that  $p(A|B) \ll 1$ : the probability of rain, given clouds, is small. The difference between  $p(A|B)$  and  $p(B|A)$  is apparent in many real world examples (see Fig. 3). It is important not to confuse these two probabilities.



**Figure 3.**  $p(A | B)$  does not in general equal  $p(B | A)$ . Rain is always accompanied by clouds [ $p(\text{cloudy} | \text{raining}) = 1$ ], but clouds are not always accompanied by rain [ $p(\text{raining} | \text{cloudy}) \ll 1$ ]. Almost all university students are unmarried, but most unmarried people are not university students. Nearly 100% of professional basketball players are tall, but the vast majority of tall people are not professional basketball players. The area of each rectangle represents the probability of the event (the areas are not drawn to scale).

### The prosecutor's fallacy

The false belief that  $p(A|B) = p(B|A)$  is called the prosecutor's fallacy or the conditional probability fallacy. In general, it is not true that  $p(A|B) = p(B|A)$ . For instance, most professional basketball players are tall, but most tall people are not professional basketball players. If  $A$  is “being a basketball pro” and  $B$  is “being tall”, then this example illustrates that  $p(B|A) > p(A|B)$ . For each of the following examples, it should be clear that  $p(A|B) \neq p(B|A)$ . In some cases, it should also be clear which of the two probabilities is greater. Consider each example carefully:

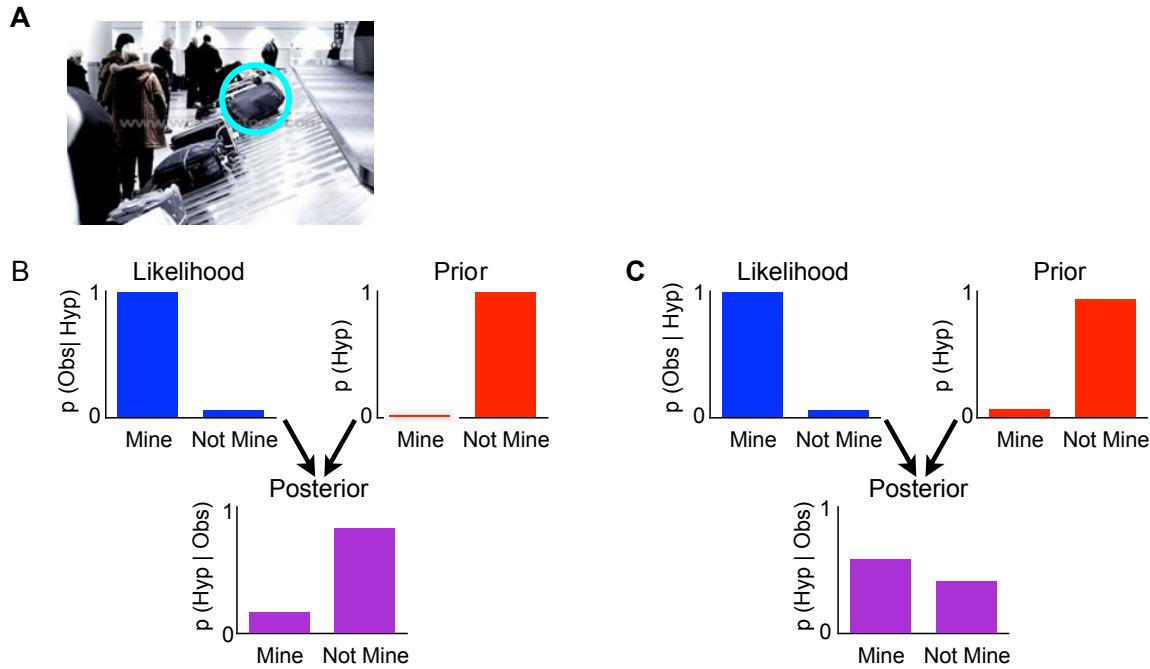
- $p(\text{rain} | \text{clouds}) \neq p(\text{clouds} | \text{rain})$
- $p(\text{speaks French} | \text{born in Paris}) \neq p(\text{born in Paris} | \text{speaks French})$

- $p(\text{patient has the condition} \mid \text{tests positive}) \neq p(\text{tests positive} \mid \text{has condition})$

The prosecutor's fallacy takes its name from the false argument, sometimes put forth in courts of law, that  $p(\text{defendant is innocent} \mid \text{evidence}) = p(\text{evidence} \mid \text{defendant is innocent})$ . For example, suppose that a partial, smudged fingerprint is found on a weapon left at a crime scene. A fingerprint database search reveals that a man who lives in the same city has a fingerprint that matches the one left on the weapon. A forensic expert testifies that only 1 in 1,000 randomly selected people would provide such a match. The prosecutor argues that, based on the forensic expert's testimony, the probability that the defendant is innocent is only 1 in 1,000. The prosecutor is confusing  $p(\text{observation} \mid \text{innocent})$  – the testimony of the forensic expert – with  $p(\text{innocent} \mid \text{observation})$ . In fact, these conditional probabilities are rarely equal. Bayes' rule (section 1.1.4) permits the correct calculation of  $p(A|B)$  from  $p(B|A)$  and other relevant probabilities, and has been used for this purpose in some courts (Fenton, 2011).

## 1.2 Conceptualizing inference through an example: the baggage claim

Many air travelers have waited expectantly in an airport baggage claim area, watching for their bags to drop down the chute into the circulating luggage carousel (Fig. 4A). Let us suppose that you are engaged in this ritual of modern-day air travel along with 99 other passengers from your flight, each of whom, like you, checked one item of luggage. A recording piped through the speakers reminds you that “Many bags look alike. Please check your bag carefully before exiting the terminal.” Indeed, your bag is one of the most popular models on the market, a black rectangular case used by 5% of all travelers. Of course, if you look at your bag close-up, you will notice individual markings — a name tag, a piece of string you have attached to the handle, etc. — that allow you to unambiguously identify your bag. But at the distance you are standing from the luggage chute, you cannot tell your bag from the 5% of bags in general that have the same shape, size and color. Now let's suppose that the first bag from your flight to enter the luggage carousel indeed has the same shape, size, and color as your bag. Is it your bag?



**Figure 4.** Expectation influences perception. **A.** The 1st bag and also the 86th bag match yours in shape, size, and color. **B.** Likelihood function, prior probability distribution, and posterior probability distribution upon viewing the 1st bag. Your posterior distribution indicates that the bag is probably not yours. **C.** Likelihood function, prior distribution, and posterior distribution upon viewing the 86th bag. The same likelihood as in (A) combined with a different prior expectation produces a posterior distribution that favors the hypothesis the bag is yours. In this and all subsequent figures in the book, likelihood functions are drawn in blue, prior distributions in red, and posterior distributions in purple.

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

### 1.2.1 The likelihood function

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

What information is contained in your observation? If the bag you are viewing is in fact your own, it will have the same shape, size, and color. Thus,  $p(\text{observed shape, size, color} \mid \text{my bag})$

bag) = 1. But even if the bag you are viewing is not your own, it has some chance of matching the shape, size, and color of your bag. Since your bag is the model used by 5% of travelers,  $p(\text{observed shape, size, color} \mid \text{not my bag}) = 0.05$ . These two conditional probabilities are known as *likelihoods*. The likelihood of a hypothesis is the probability of the sensory observations if the hypothesis were true, or in other words, how expected the observations are if the hypothesis were true. A plot of the likelihood of every possible world state, known as the *likelihood function*, summarizes the degree to which the observation favors one world state interpretation over the other (Fig. 4B). The less informative the observation, the “broader” or “flatter” will be the likelihood function; the more informative the observation, the “narrower” or “sharper” will be the likelihood function. The observation will generally favor some interpretations more than others, but how much it does so will depend on the quality of the sensory information.

### 1.2.2 The prior probability distribution

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

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

In short, perceptual inference is based not just on the observation (via the likelihood function), but also on expectation. We represent expectation by *prior probability*. The prior probability of a world state is based on all relevant information except the current observation. In the present example, your experience of waiting patiently as 85 bags emerged onto the carousel, together with your background knowledge that 100 bags were present on your flight, has informed you that the prior probability that your bag will emerge next is 1 in 15 (i.e., 6.7%), which is greater than the 1% that it was for the first bag. Although prior probabilities are

conditioned on experience and background knowledge, in the interest of brevity we usually omit the conditioning symbol ( $|$ ) and write prior probabilities simply as  $p(\text{hypothesized world state})$ , e.g.,  $p(\text{the bag is mine})$  and  $p(\text{the bag is not mine})$ . We plot the prior probability of each hypothesized world state as a *prior probability distribution* (Fig. 4B,C).

### 1.2.3 The posterior probability distribution

The brain somehow has to combine likelihoods,  $p(\text{observation} | \text{hypothesized world state})$ , with prior probabilities,  $p(\text{hypothesized world state})$ , to generate the probabilities it most wants to know:  $p(\text{hypothesized world state} | \text{observation})$ . These latter probabilities are called *posterior probabilities* to indicate that, unlike prior probabilities, they are formed *after* the observation. The *posterior probability distribution* (Fig. 4B,C) represents the brain's belief in each possible world state, based on all relevant information (i.e., observation and expectation).

How should the brain combine likelihoods with priors to generate posteriors? Remarkably, it turns out that the optimal method is based simply on the multiplication of the likelihood and the prior:

$$p(\text{hypothesized world state}_i | \text{observation}) = \frac{p(\text{observation} | \text{hypothesized world state}_i)p(\text{hypothesized world state}_i)}{\sum_{k=1}^N p(\text{observation} | \text{hypothesized world state}_k)p(\text{hypothesized world state}_k)}$$

where the sum in the denominator is over all possible world states,  $k = 1, 2, \dots, N$  (in the luggage example,  $N = 2$ ). This simple but powerful relationship is known as Bayes' rule, and provides the basis for all topics in this book. The numerator shows that the posterior probability of a given world state is simply proportional to the product of its prior probability and its likelihood. The constant term in the denominator ensures that the posterior probabilities of the different world states sum to one (indicating that exactly one of the states is the correct one).

Continuing with our perceptual example, let us evaluate the posterior probability that the first bag that you see emerge onto the luggage chute is your own. We first enumerate the possible world states, which in this case we will call hypothesis  $H_1$  (the bag is mine), and  $H_2$  (the bag is not mine). Next, we write down the prior probabilities of each hypothesis, given our knowledge that this is the first bag to appear. We then write the likelihoods that express the probability of the sensory observation, (shape, size, and color of the luggage seen) given each hypothesis

$$\begin{aligned} p(H_1) &= 0.01 & p(\text{observation} | H_1) &= 1 \\ p(H_2) &= 0.99 & p(\text{observation} | H_2) &= 0.05 \end{aligned}$$

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

$$p(H_1 \mid \text{observations}) = \frac{1 \cdot 0.01}{1 \cdot 0.01 + 0.05 \cdot 0.99} = 0.168$$

$$p(H_2 \mid \text{observations}) = \frac{0.05 \cdot 0.99}{1 \cdot 0.01 + 0.05 \cdot 0.99} = 0.832$$

There are several important considerations to appreciate at this point:

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

Now suppose that we continue to wait for our bag to appear, failing to see it among the first 85 bags to enter the carousel. To calculate the posterior probability that the 86<sup>th</sup> bag, which also matches ours in shape, size, and color, is our own, we follow the same procedure, but with new prior probabilities of 1/15 for  $H_1$  and 14/15 for  $H_2$  (Fig. 4C).

Exercise: Verify that the posterior probabilities when evaluating the 86<sup>th</sup> bag will be  $p(H_1 \mid \text{observation}) = 0.588$ , and  $p(H_2 \mid \text{observation}) = 0.412$ .

Thus, our confidence that the bag we are viewing is our own has now increased dramatically, from 16.8% (first bag seen) to 58.8% (86<sup>th</sup> bag seen), despite the fact that in the two cases the observation, and therefore the likelihood functions, are the same (Fig. 4). The posterior distribution depends not only on the sensory data but also on the prior distribution.

## Revisiting The prosecutor's fallacy

Using Bayes' rule, we can now provide a fully worked example to illustrate the prosecutor's fallacy (Box in Section 1.1). Recall that a partial, smudged fingerprint is found on a weapon left at a crime scene. Some of the people who live in the city happen to have their fingerprints on file, and a fingerprint database search reveals that a man who lives in the same city has a fingerprint that matches the one left on the weapon. A forensic expert testifies that only 1 in 1,000 people would provide such a match, and on this basis the prosecutor argues that the defendant's probability of being innocent is only 1 in 1,000.

Let's suppose that the city has 1,000,001 adult inhabitants. Given only that the defendant lives in the city, his prior probabilities of being innocent ( $H_1$ ) or guilty ( $H_2$ ) are therefore:

$$p(H_1) = 1,000,000 / 1,000,001$$

$$p(H_2) = 1/1,000,001$$

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

$$p(\text{observation} | H_1) = 1/1,000$$

$$p(\text{observation} | H_2) = 1$$

Using Bayes' theorem, we find that:

$$p(H_2 | \text{observation}) = \frac{1 \cdot \frac{1}{1,000,001}}{\frac{1}{1,000} \cdot \frac{1,000,000}{1,000,001} + 1 \cdot \frac{1}{1,000,001}} = \frac{1}{1,001}$$

The defendant is almost surely innocent, despite the prosecutor's argument! Note that our inference can be verified as follows: The city contains 1,000,001 people, 1,000,000 who are innocent and 1 who is guilty. Consequently, if we had the fingerprints of everyone in the city, we'd expect 1,001 matches, only 1 of which is from the guilty citizen. The probability of guilt, given a fingerprint match, is therefore 1/1,001.

### 1.3 Bayesian inference: a closer look

Having introduced the elements of Bayesian inference, we now explore more deeply. We first derive Bayes' rule. Next, we discuss factors that influence the likelihood function and prior distribution. Lastly, we provide a brief historic perspective on the origins of inference.

#### 1.3.1 Derivation of Bayes' rule

Where does Bayes' rule come from? There are several derivations that all lead to the same rule. Here we derive Bayes' rule using two basic rules of probability: the product and sum rules. We use the baggage claim scenario in our derivation, but the derivation can easily be extended to scenarios with any number of world states.

Suppose you observe a bag that looks like yours. It is possible that the bag is yours and it looks like yours, or that the bag is not yours but it looks like yours anyway. These two situations can be thought of as pairs of events (world state, observation) that might have occurred:

1. (the bag is mine, it looks like mine) =  $(H_1, \text{observation})$
2. (the bag is not mine, it looks like mine) =  $(H_2, \text{observation})$

We can express the probability of each event pair by applying the *product rule*, which states simply that  $p(A, B) = p(B|A)p(A)$ . Thus,

1.  $p(H_1, \text{observation}) = p(\text{observation}|H_1)p(H_1)$
2.  $p(H_2, \text{observation}) = p(\text{observation}|H_2)p(H_2)$

Because  $p(A, B) = p(B, A) = p(A|B)p(B)$ , we can also write:

1.  $p(H_1, \text{observation}) = p(H_1|\text{observation})p(\text{observation})$
2.  $p(H_2, \text{observation}) = p(H_2|\text{observation})p(\text{observation})$

Dividing expression 1 by expression 2 in each case, we obtain:

$$\frac{p(H_1|\text{observation})}{p(H_2|\text{observation})} = \frac{p(\text{observation}|H_1)p(H_1)}{p(\text{observation}|H_2)p(H_2)}$$

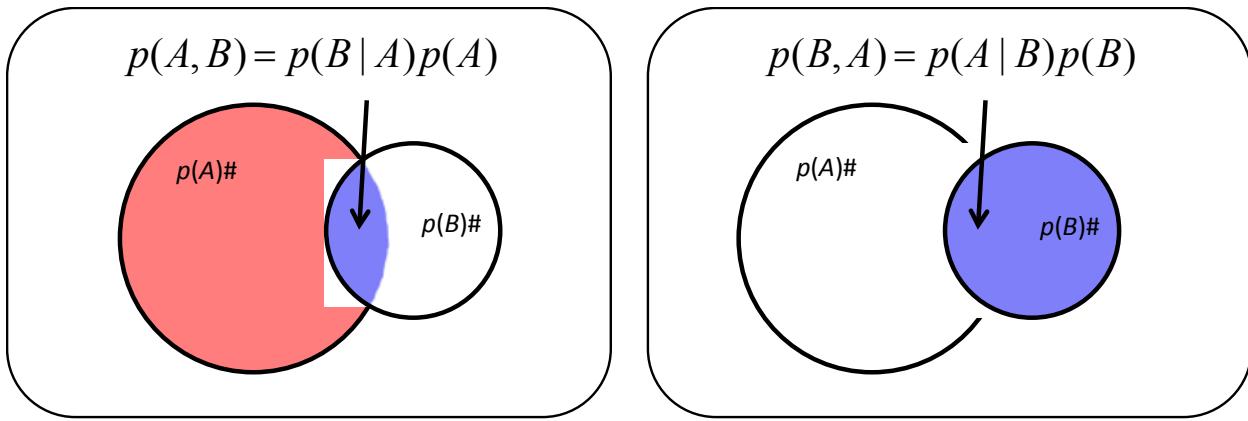
Since the bag must either be yours or not, these two mutually exclusive probabilities must sum to 1. This is a result of the *sum rule* for probabilities. Thus:

$$p(H_1|\text{observation}) + p(H_2|\text{observation}) = 1$$

The solution to these two equations is Bayes' rule:

$$p(H_1|\text{observation}) = \frac{p(\text{observation}|H_1)p(H_1)}{p(\text{observation}|H_1)p(H_1) + p(\text{observation}|H_2)p(H_2)}$$

$$p(H_2|\text{observation}) = \frac{p(\text{observation}|H_2)p(H_2)}{p(\text{observation}|H_1)p(H_1) + p(\text{observation}|H_2)p(H_2)}$$



**Figure 5.** Bayes' rule can be derived by expressing the overlap area (purple) in two equivalent ways.

### An Alternate Form of Bayes' Rule

Bayes' rule can be written in several forms, which are all mathematically equivalent. One common form results from the two expressions for  $p(H_1, \text{observation})$  shown above:  $p(H_1, \text{observation}) = p(\text{observation}|H_1)p(H_1)$ , and  $p(H_1, \text{observation}) = p(H_1|\text{observation})p(\text{observation})$ .

It follows that:  $p(\text{observation}|H_1)p(H_1) = p(H_1|\text{observation})p(\text{observation})$

Dividing by  $p(\text{observation})$  yields a compact form of Bayes' rule:

$$p(H_1|\text{observation}) = \frac{p(\text{observation}|H_1)p(H_1)}{p(\text{observation})}$$

By comparing this form of Bayes' rule to the one we derived earlier, we see that:

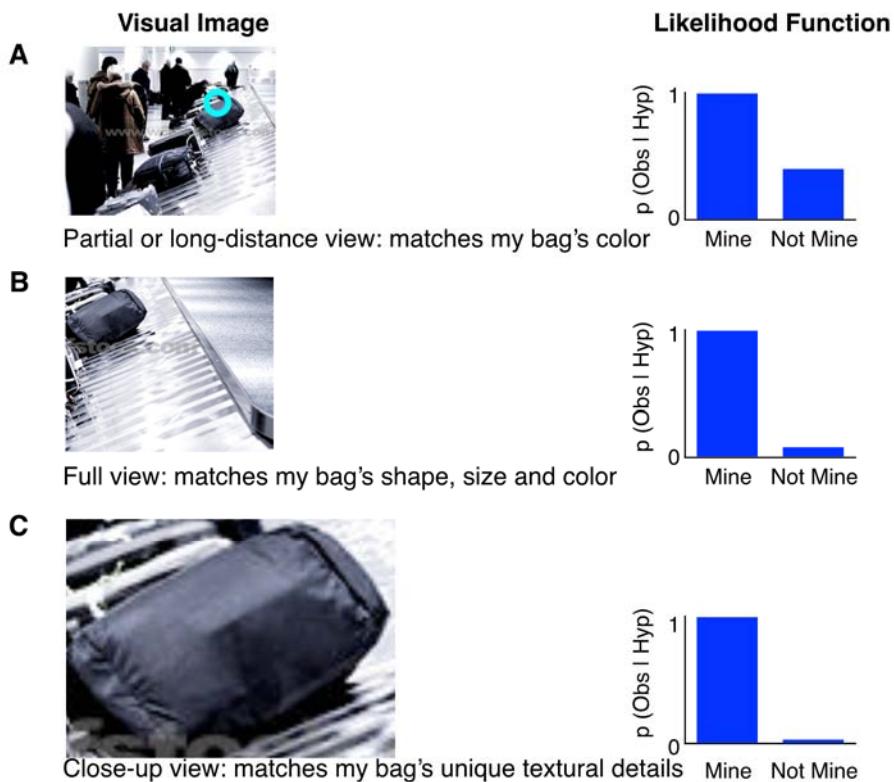
$$p(\text{observation}) = p(\text{observation}|H_1)p(H_1) + p(\text{observation}|H_2)p(H_2)$$

In words, the probability of the observation is the probability that  $H_1$  is true AND that the observation would occur if  $H_1$  were true, OR that  $H_2$  is true AND that the observation would occur if  $H_2$  were true (following the product and sum rules of probability, each “AND” is a multiplication, and the “OR” is an addition). Thus, the probability of the data is a weighted sum of the likelihoods of the hypotheses, where the weights are the prior probabilities of the hypotheses. This type of summation is an example of a procedure called marginalization, which we consider in more detail later in the book (e.g., Ch. 6).

### 1.3.2 Factors affecting the likelihood function

In real life, many factors influence the likelihood function. When you first glimpsed the 86<sup>th</sup> bag, your view of it may have been partially blocked by other baggage circulating on the

carousel, or by people standing in front of you. Your partial view may have revealed only that the bag was black, but nothing about the bag's shape or size. Based on this initial scant visual information, your likelihood function would have been broader, reflecting the fact that, for example, 40% of travel bags are black (Fig. 6A). When the bag later came into clear view, your likelihood function sharpened as you gained access to the bag's shape and size, in addition to its color (Fig. 6B; this is the likelihood function that we used above). As the bag comes closer, you can distinguish textural details such as wrinkles and scratches. This will result in further sharpening of the likelihood function (Fig. 6C). Many environmental conditions, including distance, obstructions, and others can reduce the quality of sensory data and thereby broaden the likelihood function (Fig. 7A).

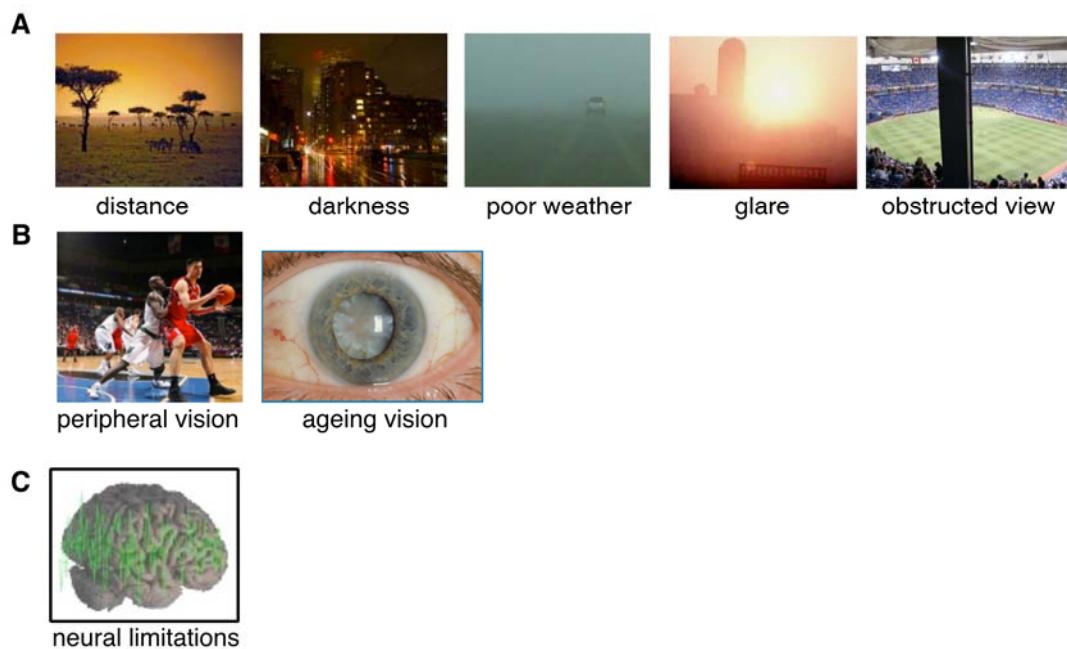


**Figure 6.** A closer look at the likelihood function. A closer view of the bag reveals features that were not apparent from a distance. This improvement in the quality of the observation can dramatically sharpen the likelihood function. **A.** 40% of bags match yours in color. **B.** 5% of bags match yours in shape, size, and color. **C.** Fewer than 1% of bags have the same textural details (wrinkles, bulges, scratches) as your bag.

In addition to environmental conditions, the sensory capabilities of the observer play an important role in shaping the likelihood function. A different viewer, with worse eyesight than yours, would experience more sensory uncertainty. In short, the “eyes of the beholder” affect the quality of sensory inputs and therefore the shape of the likelihood function (Fig. 7B).

More subtly, the observer’s background knowledge also plays an important role in

shaping the likelihood function. To an observer who remembered only the shape and size of his bag, but not its color, the same visual scene would be less informative (perhaps 5% of bags match yours in shape, size, and color, but 12% match it in shape and size); to an observer who had never before noticed the scratch on his bag, the sight of a scratch might have misinformed the likelihood calculation; an observer with little travel experience might have employed in the construction of the likelihood function an inaccurate estimate of the proportion of bags that look like his own; and so on. Thus, the shape of the likelihood function, like the shape of the prior distribution, depends on the background knowledge of the observer.



**Figure 7.** Sources of sensory degradation. These factors reduce the quality of visual inputs, causing likelihood functions to broaden. **A.** Physical features of the environment. **B.** The observer's sensory acuity. **C.** The observer's central nervous system. Most of the factors shown in A and B have analogs in the other senses. For example, in the case of audition, distance, soft speech, ambient noise, and ageing ears all result in low-quality inputs. In every sensory system, neural limitations such as faulty background knowledge and neural noise (see Chapters 11 and 12) also pose a challenge to perception.

### 1.3.3 Factors affecting the prior

Like likelihood functions, prior distributions are based on background knowledge. Prior probabilities therefore evolve over time as the observer acquires new knowledge, and because of this priors also tend to differ from one observer to another. To take an obvious example, an adult will in general have more knowledge about the dangers of crossing a busy street, or being bitten by a snake or a dog, than will a young child, because the adult has more experience with the world.

Priors can change on multiple timescales. They can evolve gradually as the observer gains experience with the world, or much more rapidly during the course of an evolving situation. In the luggage example, the prior probability that the next bag would be yours changed after every observation of a bag entering the carousel. In general, our priors update as we observe a changing situation.

Differences in prior probabilities can arise between one person and another due to differences in a lifetime of experience, or they can arise in a momentary, situation-dependent fashion. Consider a person who arrives late at the baggage claim, after the bags from her flight have already begun to circulate. This person can look at the bags on the carousel, but will not be able to know how many bags have already been retrieved and taken away by other passengers who have left the area. Consequently, she will not have as accurate a prior distribution as will a person who arrived early enough to see each bag enter the carousel. In general, those with greater relevant knowledge have more realistic priors, facilitating accurate perception.

### 1.3.4 Historic background: perception as unconscious inference

Bayes' rule is so-named after the English mathematician Thomas Bayes (1702-1761), who was interested in problems of inverse probability, essentially how to calculate  $p(B|A)$  when we know  $p(A)$  and  $p(A|B)$ . Bayes' *An Essay Towards Solving a Problem in the Doctrine of Chances*, published posthumously in 1763, introduced the foundation for the conditional probability calculus, a field of statistical reasoning now called Bayesian inference.

Bayes' rule was later derived independently by the French mathematician and physicist, Pierre Simon Marquis de Laplace (1749-1827). Laplace applied the formula with great effect to problems in a wide range of disciplines. Importantly, Laplace also recognized the pervasiveness of probability, stating that "the most important questions of life ... are indeed, for the most part, only problems in probability. One may even say, strictly speaking, that almost all our knowledge is only probable" (Laplace, 1995). Indeed, today Bayesian statistical inference is playing a rapidly growing role in an extraordinarily diverse set of disciplines covering nearly all fields of science and engineering: neuroscience, psychology, evolutionary and molecular biology, geology, astronomy, statistical data analysis, economics, robotics, and computer science, to name but a few.

The idea that perception is a form of unconscious inference, however, arose independently of Bayes and Laplace. Several scientists contributed to this notion. The early Arab physicist and polymath, Ibn Alhacen (965-c.1040 CE), recognized presciently that "...not everything that is perceived by sight is perceived through brute sensation; instead, many visible characteristics will be perceived through judgment...in conjunction with the sensation of the form that is seen." Thus, "...familiar visible objects are perceived by sight through defining features and through previous knowledge..." (Alhacen, *De aspectibus*, Book 2, translated by Smith, 2001).

Much later, the German physician and physicist Hermann von Helmholtz (1821-1894) again expressed the idea that perception is a form of unconscious inference, stating eloquently

that “Previous experiences act in conjunction with present sensations to produce a perceptual image” (Physiological Optics, 1867). The ideas of Alhacen and Helmholtz fit beautifully with the view that perception is a form of Bayesian inference (Fig. 8).



Thomas Bayes, 1702 - 1761



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



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



Hermann Ludwig von Helmholtz, 1821-1894  
“**Previous experiences** act in conjunction with **present sensations** to produce a **perceptual image**” - *Physiological Optics*

$$P(H) \times P(Obs|H) \propto P(H|Obs)$$

prior      Likelihood      Posterior

**Figure 8.** Luminaries in the development of Bayesian inference and the view that perception is unconscious inference.

## 1.4 Bayesian inference in visual perception

We will now further illustrate perceptual inference with a variety of examples, drawn from everyday life. Our goal is for the reader to develop an intuitive understanding of likelihoods, priors, and posteriors, and an appreciation for the remarkable explanatory power of Bayesian inference as a model of perception. We will not use mathematics in these examples, but will instead explore each example qualitatively and graphically. We will see that each has unique features, yet each is based upon the joining of likelihood function and prior through Bayes’ rule to generate a posterior perceptual inference. We hope that these examples begin to reveal both the richness of perceptual inference, and the wide applicability of the Bayesian perceptual framework.

### 1.4.1 Recognizing a friend

Suppose you see a person walking in the distance, and wonder whether he is your friend (Fig. 9A). Whatever conclusion you reach, you will have some degree of confidence, and your degree of confidence may change over time as you continue to view the scene. We perceive visual scenes with little conscious effort, but the processing the brain engages in is sophisticated. Even the best computer vision systems fail to match the accuracy of human visual perception.

Why is scene recognition so challenging? The sensory input captured by the nervous

system (the visual image in this case) is compatible with multiple interpretations. The visual image could be that of your friend or of another person. The image is not entirely uninformative - it provides sufficient information to recognize that the object in question is in fact a person, and it provides information regarding the approximate shape (height, girth, etc.) of the person. Over time, the moving image also provides information about the person's gait. Nevertheless, the person is far away and your view of him is partially blocked by other people. Thus, recognition is difficult in part because the sensory information is limited, and also in part because many pieces of distributed information must be combined into a joint estimate.

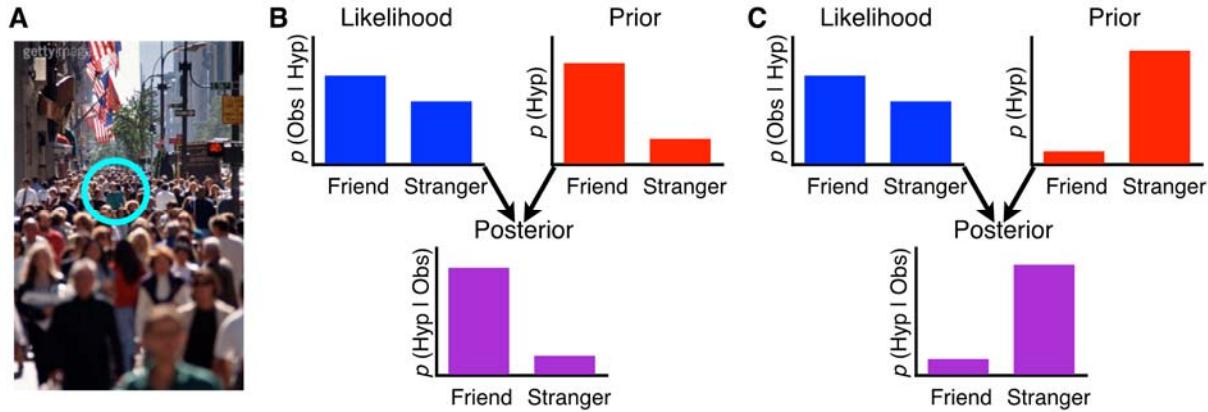
The likelihood function summarizes how well a visual image allows you to distinguish one possible world state from another (friend or not, Fig 9). Many factors affect this likelihood function, including height (your friend is tall) hair color (brown), way of holding the head (tilted). We will leave aside at present how we might arrive at the exact form of the likelihood function. For now, it is sufficient to understand that the likelihood function represents the full information content of the image relevant to the question at hand (is that my friend?). Specifically, it represents the probability that your friend would give rise to the visual image you currently sense, compared to the probability that another person would give rise to the same visual image. The width of the resulting likelihood function defines the difficulty of a perceptual problem.

Importantly, as we have already seen, the likelihood function is not sufficient to solve the problems we want to solve. The likelihood function plots the probability of the observation given each hypothesized world state:  $p(\text{observation} | \text{world state})$ . What we want to know is the posterior distribution: the probability of each possible world state, given the observation:  $p(\text{world state} | \text{observation})$ . To determine the posterior probability, we combine the likelihood function with a prior distribution. Let's consider two different scenarios, each one associated with the same visual image:

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

The sensory input is identical (Fig. 9A), but your perceptual inference would differ dramatically under these two scenarios. Under Scenario 1, you would probably conclude that the person walking towards you is your friend; under Scenario 2, you would probably conclude that he is not. Clearly, as we have already seen, expectation plays a crucial role in the perceptual inference process.

Bayes' rule shows how to optimally combine expectation, represented by the prior distribution, with the observation, represented by the likelihood function, to calculate a posterior distribution (Fig. 9). The posterior distribution represents our perceptual inference, and is based on all knowledge we have (present observation and previous experience).



**Figure 9.** Recognizing a friend. **A.** A crowded visual scene offers a low-resolution view of a person who resembles your friend. **B.** You consider the probability that the visual image would result from your friend to be greater than the probability that it would result from a stranger (likelihood function). You expected to meet your friend (prior distribution). Therefore, you believe the person in question is probably your friend (posterior distribution). **C.** In this alternate scenario, you thought your friend was out of town, so your prior distribution sharply favors the *stranger* hypothesis. Given the same observation (likelihood function), you conclude that the person in question is probably not your friend.

#### 1.4.2 Slippery when wet

As humans move through the world, we rely on our senses, particularly vision, to avoid hazards. In the modern world, hazards come in many forms, for instance an object in our path, a rapidly approaching car, or a downward step such as a curb. Another hazard of modern life is the slippery floor. Is the floor slippery (Fig. 10)? If it is – or might be – caution is warranted: small and slow steps. If it is not, we can safely proceed with long, purposeful strides. An important question is how perception can distinguish the two cases.

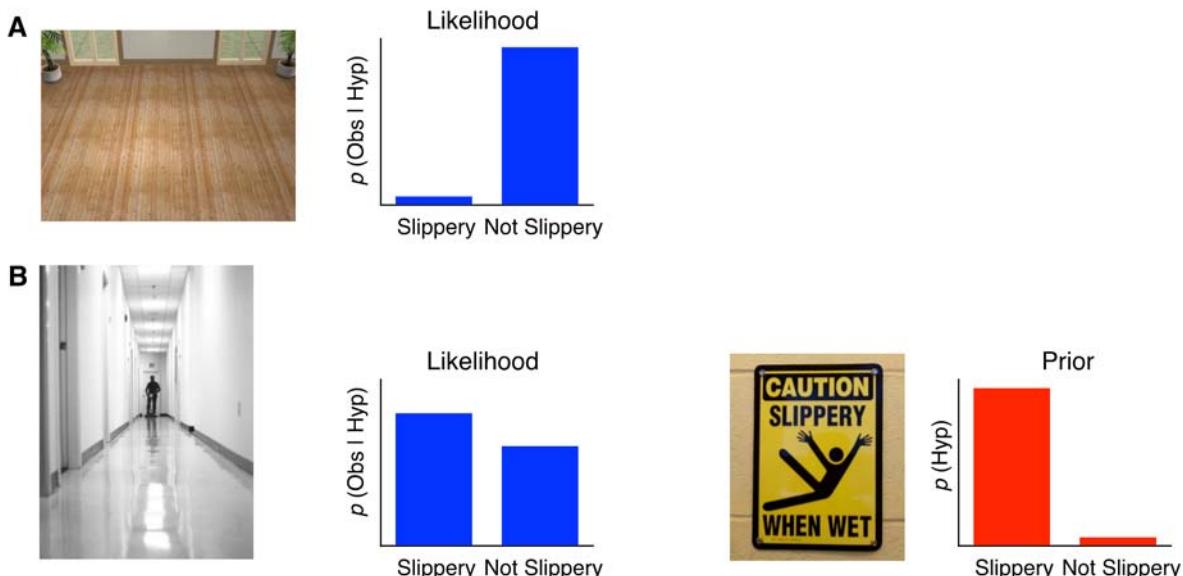
Many floors are both shiny and slippery, whereas many other floors are neither. Observing a shiny surface thus results in a relatively sharp likelihood function. The construction of the likelihood function requires background knowledge – in this case, our knowledge of the way various floors tend to reflect light. In general, to determine likelihoods, the observer needs to have an (implicit) understanding of the process by which different world states generate sensory data. In this case, the observer needs an intuitive understanding of optics: that a slippery floor tends to reflect light to a greater degree. To recognize the dependence of the likelihood on the observer’s background knowledge, we sometimes write the likelihood as  $p(\text{observation} | \text{world state}, B)$ , where  $B$  signifies information obtained through previous experience. This makes explicit that likelihood functions depend on background knowledge.

As in the problems discussed above, we need to calculate the posterior probability of each hypothesized world state,  $p(\text{world state} | \text{observation})$ . This calculation involves multiplying priors and likelihoods. Recall that the prior probability,  $p(\text{slippery})$ , reflects the observer’s expectation regarding the slipperiness of the floor, independently of the visual observation.

Before even entering a room, what probability would the observer assign to the hypothesis that the floor will be slippery? How does the observer acquire such priors?

The background knowledge that informs priors may have been acquired over a lifetime of previous experience, or very recently. If the observer has entered the same room in the recent past, and it has not been slippery, the observer's prior distribution will sharply favor the not-slippery hypothesis; if the observer has no previous experience with the room, but on the way to it passed through other rooms in the same building, and these were not slippery, the prior will again sharply favor the not-slippery hypothesis. Even if the observer is entering a building for the first time, the prior will favor (but not as sharply) the not-slippery hypothesis, provided that the majority of floors in the observer's experience are not slippery. By contrast, if the observer sees a newly posted *slippery when wet* sign, the prior will favor the opposite hypothesis. To recognize the dependence of the prior on the observer's background knowledge, we sometimes write the prior as  $p(\text{world state} \mid B)$ , where  $B$  again signifies information obtained through previous experience.

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

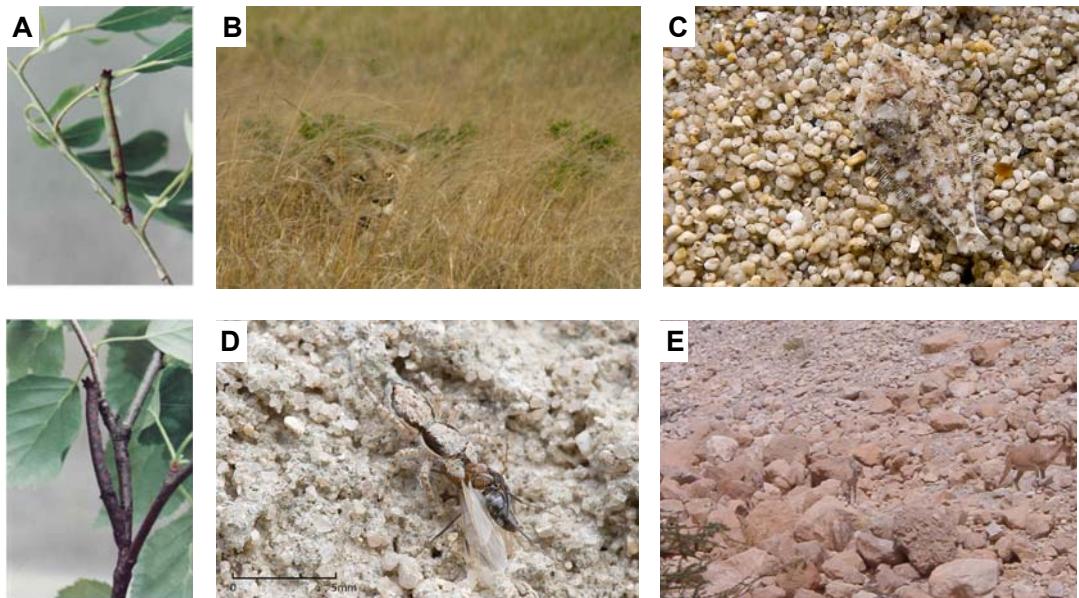


**Figure 10.** Perception of slipperiness. **A.** The likelihood function resulting from the visual image of this wood floor favors the “not slippery” world state:  $p(\text{observation} \mid \text{not slippery}) \gg p(\text{observation} \mid \text{slippery})$ . **B.** The shiny floor results in a likelihood function that favors the “slippery” world state:  $p(\text{observation} \mid \text{slippery}) > p(\text{observation} \mid \text{not slippery})$ . A “slippery when wet” sign would result in a sharp prior in favor of the “slippery” world state.

### 1.4.3 Camouflage

Flat likelihood functions present challenges to perception, and camouflage and mimicry can be

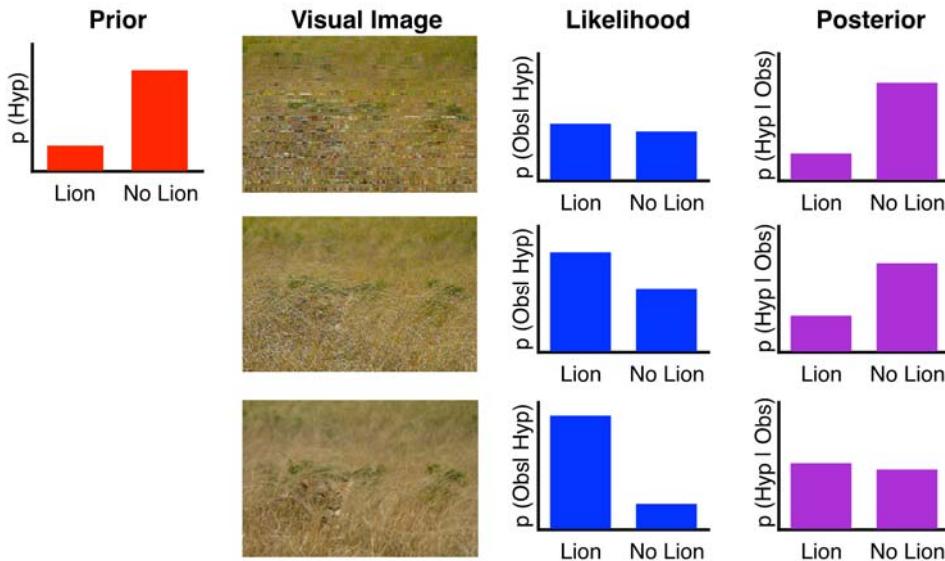
seen as strategies aimed at flattening likelihood functions. Many species have evolved traits and behaviors that serve to disguise their presence or their identity. Consider, for instance, the peppered moth caterpillar. Remarkably, individuals of this species assume the color of the tree bark on which they live (Fig. 11A). By blending in with the background, these caterpillars protect themselves from predatory birds. The visual image provides little indication of the caterpillar's presence. Camouflage is not exclusive to prey; predators, too, benefit from it. Consider the image of a lioness as she lies in waiting for her prey. Crouching low in the high golden grass, whose color closely resembles her own, she is nearly invisible until the moment she strikes. Although they can run fast, lions and other large cats lack stamina for long chases. Their success in hunting therefore depends on their ability to approach prey unnoticed. Examples of camouflaged predators and prey abound in the animal kingdom (Fig 11).



**Figure 11.** Likelihood function-flattening features in the animal kingdom. **A.** The peppered moth (*Biston betularia*) caterpillar changes its color to blend in with the background (above: willow; below: birch); from Noor et al., 2008. **B.** In tall golden grass of Kenya's Masai Mara National Reserve, a well-camouflaged lioness lies in waiting for wildebeest prey. **C.** A flounder against the sea floor. **D.** A well-camouflaged jumping spider with its captured ant prey (Dar es Salaam, Tanzania). **E.** Ibex in the Israeli desert.

When a well-camouflaged animal is viewed, (Fig. 12A, the observer's likelihood function does not clearly favor the animal's presence. Of course, as is always the case, the shape of the likelihood function results, not exclusively from the visual image, but also from the sensory abilities and acumen of the observer. A lion that to one observer is nearly perfectly camouflaged may be visible to another observer who has better visual acuity (Fig. 12). To an observer who knows from experience that peppered moth caterpillars tend to be slightly wider than the twigs of the tree they inhabit, the same visual scene (Fig 11A) will result in a sharper likelihood function

than it does for an observer who does not have this background knowledge. As long as the observer's likelihood function is not perfectly flat, the observer will learn something from the sensory input.



**Figure 12.** Effect of visual acuity on the posterior distribution. Three prey who hold identical (20%) prior expectation for the presence of a lion (upper left) differ in visual acuity, and therefore experience different likelihood functions when confronted with the same visual scene. The flatter the likelihood function, the more the posterior distribution resembles the prior distribution. **Top:** To this animal with poor visual acuity, the visual scene evokes a nearly flat likelihood function. The animal's posterior distribution is therefore similar to its prior distribution; it has learned little from the visual observation. **Middle:** An animal with intermediate visual acuity has a likelihood function that is not flat. This animal's posterior distribution differs slightly from its prior distribution. **Bottom:** For this animal with excellent visual acuity, the scene results in a sharp likelihood function in favor of the Lion's presence. The animal's posterior distribution indicates slightly greater than 50% probability that a lion is present..

Indeed, along with camouflage, evolution has given rise to sophisticated sensory systems – and cognitive abilities – that function to reduce uncertainty about the locations of other animals. In an arms race of sorts, animals have evolved progressively more sophisticated sensory systems to detect their (often progressively better-hidden) opponents. The evolution of mammalian visual, auditory and olfactory systems are cases in point, as is the evolution of highly specialized detection systems such as the ultrasonic echolocation used by insect-eating bat species. It can be argued that animals' perceptual systems have evolved to produce sharper likelihood functions.

## 1.5 Bayesian inference in auditory perception

So far, we have considered visual examples. However, nothing about inference is specific to vision. In this section and the next, we consider audition. Humans live in an acoustically rich

environment: birds chirp, the wind howls, dogs bark, car horns blare, music plays, and, perhaps most importantly, we talk to one another. Whether we are identifying the source of a sound (is that a dog barking?), perceiving its location (where is that barking dog?), or interpreting its meaning (what was that word you just said?), we use perceptual inference, combining likelihoods and priors to generate posterior probabilities.

### 1.5.1 Birds on a wire

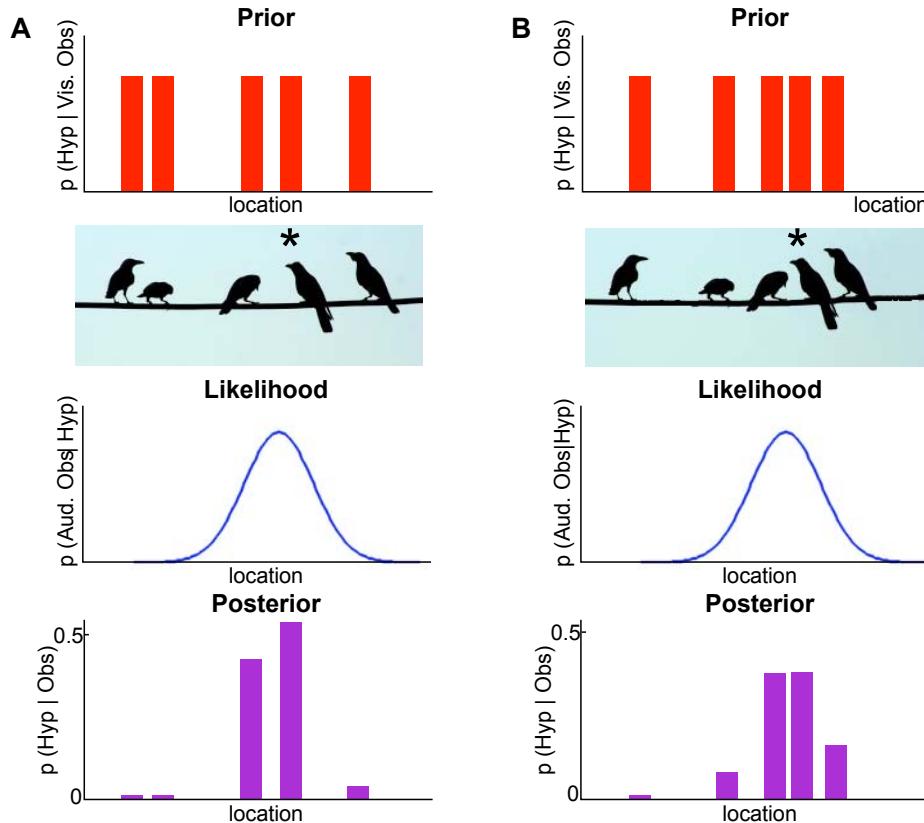
Humans often rely at least in part on our sense of hearing to locate objects. We and other mammals localize sounds sources by using sophisticated yet unconscious calculations, including comparing the intensity and time of arrival of sounds at the two ears. Nevertheless, our ability to localize sounds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

Suppose that you are walking outside on a beautiful sunny morning, when you notice the silhouettes of 5 birds perched on a wire (Fig. 13A). Suddenly, one of the birds (you cannot see which) bursts into melodious song. Which bird sang? Your auditory system rapidly processes the acoustic observation into a broad likelihood function. This likelihood function is a continuous function over location; that is, the sound you heard is compatible with a source at many possible (indeed, an infinite number of) locations. Nevertheless, certain locations are associated with higher likelihoods than others. Interestingly, the location of highest likelihood may well not coincide with the exact location of any bird. This situation is common in acoustic perception, and can be due to many factors. For instance, the bird that chirped was probably not facing you directly; sound can deflect off nearby objects before it reaches your ears; and noise in your own nervous system can cause the likelihood function to shift slightly in location from trial to trial even when the identical sound is repeated.

Unlike the likelihood function, the visual information is not continuous, but rather discrete. You see five individual birds. Thus, the prior distribution, based on your visual observation, is nonzero at just five discrete locations (here we are assuming that your visual perception is highly accurate for this high-contrast scene). Note that the prior probabilities are taken to be equal across the five birds, and the prior probability that the sound source would occupy an empty location on the wire is zero. This simply means that, prior to hearing the song, we consider it equally probable that any one of the birds will sing. Using Bayes' rule, we can now easily calculate the posterior distribution for the location of the sound source. For each of the hypothesized sound source locations, we multiply the likelihood by the prior. The resulting posterior distribution indicates that the singing bird was probably the second one from the right.

Intuition tells us that if the birds had been closer together on the wire, our inference would be less certain. This result indeed emerges from our Bayesian procedure, as shown in Fig. 13B. Here we show the singing bird at the same location, but with three of the other birds closer to it than they were before. Our prior distribution reflects the new positions of the birds, but the acoustic observation, and therefore the likelihood function, are the same as before. The posterior

distribution in this case is broader and lower than before, indicating that, although the same bird is still the most probable singer, our uncertainty has grown.



**Figure 13.** Sound source localization. **A.** The visual image of the birds provides the basis for a prior distribution over sound source location. The broad likelihood function reflects the imprecision of the acoustic observation. The posterior distribution favors the hypothesis that the second bird from right sang (\*). **B.** Perceptual uncertainty increases if the birds crowd closer together.

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

We will learn about a second alternative approach to this problem in Chapter 4, which again would reach the same answer: Starting with a flat prior over position, we could incorporate simultaneously both the visual and the acoustic observations as likelihood functions, in a procedure known as *cue combination*. In this approach, we would not use the visual information to generate a prior distribution for the subsequent auditory observation, but would instead combine a visual likelihood function with five discrete peaks with a continuous acoustic likelihood function. In essence, when we have two or more independent sources of information we can choose whether to incorporate the different sources sequentially, with the posterior from each observation being used as the prior for the next, or all at once, with all the observations entering through a likelihood function. Thus, there is often a blurring of boundaries between likelihood functions and priors, with the choice of how to incorporate the information left up to the Bayesian modeler. This flexibility is not a problem but a benefit of Bayesian inference. The internal consistency of the rules of Bayesian inference ensures that, as long as all the information is incorporated, the resulting posterior distribution will be the same regardless of the route taken. Within the Bayesian framework, there are often multiple ways of arriving at the same, useful solution.

### 1.5.2 Mondegreens

Although our brains do it automatically and apparently effortlessly, speech perception requires sophisticated inference on a variety of levels. Most obviously, we must correctly perceive the spoken word. It is easy to misinterpret even a single word spoken in isolation, particularly in the presence of ambient noise (the drone of a car engine, street sounds, chatter from other nearby speakers, and so on) and/or when the speaker is soft-spoken. Under such conditions, akin to low-contrast vision, likelihood functions are broad. It is instructive to keep a list of such occurrences. In recent conversations with others, we have misheard *Mongolia* as *magnolia*, *fumaroles* as *funerals*, *hogs* as *hawks*, *census* as *senses*, *a moth* as *I'm off*, *maple leaf* as *make believe*, *peaches and strawberries too* as *peaches and strawberries stew*.

As the last three of these examples illustrate, an additional challenge to speech perception arises because the pauses between spoken words are often no longer than the pauses between syllables within a single word. Thus, it is by no means a trivial task to infer where one word ends and the next begins. This difficulty leads to errors in which syllables from different words combine improperly in our perception. As a child, the author Sylvia Wright enjoyed listening to the popular 17<sup>th</sup>-century Scottish ballad, *The Bonny Earl o'Moray*, read to her frequently by her mother. She was particularly fond of the sad but beautiful lines describing the murder of the Earl and of his love, the Lady Mondegreen:

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

In fact, the words heard by the young Sylvia Wright were not those that her mother spoke. The last line of the ballad actually reads: “And laid him on the green.” The unfortunate dead Earl was placed on the grass, alone – Lady Mondegreen existed only in Sylvia Wright’s mind! Sylvia Wright’s creative but mistaken interpretation of the spoken ballad reflects a parsing error. She heard the sounds “laid hi-” as “lady,” and “-m on the green” as “Mondegreen.”

Sylvia Wright later coined the term “mondegreen” to refer to a misheard word or phrase. Given the inherent phonetic ambiguity of spoken language, examples of mondegreens abound. When Queensland, Australia was inundated by tropical cyclone Tasha, the Morning Bulletin of Rockhampton (Jan 6, 2011) reported that, as a result of the flooding, “More than 30,000 pigs have been floating down the Dawson River since last weekend.” This information, based on an interview between the reporter and the owner of a local piggery, was staggeringly incorrect. The owner had spoken, not of “30,000 pigs,” but of “30 sows and pigs” floating downstream! The Morning Bulletin published a correction the next day.

Books and many websites are devoted to listing peoples’ favorite mondegreens, particularly those resulting from misheard song lyrics, which we can all enjoy because we share access to the songs. It is instructive to visit websites on which listeners post their particular misheard versions of the same songs. The many different misheard versions of a line such as “Lucy in the sky with diamonds” presumably reflect both the phonetic ambiguity (broad likelihood function) and improbable content (low prior probability) of those lyrics. With respect to prior probabilities, “There’s a bathroom on the right” is surely a more common sentence than “There’s a bad moon on the rise”, and “submarine” is arguably more plausible than “summer breeze” as a mode of transport (Fig. 14).

**Lucy in the sky with diamonds**  
The Beatles

“Lucy in disguise with lions”  
“Lucy and this guy eat ions”  
“Lucy and this guy are dying”  
“Lucy and this guy at Dinah’s”  
“You’ll see in the sky McDonald’s”

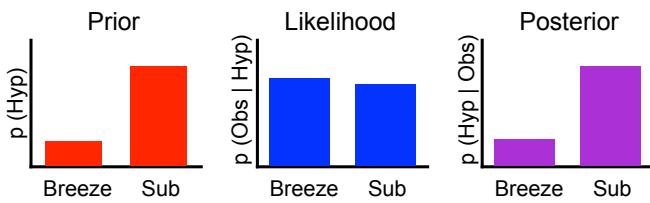


There's a bad moon on the rise  
Creedence Clearwater Revival

And you come to me on a summer breeze  
Bee Gees (How Deep is Your Love)

“There's a bathroom on the right.”

“And you come to me on a submarine”



The Death of Lady Mondegreen  
(Harpers Magazine, Nov. 1954)

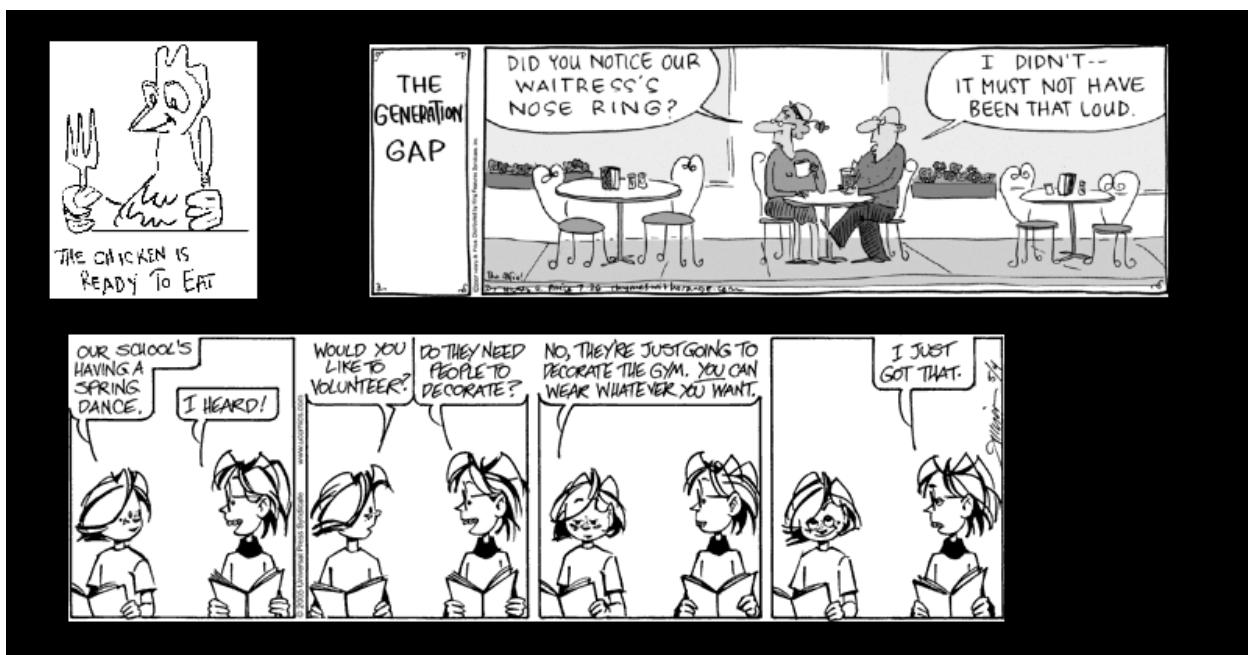
**Figure 14.** Mondegreens result from phonetic ambiguity (broad likelihood functions) coupled with low expectation for the actual phrase that was sung or spoken (prior distribution in favor of the “wrong” hypothesis).

An important lesson to take away from mondegreens is that speech perception is based, as is any perceptual inference, in the combination of likelihood functions and prior distributions. We generally perform this task very well, but of course occasional mistakes are inevitable. Indeed, “The more unintelligible the original lyrics, the more likely it is that listeners will hear what they want to hear” (O’Connell, 1998). In the terms of Bayesian inference, the flatter the likelihood function, the greater will be the influence of the prior distribution on the resulting posterior distribution (just as in Fig. 12). Thus, the same feature of perception that usually serves us so well – our incorporation into perceptual inference of our prior expectation – backfires to create mondegreens when we are faced with an unexpected word (low prior) that sounds like (broad likelihood) another, more expected (high prior) word.

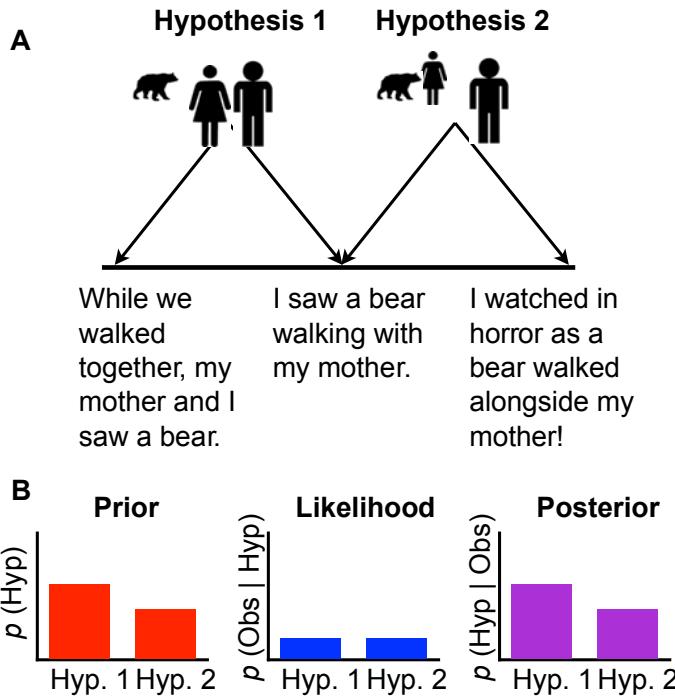
Keeping this in mind, it is easy to evoke mondegreens in others: simply select two different words or phrases that sound alike, ensure that your listener has a prior distribution in favor of one of the words or phrases, then speak the other. You may wish to try the following demonstration with a friend.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

broad likelihood function experienced by your friend will combine with a sharp prior distribution (given the previous content of your discourse) to favor the “recognize speech” hypothesis.

Even when the listener perceives every word correctly, she faces a final crucial challenge: to identify the intended meaning of the string of words. Once again, this often requires evaluating multiple hypotheses. Consider the sentence, “The bridge is being held up by red tape.” This sentence, even when perfectly heard, is nevertheless consistent with two interpretations; that is, it evokes a broad likelihood function, due not to phonetic ambiguity but rather to syntactic ambiguity. In fact, sentences such as this occur quite commonly in English (Fig. 15). When we hear such a sentence, or read it, we naturally combine the likelihood functions with a prior distribution, and usually reach the correct perception. We are sometimes bemused – and amused – momentarily, however, as both interpretations cross our minds. This occurred recently to one of the authors when a friend told him of a wilderness trip he took with his parents. “There was wildlife everywhere,” he exclaimed, “In fact, I saw a bear walking with my mother” (Fig. 16). Although this book will not focus on this type of semantic inference, we point it out to illustrate that uncertainty and inference play a role at all levels of brain function.



**Figure 15.** Syntactic ambiguity in language.



**Figure 16.** Perceptual inference under syntactic ambiguity. **A.** Each world state could be described in many different ways, a few of which are shown. The speaker happened to choose an expression that could describe both world states: “I saw a bear walking with my mother.” **B.** Bayesian perceptual inference. Background knowledge suggests that bears are less likely to walk alongside people than to be seen by them at a distance, so the prior distribution favors Hypothesis 1. The likelihood function shows that the spoken sentence has about equal probability under the two hypotheses. The posterior distribution therefore favors Hypothesis 1.

## 1.6 Concluding remarks

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

- The likelihood function summarizes the information content of the sensory observation, relevant to distinguishing one world state from another.
- Perception is not based entirely on sensory observation, but also on expectation grounded in previous experience. We express expectation as a prior distribution over world states.
- Bayes’ rule calculates the posterior probability of each possible world state from the likelihoods and prior probabilities of the world states.

- The posterior probability of a world state is not the same as the likelihood of the world state. In general,  $p(A | B) \neq p(B | A)$ .
- Flat likelihood functions pose a challenge to perception. The flatter the likelihood function, the more the posterior distribution resembles the prior distribution. If the likelihood function is perfectly flat, then the posterior distribution is identical to the prior distribution, and the observer has learned nothing from the observation.
- The procedures of Bayesian inference apply equally to situations in which the hypothesized world states are discrete or in which they are continuous.
- Like visual perception, auditory perception is well described as a process of unconscious Bayesian inference.
- Bayesian inference can be done iteratively, a process in which the posterior distribution from one inference is used as the prior distribution for the next. For example, a posterior distribution based on a previous observation in one modality (e.g., vision) can be used as a prior distribution for a subsequent inference based on a later observation in another modality (audition).
- Speech is fraught with phonetic and syntactic ambiguity, giving rise to flat likelihood functions in many instances. As with other perceptual inference, posterior distribution s in speech perception reflect the influence of likelihoods and priors. When likelihoods are nearly flat, priors exert a greater influence on the posterior distribution. This can cause misinterpretations such as mondegreens.

## 1.7 References

Bayes, T. (1763). An Essay Towards Solving a Problem in the Doctrine of Chances. Philosophical Transactions of the Royal Society of London, 53: 370–418.

Brugger P, Brugger S (1993) The Easter bunny in October: Is it disguised as a duck? Perceptual and Motor Skills 76: 577-578.

Burdon D (Jan. 6, 2011) Pigs float down the Dawson. The Morning Bulletin.

Fenton N (Nov. 3, 2011) Improve statistics in court. Nature 479:36-37

Helmholtz, Hermann Ludwig von (1925) Treatise on Physiological Optics, III: The Perceptions of Vision (1910); Southall JPC, ed. RochesterN.Y.: Optical Society of America.

Laplace, Pierre Simon, *Philosophical Essay on Probabilities*, translated from the fifth French edition of 1825 by Andrew I. Dale. Springer-Verlag: New York (1995).

Noor MA, Parnell RS, Grant BS (2008) A reversible color polyphenism in American peppered moth (*Biston betularia cognataria*) caterpillars. *PLoS One* 3:e3142.

O'Connell, PL. Sweet Slips Of the Ear: Mondegreens. *New York Times* (Aug. 9, 1998).

Rosenhouse, Jason, *The Monty Hall Problem: The remarkable story of math's most contentious brain teaser*. Oxford University Press: New York (2009).

Smith, A. Mark, Ed. (2001) *Alhacen's Theory of Visual Perception: A Critical Edition, with English Translation and Commentary, of the First Three Books of Alhacen's *De aspectibus*, the Medieval Latin Version of Ibn al-Haytham's *Kitab al-Manazir**. Transactions of the American Philosophical Society, volume 91, parts 4-5.

Wright, S. (Nov, 1954) The Death of Lady Mondegreen. *Harper's Magazine*, 209: 48-51.

## 1.8 Further reading

### History and applications of Bayesian Inference

McGrayne, Sharon Bertsch, *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines & emerged triumphant from two centuries of controversy*. Yale University Press: New Haven (2011).

### Evolution of perceptual systems and camouflage: a Bayesian perspective

Geisler, W. S. and Diehl, R.L. (2003) A Bayesian approach to the evolution of perceptual and cognitive systems. *Cognitive Science* 27: 379-402.

### Perception as Unconscious Inference

Hatfield, G. (2002). *Perception as Unconscious Inference*. In *Perception and the Physical World: Psychological and Philosophical Issues in Perception*, ed. by Dieter Heyer and Rainer Mausfeld (New York: Wiley), 115–143.

### Phonetic and syntactic ambiguity

Smith, R. *Milk drinkers turn to powder and other pun-ishing headlines*. *Globe and Mail* (Sept. 24, 2009)

*Red tape holds up new bridge, and more flubs from the nation's press*, Gloria Cooper, ed.; collected by the Columbia journalism review (1987)

## 1.9 Problems

1. If  $A$  is the event “a person is old,” and  $B$  is the event “a person suffers from Alzheimer’s disease,” is  $p(A|B)$  less than, equal to, or greater than  $p(B|A)$ ? Why?
2. Generate three comparisons of your own, of the type  $p(A|B) \neq p(B|A)$ . In each case, state which probability is greater, and explain why.
3. At a particular university, 15% of all students are in humanities, 58% of all students are undergrads, and 19% of undergrads are in humanities. What is the probability that a random humanities student is an undergrad?
4. 1% of the population suffers from disease D. A diagnostic test for D is being piloted. The probability that someone without D tests positive (false-alarm rate) is 2%. The probability that someone with D tests negative (miss rate) is 3%.
  - a) Make a quick guess of the probability that someone who tests positive actually has D.
  - b) Calculate this probability. If it is very different from your answer to a), what went wrong in your intuition?
  - c) (\*) (Due to Huihui Zhang, Beijing University) Suppose now that there is an extra variable we have ignored, namely whether someone goes to the doctor to have a diagnostic test done. This probability is higher if someone has the disease (because there will likely be symptoms) than if someone does not have the disease. Assume a 5-to-1 probability ratio for this. Now recalculate the probability that someone who tests positive actually has D. Is it closer to your original intuition?
5. If you look carefully at the probability graphs in the figures of this chapter, you will notice that the prior probabilities of the different hypotheses sum to one, as do the posterior probabilities of the different hypotheses. The likelihoods, however, do not generally sum to one. Why do likelihoods not have to sum to 1, whereas priors and posteriors do? Hint: think carefully about the definitions of likelihood, prior and posterior.
6. Prove using Bayes’ rule that when the likelihood function is perfectly flat, the posterior distribution is identical to the prior distribution.
7. (See Section 1.2.) Prove using Bayes’ rule that if you see a bag on the luggage carousel that does not match yours (for instance, a small red bag, when yours is large and black), the posterior probability that it is yours is zero.

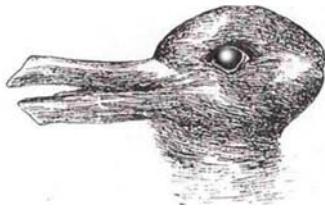
8. \* (See Section 1.2.) You are one of 100 passengers waiting for your bag at an airport luggage carousel. Your bag looks the same as 5% of all bags. Derive a general expression for the probability that the bag you are viewing (which matches your bag visually) is your own, as a function of the number of bags you have viewed so far. How many bags must you view (without finding your own) before the posterior probability that the bag you are viewing (which matches your own visually) is greater than 70%?
9. (See Section 1.2.) In the luggage carousel example, we defined the visual observation as the shape, size, and color of the bag seen, and we therefore took  $p(\text{observation} | H_1)$  to equal 1 when the observation matched the shape, size, and color of your bag. But the exact “look” of the bag on the luggage carousel involves more than just its shape, size, and color. For instance, as the bag enters the carousel, it may come to rest at any one of many different orientations. Now suppose we were to redefine the “observation” to be shape, size, color and *orientation* of the bag seen. To keep things simple, let’s assume that there are 360 possible angles (one for each degree around the circle) and two possible sides (right-side up or upside down), for a total of 720 possible orientations with which a bag may come to rest on the carousel. If we further assume that each orientation is equally likely, then the probability of the observation given hypothesis 1 is no longer one, but rather  $1/720$ . Similarly, the probability of the observation given hypothesis 2 would no longer be 0.05, but rather  $0.05/720$ . Since the likelihoods have changed, must not the posterior distribution change as well? Explain.
10. Suppose you are waiting to catch a particular bus in a city that has just 10 bus routes; the route followed by each bus is indicated by an integer in the corner of its front display. You see the bus below from a distance, and naturally wonder whether this is the bus you are waiting for. A) Based on the visual image of the difficult-to-discriminate bus route number (see arrow), and your intuitive understanding of how different numbers might appear, construct a likelihood function that plots  $p(\text{visual observation} | \text{hypothesized bus route})$ , for all numbers from 1 to 10. B) As it turns out, you happen to know that only buses 3, 4, 5, and 6 travel down the street you are on. Furthermore, you know that buses 3 and 4 come twice as frequently as buses 5 and 6. Based on this background knowledge, construct your prior distribution for the bus number. C) Use Bayes’ rule to calculate your posterior distribution for the number of the bus.



11. Rephrase in terms of Bayesian perceptual inference the following statement written by Ibn Alhacen approximately 1,000 years ago: "...when sight perceives a rose-red color among the flowers in some garden, it will immediately perceive that the things in which that color inheres are roses because that color is specific to roses...But this does not happen when sight perceives a myrtle-green color in the garden. For when sight perceives only the myrtle-green in the garden, it will not perceive the myrtle-green to be myrtle simply from the perception of the green, because several plants are green, and, in addition, several plants resemble myrtle in greenness and shape." (*De aspectibus*, book 2, as translated by Smith, 2001).
12. Why is it that we identify ourselves at the very beginning of a phone conversation, even to people we already know, but we do not do this when we meet in person? Express your answer within the framework of Bayesian perceptual inference.
13. When a conversation companion speaks softly, or when a conversation occurs in the presence of significant ambient noise, we sometimes cup our ears and/or look carefully at the speaker's lips. Why, in Bayesian perceptual terms, do we do this?
14. To explore how a noisy environment engenders uncertainty, consider the word "lunch." Suppose that you see this word written (or hear it spoken), with the letter "l" blocked out: \_unch (e.g., by ambient auditory noise). List all source words that are compatible with what you see. Now repeat, but with the letter "n" blocked: lu\_ch. Finally, consider the case in which both the l and the n are blocked: \_u\_ch. In terms of conditional probabilities relevant to perception, what is the effect of blocking out the l, n, and both?
15. The NATO phonetic alphabet, used by many military, maritime, and other organizations during radio communications, represents each letter with a word: A (Alpha), B (Bravo), C (Charlie), D (Delta), E (Echo), F (Foxtrot), and so on. What purpose does this serve in radio communications? Explain with respect to conditional probabilities. In particular,

consider a radio communication under conditions of considerable background noise, in which the sender wishes to spell the word “FACE.” Compare  $p(\text{auditory signal heard by the receiver} \mid \text{FACE spelled by the sender})$  vs.  $p(\text{auditory signal heard by the receiver} \mid \text{another word, such as FADE, spelled by the sender})$ , when the sender uses the regular alphabet, and again when the sender uses the NATO phonetic alphabet.

16. English speakers sometimes incorrectly perceive English words when they listen to songs sung in a foreign language with which they are unfamiliar, and listeners also mistakenly perceive words in music that is played backwards. Provide a Bayesian explanation for these phenomena.
17. Suppose you see someone you do not know, getting only a brief look at him from a distance of about 10 meters. If your interest is in estimating this person’s age, how would you proceed? What factors, including and in addition to the person’s appearance, would affect your estimation? Provide a Bayesian description of your reasoning. As part of your answer, draw examples of your likelihood function, prior distribution, and resulting posterior distribution.
18. A research article entitled “The Easter bunny in October: Is it disguised as a duck?” explained that “Very little is known about the looks of the Easter bunny on his non-working days.” To investigate, the authors showed an “ambiguous drawing of a duck/rabbit...to...265 subjects on Easter Sunday and to 276 different subjects on a Sunday in October of the same year.” The authors report that “Whereas on Easter the drawing was significantly more often recognized as a bunny, in October it was considered a bird by most subjects.” The drawing shown by the authors in their study was similar to the following:



Provide a Bayesian perceptual explanation for the authors’ results.

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



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

## Contents

2	Chapter 2: Building a Bayesian model .....	2-1
2.1	An example of a psychophysical experiment.....	2-1
2.2	The three steps of Bayesian modeling.....	2-2
2.3	Step 1: The generative model.....	2-4
2.3.1	The stimulus distribution .....	2-5
2.3.2	The measurement.....	2-8
2.3.3	The measurement distribution.....	2-8
2.3.4	Joint distribution .....	2-9
2.4	Step 2: The inference process.....	2-10
2.4.1	The prior distribution .....	2-10
2.4.2	The likelihood function.....	2-11
2.4.3	The posterior distribution.....	2-14
2.4.4	The maximum-a-posteriori (MAP) estimate.....	2-20
2.5	Step 3: The distribution of the MAP estimate.....	2-21
2.6	Sanity checks.....	2-24
2.7	Concluding remarks .....	2-25
2.8	Problems.....	2-26

## 2 Chapter 2: Building a Bayesian model

In Chapter 1, we introduced the concepts of probability, inference, and Bayes' rule in a variety of daily-life examples. Although these concepts make great intuitive sense, there is much more to Bayesian modeling than intuitive explanations. One of the great strengths of the Bayesian modeling approach is that it allows for precise mathematical predictions for behavior. The goal of this chapter is to do that for a simple task. This chapter differs from Chapter 1 in that we use a more mathematical formalism; for an introduction to probability theory, you may wish to read the first few sections of Appendix A at this point.

### 2.1 An example of a psychophysical experiment

Psychophysics is the study of how controlled stimuli are perceived or acted upon by organisms. For example, an experimenter might show you two lines on a computer screen and ask you which is longer. When the lines are very similar in length, this is a difficult task and you will make mistakes. These mistakes can tell the experimenter about the way you are solving the task. Researchers have used psychophysics for more than a century to probe the nature of perceptual processing.

The psychophysical task we will use as the leading example in this chapter is an auditory localization task. Imagine that you are facing a projection screen that displays a horizontal line stretching across the width of the screen. Located behind the screen, at the same elevation as the line, is a densely spaced array of very many tiny loudspeakers. A tone will originate from one of these speakers. Your task is to report with a cursor the location from which you perceived the tone to emanate. This task is repeated many times; each repetition is called a trial. In this experiment, you are estimating a continuous quantity, namely the position of a sound source along a line. You have sensory observations, but possibly also prior knowledge. The steps involved in building a Bayesian model for this simple task provide a complete recipe for building and applying any Bayesian model.

## 2.2 The three steps of Bayesian modeling

The observer's task is to infer the value of a world state of interest (here sound location) from a given sensory observation or multiple sensory observations (here the sound waves impinging on your eardrum). This inference process can be modeled using a Bayesian model. Every Bayesian model consists of three steps that must be followed in a particular order (see also inside front cover). These steps are:

- 1) Defining the *generative model*: the probabilities of world states and of sensory observations given world states
- 2) Specifying the *inference process*: the observer's estimation of the state of the world based on the sensory observations and prior expectations
- 3) Computing the observer's *estimate distributions*: the predicted distribution of the observer's estimates.

Mistakes can easily arise from not clearly separating these three steps. We now preview each step.

**Step 1: Defining the generative model.** The generative model (also called forward model) describes the statistics of the task-relevant variables, including the observer's sensory observations. It is a full statistical description of what is happening in a task. Variables in the generative model always include the world state of interest and the observer's sensory observations, but could also include other variables. The generative model specifies the probability distributions over all variables in the task. In the luggage example of Chapter 1, the generative model specifies the probability that a random bag is yours, the probabilities of different visual images produced by your bag, and the probability of different visual images produced by bags that are not yours.

Many of the distributions in the generative model are specified by the experimental design. For instance, the probability of a sound occurring at a particular location is specified in the experimental design. However, we need to make an assumption about the distribution of the sensory observations. Most examples in this book will allow for the possibility that sensory observations are noisy. "Noise" has a diverse set of meanings across distinct scientific fields, but

in this book, we mean that the same stimulus does not always produce the same internal representation in the brain. Noise thus implies random variability of the sensory observations from trial to trial. Noise can be due to a wide variety of factors, both external (in the world) and internal (in the brain).

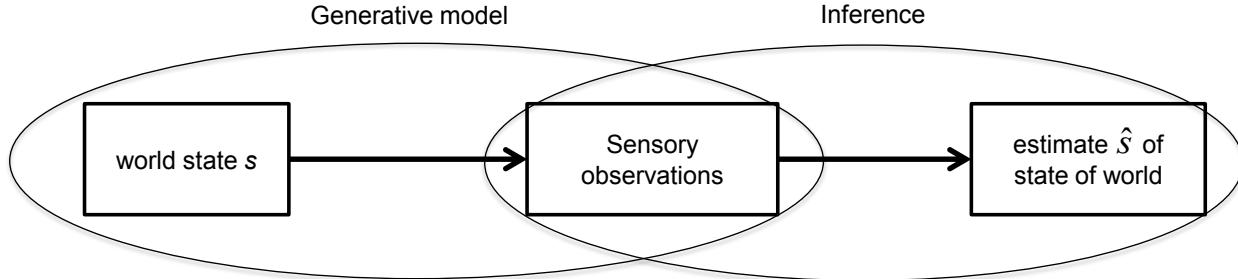
**Step 2: Deriving how the observer performs inference.** In this step, we compute the observer's posterior distribution, i.e. the observer's probability distribution over the world state of interest, given the sensory observations. The observer's inference process "inverts" the generative model, in order to reach a conclusion about the world state in light of the sensory observations. In the luggage example, the inference process consisted of computing the probability that the bag is yours from the sensory observations and prior information. The generative model as understood by the observer, along with the sensory observations, completely defines the posterior distribution; no additional information is needed.

In this chapter, we assume that the generative model believed by the observer is the same as the actual generative model specified in step 1. This does not have to be the case; for example, you might believe that 10%, instead of 5%, of all bags look like yours. We will reconsider this assumption in Chapter 3.

The last step in specifying the observer's inference process is to specify how the observer obtains an estimate of the world state from the posterior distribution. The most common recipe is to choose the hypothesis that has the highest posterior probability – this recipe is known as maximum-a-posteriori estimation. The inference process is typically, but not always, a deterministic function of the sensory observations: for given sensory observations, the estimate of the world state is always the same.

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

**Figure 2.1.** Schematic of a Bayesian model.



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

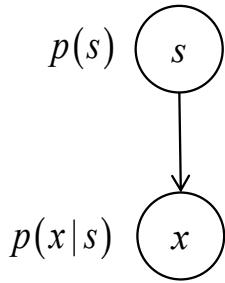
The generative model describes the input into the decision-making process (Fig. 2.1). The inference process describes the observer's calculation of a posterior probability distribution over world states, and selection of a world state estimate; in the end, the inference process is summarized as an input-output relationship between the sensory observations and the world state estimate. The estimate distribution, how frequently the subject will exhibit each possible behavior, is obtained by combining this input-output relationship with the distribution of the sensory observations.

In the following sections, we undertake the full Bayesian modeling procedure for the example auditory localization task described above. Despite its simplicity, this example illustrates the three steps and captures many of the subtleties of Bayesian modeling.

### 2.3 Step 1: The generative model

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

Thus, the problem contains two variables: the stimulus (true sound location,  $s$ ) and the observer's internal measurement of the stimulus,  $x$ . These two variables appear in the generative model, which is depicted in Fig. 2.2.



**Figure 2.2.** The first step in Bayesian modeling is to define the generative model. This diagram is a graphical representation of the generative model discussed in this chapter. Each node represents a random variable, each arrow an influence. Here,  $s$  is the true stimulus and  $x$  the noisy measurement of the stimulus.

A diagram like this – also called a graphical model – consists of nodes that contain the random variables and arrows that represent stochastic dependencies between variables. Each node is associated with a probability distribution. The variable at the end of an arrow has a probability distribution that depends on the value of the variable or variables at the origin of the arrow. In other words, an arrow can be understood to represent the *influence* one variable has on another. The arrow can be read as “produces” or “generates” or “gives rise to”, e.g. “the sound location  $s$  gives rise to a measurement  $x$ ”.

In our problem, there are only two variables: the world state or stimulus  $s$  and its measurement  $x$ . No arrow points to  $s$ , and therefore the distribution of  $s$  has a prior distribution  $p(s)$ . This distribution represents the overall frequency of occurrence of each possible value of the stimulus. The arrow pointing from  $s$  to  $x$  indicates that the distribution of  $x$  depends on the value of  $s$ . Mathematically, this is expressed as a *conditional probability distribution*  $p(x|s)$ . Conditional probability distributions are formally defined in Appendix A, Section 10. We will now describe the components of the generative model in detail.

### Study tip

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

### Box 2.1. Notation for probability distributions

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

### 2.3.1 The stimulus distribution

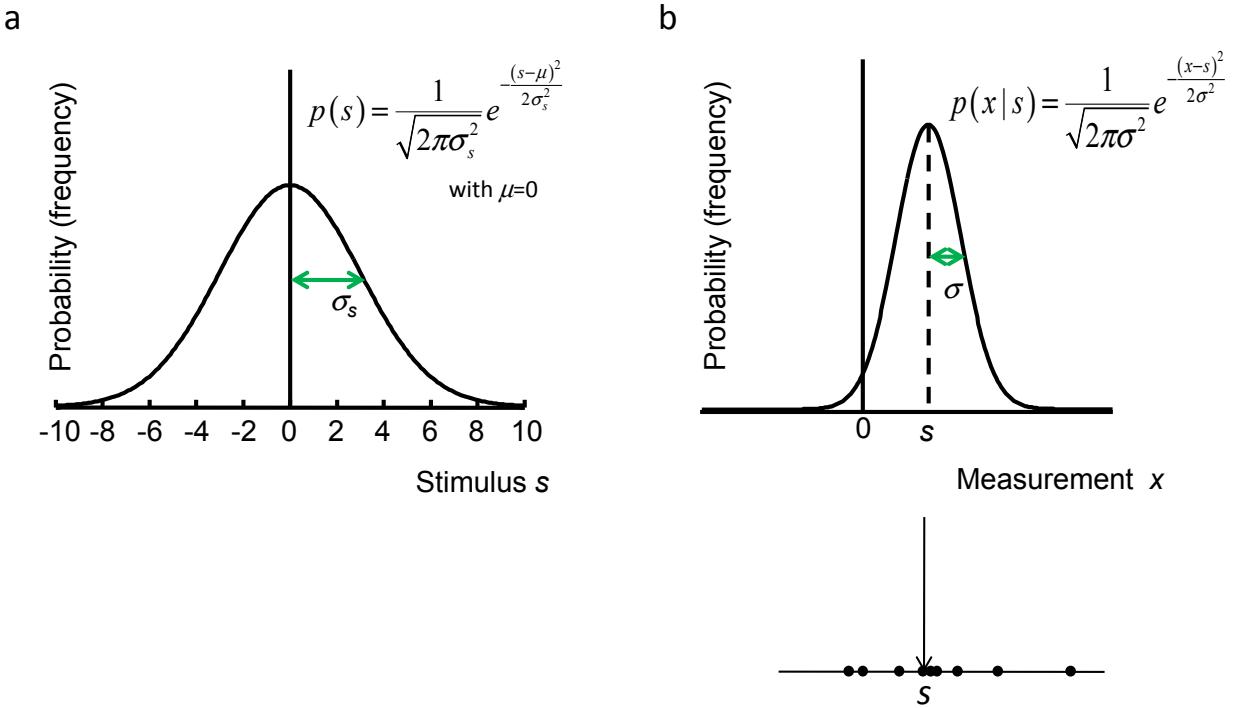
The distribution associated with the stimulus  $s$  is denoted  $p(s)$ . This *world state distribution* or in our current example, *stimulus distribution*, reflects how often each possible value of  $s$  occurs. In

models of cognition,  $p(s)$  would be called the *base rate* of  $s$ , but this terminology is less common in perceptual modeling.

In our example, the experimenter has programmed a computer to draw the stimulus on each trial from a Gaussian distribution with a mean  $\mu$  and variance  $\sigma_s^2$ . This distribution is defined by the density (see Box):

$$p(s) = \frac{1}{\sqrt{2\pi\sigma_s^2}} e^{-\frac{(s-\mu)^2}{2\sigma_s^2}} \quad (2.1)$$

and is depicted in Fig. 2.3a for  $\mu=0$  and  $\sigma_s=3$  (as units, you can think of degrees of visual angle, but this is not essential). This Gaussian shape with mean zero implies that the experimenter more often presents the tone straight ahead than at any other location.



**Figure 2.3.** The probability distributions that belong to the two random variables in the generative model. (a) A Gaussian distribution over the stimulus,  $p(s)$ , reflecting the frequency of occurrence of each stimulus value in the world. The unit of  $s$  is arbitrary but could be degrees of visual angle in the localization example. In many plots, we will leave out numerical values altogether. (b) Suppose we now fix a particular value of  $s$ . Then the measurement  $x$  will follow a Gaussian distribution around that  $s$ . The diagram at the bottom shows a few samples of  $x$ .

### Box 2.2: The normal/Gaussian distribution

The most frequently used continuous probability distribution is the normal or Gaussian distribution. Its density function is

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \quad (2.2)$$

This is the famous “bell-shaped” distribution (or bell curve). We will sometimes use the notation  $p(y) = \text{Normal}(y; \mu, \sigma^2)$  as short-hand for Eq. (2.2). The parameters  $\mu$  and  $\sigma^2$  are the mean and variance of the distribution, respectively. The factor  $\frac{1}{\sqrt{2\pi\sigma^2}}$  is needed so that the total probability – the integral of  $p(y)$  – is equal to 1. The exponent,  $-(y-\mu)^2/(2\sigma^2)$ , has a maximum value of zero at  $y=\mu$ , which is therefore the maximum of the Gaussian distribution. From this maximum outward, the exponent decays. It will be -1 once the difference between  $y$  and  $\mu$  has reached  $\sigma\sqrt{2}$ . There, the Gaussian will have decreased by a factor of  $e$ .

Gaussian distributions result when many randomly occurring fluctuations can affect the variable of interest. The more formal version of this statement is called the *Central Limit Theorem*. A typical example is the height of people, which follows a roughly Gaussian distribution, presumably because many factors contribute to height. Suppose that the average height for males is 175 cm, with a standard deviation of 10 cm. In this case we would find many males between 165 and 185 cm (within one standard deviation from the mean), fewer between 155 and 165 and between 185 and 195 cm, and very few above 195 cm or below 155 cm (more than two standard deviations away from the mean).

Exercise 2.1: If one substitutes  $y=\mu$  and  $\sigma=0.1$  in Eq. (2.2), one finds  $p(y) = 3.99$ . How can a probability be larger than 1? If the answer to this question is not immediately clear, read Section 5 of the Appendix, on the difference between probability mass and density functions.

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

Gaussian distributions are convenient for analytical calculations; for example, multiplying two Gaussians produces another Gaussian (see Box 2.X). Gaussian distributions are also convenient for simulations; for example, to draw samples from a Gaussian distribution

with mean  $\mu$  and standard deviation  $\sigma$ , one can draw samples from one with mean 0 and standard deviation 1 (a *standard normal distribution*), multiply them by  $\sigma$ , and add  $\mu$ .

### 2.3.2 The measurement

A physical stimulus elicits activity in the nervous system. This activity will vary randomly from trial to trial even when the physical stimulus itself is identical each time. Such variability originates from many sources. Our sensors are subject to random variability due to intrinsic stochastic processes. For instance, thermal noise affects the responses of hair cells in the inner ear that sense sound waves. The transduction process by which the nervous system captures physical energy and converts it into an electrical response is also stochastic. For instance, the absorption of photons by photoreceptors is a stochastic process; only sometimes is there a response to a single photon. At the subcellular level, neurotransmitter release and ion channel opening and closing are stochastic processes.

In some cases, this noise can be easily illustrated. For example, if we place the index finger of our right hand on top of a table and try to place the index finger of our left hand at the matching location underneath the table, we often observe quite a difference (typical variability is about 2 cm in this task). This indicates noise in our internal proprioceptive representations of limb location. Similarly, it is difficult to estimate whether one object is heavier than another based on our sense of force because the internal measurement of force is noisy – necessitating the use of scales to compare weights. These examples suggest that the relationship between stimulus and sensor response is stochastic.

We define a *measurement* as the representation in stimulus space of the internal neural representation of the stimulus. For example, if the true location  $s$  of the sound is  $3^\circ$  to the right of straight ahead, then its measurement  $x$  could be  $2.7^\circ$  or  $3.1^\circ$ . The terminology “measurement” stems from the analogy with making physical measurements. If a stick is 89.0 cm long, you might measure its length to be 89.5, 88.1, 88.9 cm, or so on. We say that the measurement “lives” in the same space as the stimulus, because it has the same units as the stimulus.

### 2.3.3 The measurement distribution

The *measurement distribution* is the distribution of the measurement  $x$  for a given stimulus value  $s$ . This *conditional distribution*,  $p(x|s)$ , describes the frequency of occurrence of each value of the measurement when the same stimulus value  $s$  is repeated many times.

If many sources contribute to the variability of the measurement, we will end up with a measurement distribution that is roughly Gaussian. This assertion is – loosely – a consequence of the Central Limit Theorem (see Box 2.2). While the Gaussian form of the stimulus distribution, Eq. (2.1), is often chosen simply because it facilitates calculations, the Gaussian form of the measurement distribution is quite fundamental, independent of the experimental design, and common to most Bayesian models we discuss in this book.

Thus, the equation for the measurement distribution is

$$p(x|s) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-s)^2}{2\sigma^2}}, \quad (2.3)$$

where  $\sigma$  is the standard deviation of the noise in the measurement, also called *measurement noise level* or *sensory noise level*<sup>1</sup>. This Gaussian distribution is shown in Fig. 2.3b for one value of  $s$ . The higher  $\sigma$ , the noisier the measurement and the wider its distribution. Lowering the light level in a room, increasing the distance to an object, decreasing the presentation time, or removing your corrective eyewear are all ways to increase the standard deviation of the measurement noise of a visual stimulus. For an auditory or tactile stimulus, the same is achieved by introducing background noise, decreasing the intensity of a stimulus, or decreasing the presentation time. The inverse of the variance of the measurement distribution,  $1/\sigma^2$ , is sometimes called the *reliability* or the *precision* of the measurement  $x$ .

### Box 2.1. Noise and Ambiguity

As explained in Chapter 1, there are many sources of uncertainty in perception. In this chapter, we consider uncertainty that arises from noise in the observer's sensory measurement. Because our sensory systems are universally subject to measurement noise, this form of uncertainty is always present to some degree in perception. However, uncertainty can additionally arise from ambiguity in the stimulus itself: different world states can produce the same sensory stimulus. An example of an ambiguous image is a shiny floor, which may or may not be slippery (Chapter 1, Figure 10). Later in the book, we will encounter further examples of ambiguous images (e.g., size-distance ambiguity: see Chapter 6). Whether uncertainty is caused only by sensory noise, or also by stimulus ambiguity, the end result is that the observation has nonzero probability given more than one hypothesized world state.

### 2.3.4 Joint distribution

Together, the two distributions  $p(s)$  and  $p(x|s)$  completely specify the generative model. One could combine them into a single, *joint distribution* (Appendix A.?) which expresses the frequency of occurrence of every combination of  $s$  and  $x$ :

$$p(s,x) = p(s)p(x|s).$$

This is true for every generative model: it specifies the joint distribution of all variables in the task, including the observations.

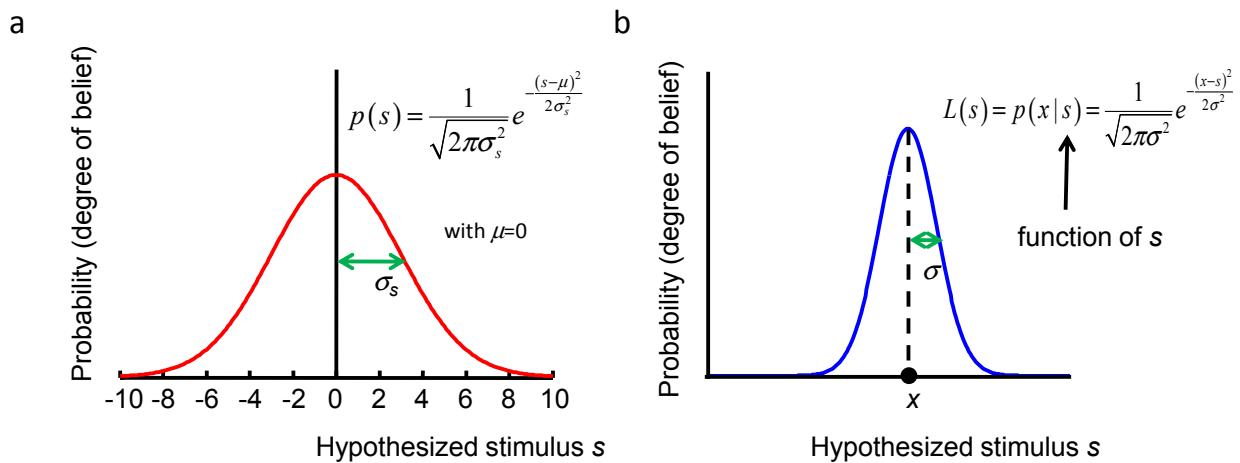
---

<sup>1</sup> We assume throughout this chapter that  $\sigma$  does not depend on  $s$ . In many real tasks, this assumption is not justified. For example, the internal measurement of sound location will be noisier for locations farther out from straight ahead.

## 2.4 Step 2: The inference process

Organisms do not have direct knowledge of world states. The observer's brain has to infer the value of a world state of interest based on the observations. Performing inference in a Bayesian/probabilistic way consists of computing the posterior distribution, and using that to produce an estimate of the world state of interest. Bayesian inference is sometimes referred to as "inverting" the generative model or as an "inverse" problem, because the observer starts with the final product of the generative model (the observations) and reasons back to the cause (the world state of interest).

Two functions play a role in the posterior distribution: the prior distribution and the likelihood function.



**Figure 2.4.** Consider a single trial on which the measurement is  $x$ . The observer is trying to infer which stimulus  $s$  produced this measurement. The two functions that play a role in the observer's inference process (on a single trial) are the prior and the likelihood. The argument of both the prior and the likelihood function is  $s$ , the hypothesized stimulus. (a) Prior distribution. This distribution reflects the observer's beliefs about different possible values the stimulus can take. (b) The likelihood function over the stimulus based on the measurement  $x$ . The likelihood function is centered at  $x$ .

### 2.4.1 The prior distribution

In Section 2.3.1, we introduced the stimulus distribution  $p(s)$ , which reflects how often each stimulus value (auditory location) occurs in the experiment. Suppose that the observer has learned this distribution through extensive training on this experiment. Then, the observer will already have an *expectation* about the stimulus before it even appears, namely that  $s=\mu$  will be most probable, and that the probability falls off with the magnitude of  $s$  according to the learned Gaussian curve. The expectation that the observer holds about the stimulus without having received any evidence on the given trial constitutes prior knowledge, and therefore, in the

inference process,  $p(s)$  is referred to as the *prior distribution* (Fig. 2.4a). The prior distribution is mathematically identical to the world state distribution  $p(s)$ , but unlike it, the prior distribution exists on *an individual trial*: it reflects the observer's beliefs on that trial. It is therefore an example of a subjective distribution. The argument of the prior distribution is the *hypothesized world state (hypothesized stimulus)* rather than the true world state (true stimulus), and probability is interpreted as *degree of belief* rather than frequency of occurrence. We will often use the terms "hypothesized world state" or "hypothesis", and "degree of belief" when referring to the functions used in the inference process.

### Box 2.3. Objective and subjective probabilities

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

#### 2.4.2 The likelihood function

The likelihood function represents the observer's belief about a variable given the measurements only – absent any prior knowledge. When the measurement distribution is known, so is the likelihood function. In our current example, we know the measurement distribution of  $x$  given  $s$ . This means that we know the likelihood function over  $s$ , which we denote  $L(s;x)$ , or  $L(s)$  for short when it is clear that the likelihood is based on  $x$ :

$$L(s;x) = p(x|s).$$

At first sight, this definition seems strange: why would we define  $p(x|s)$  under a new name? The key point lies in the fact that the likelihood function is a function of  $s$ , not of  $x$ . The notation " $;x$ " in fact indicates that  $x$  is a parameter of the function, and on a given trial, it is simply a fixed number, not a variable.

### Box 2.4: The likelihood of what?

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

### Box 2.5: A daily-life example of a likelihood

Typically, different provinces in a country have different proportions of farmers: some provinces are more rural, others more urban. What is the likelihood that a randomly chosen farmer lives in a particular province? This likelihood is numerically equal to the probability that the person is a farmer given that they live in that province, which is equal to the proportion of farmers in that province. To relate back to the main text, the measurement is analogous to the statement “This person is a farmer” whereas the world state we are interested in is “the province where this person lives”. Note that the total likelihood summed over all provinces is not equal to 1. For example, if the country has 12 provinces and each province has 10% farmers, the likelihoods would sum up to 1.2. Because likelihoods do not sum to 1, the likelihood function is not a probability distribution.

The interpretation of the likelihood function is in terms of *hypotheses*. When an observer is faced with a particular measurement, what is the probability of that measurement when the world state (here  $s$ ) takes a certain value? Each possible value of the world state is a hypothesis, and the likelihood of that hypothesis is the observer’s belief that the measurement would arise under that hypothesis. Thus, a fundamental difference between the measurement distribution and the likelihood function is that the former is an objective probability distribution, while the latter represents the subjective beliefs of an observer. This is analogous to the distinction between the world state distribution and the prior distribution.

As illustrated in the farmer example (Box 2.3), the likelihood function is not necessarily normalized (i.e., does not generally integrate to one). The reason is that it is a function of the variable after the “given” sign, not of the one before it. This is why the likelihood function is called a *function* and not a distribution (a distribution is always normalized).

For the measurement distribution given by Eq. (2.3), the likelihood function over the stimulus is

$$L(s; x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(s-x)^2}{2\sigma^2}} \quad (2.4)$$

This likelihood function, shown in Fig. 2.4b, happens to be normalized (over  $s$ ), since in a Gaussian distribution, argument and mean can be interchanged without changing the distribution. However, in Chapter 5, we will encounter likelihood functions over other world states than  $s$ , which turn out to be not normalized. Moreover, when we discuss a neural measurement distribution in Chapter 11, we will see an example of a non-normalized likelihood function over the stimulus.<sup>2</sup>

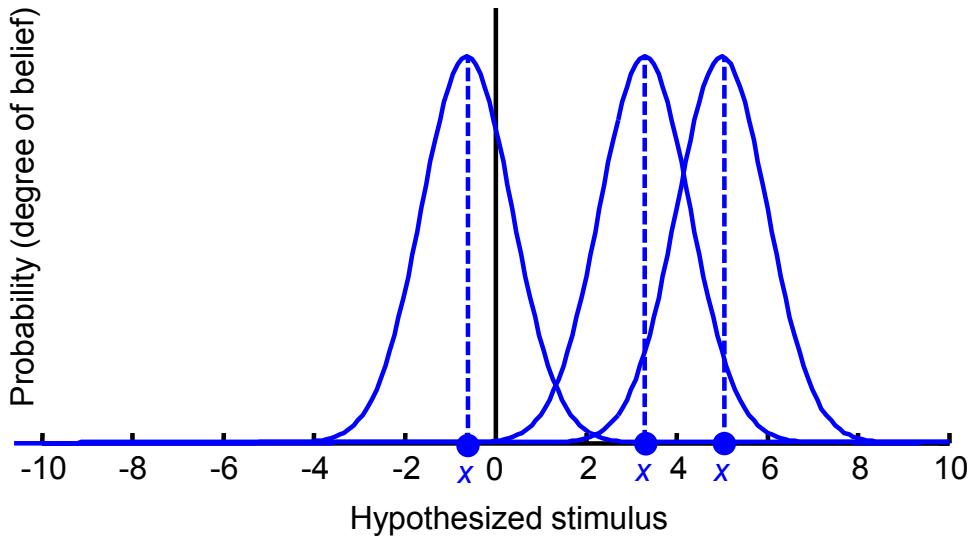
<sup>2</sup> Because  $L(s; x)$  happens to be normalized, we can talk about the variance of this likelihood function, which is  $\sigma^2$ . If  $L(s; x)$  had not been normalized, it would strictly speaking not have been appropriate to talk about variance,

Using the likelihood function in Eq. (2.4), one could make a best guess of the value of the stimulus in the world. This is called the *maximum-likelihood estimate* of  $s$ , and we denoted it  $\hat{s}$ ; the hat is common notation for an estimate. Formally, we can write the definition of the maximum-likelihood estimate as

$$\hat{s}_{\text{ML}} = \underset{s}{\text{argmax}} L(s; x). \quad (2.5)$$

“Argmax” stands for “the argument of the maximum”: the value of the variable written below it for which the function following it takes its largest value. In this case, the maximum-likelihood estimate is simply equal to the measurement,  $\hat{s}_{\text{ML}} = x$ .

In the context of our auditory localization task, the likelihood function in Fig. 2.4b reflects the observer’s belief that the measurements would arise from each hypothesized sound location. The peak at  $x$  indicates the location of a sound source that would with highest probability produce the measurement, the maximum-likelihood estimate (MLE). Moreover, the width of the likelihood function is interpreted as the observer’s level of uncertainty. A narrow likelihood means that the observer is very certain, a wide likelihood that the observer is very uncertain. Although it follows from Eq. (2.3) that the width of the likelihood function is identical to the width of the measurement distribution, these widths have different interpretations. The latter quantifies the spread of the measurements, the former the level of uncertainty based on a single measurement.



**Figure 2.4.** Likelihood functions on three example trials on which the true stimulus was identical. The likelihood function is not fixed: it “moves around” from trial to trial because the measurement  $x$  does.

---

but it is still common to do so. The precise but cumbersome verbiage would be “variance of the normalized likelihood function”.

Importantly, the likelihood function changes from trial to trial, since it depends on the measurement  $x$ , which itself is randomly drawn on each trial. This is shown in Fig. 2.4. The blue dots represent several random draws  $x$ , each on a separate trial, from a Gaussian distribution centered at the true stimulus (not shown). For each draw  $x$ , the likelihood function  $L(s;x)$  is shown in blue. It is a function of the *hypothesized* stimulus value  $s$ . From trial to trial, the likelihood function “wiggles around”.

### 2.4.3 The posterior distribution

An optimal Bayesian observer computes a posterior distribution over a world state from measurements, using knowledge of the generative model. In the example central to this chapter, the relevant posterior distribution is  $p(s|x)$ , the probability density function over the stimulus variable  $s$  given a measurement  $x$ . Bayes’ rule states

$$p(s|x) = \frac{p(x|s)p(s)}{p(x)}.$$

It is also commonly written as

$$p(s|x) \propto p(x|s)p(s). \quad (2.6)$$

where we use the proportionality sign and remove the factor  $1/p(x)$ , since this factor simply acts as a normalization constant. If we do not know the normalization factor, we still know the full shape of the posterior probability distribution, including which value of  $s$  has the highest probability. We refer to the Box for a more detailed explanation.

#### Box 2.6: Why that proportionality sign?

It is very common to see the form of Bayes’ rule in Eq. (2.6), with a proportionality sign. How is this particular form justified? It is because the denominator of Bayes’ rule, here  $p(x)$ , does not depend on the argument of the posterior, here the world state of interest. Thus, it simply acts as a multiplicative constant. A multiplicative constant does not change the shape of a function or where that function is maximal. Of course, the multiplicative constant is not an arbitrary number. It has to be such that the total integrated probability equals 1. For this reason,  $1/p(x)$  is also called a *normalization constant*. One can write  $p(x)$  as the sum or integral of the numerator over all possible values of the world state:

$$p(x) = \int p(x|s)p(s)ds.$$

This is in analogy to Chapter 1 (Box: Derivation of Bayes' formula). However, there, the world state variable was discrete and therefore the normalization consisted of a sum rather than an integral.

The common way of dealing with the normalization constant is to first calculate the numerator,  $p(x|s)p(s)$ , and then normalize at the end if desired. There is nothing wrong with explicitly writing  $1/p(x)$ . However, this factor would just stand there until the end of the computation, unless you choose to write it in the integral form during the computation and evaluate the inside of the integral along with evaluating the numerator. This would be cumbersome, however, since you would have to write down the same expression twice, once in the numerator, and once inside the integral in the denominator. The effect of working with the proportionality sign is that you first evaluate the entire numerator, and then in the end evaluate the denominator by plugging the final expression of the numerator into the integral. Sometimes that final evaluation of the integral is not even needed, because the integral has a standard known form (e.g. Gaussian) or because one is only interested in where the function peaks (the maximum-a-posteriori estimate).

In Chapter 4, we will start encountering ratios of two posterior probabilities,  $p(s_1|x)/p(s_2|x)$ . For those, the normalization is common to numerator and denominator and therefore cancels out.

All we have done so far is apply a general rule of probability calculus. What is important, however, is the interpretation of Eq. (2.6). As we discussed before for prior distribution and likelihood function, the value  $s$  in this equation is interpreted as a *hypothesized value* of the random variable representing the stimulus. Eq. (2.6) expresses that the observer considers all possible values of  $s$ , and asks to what extent each of them is supported by the measurement  $x$ , and prior beliefs. The answer to that question is the posterior distribution,  $p(s|x)$ . The conditional distribution  $p(x|s)$  is now not the measurement distribution, which is a probability distribution over  $x$ , but the observer's likelihood function, which is a function of  $s$ . Since the optimal Bayesian observer uses the exact, correct distributions of the generative model, the likelihood function is given by  $L(s;x)$  from Eq. (2.4) and the prior by  $p(s)$  from Eq. (2.1). For the optimal observer, we can therefore rewrite Eq. (2.6):

$$p(s|x) \propto L(s;x) p(s). \quad (2.7)$$

### Box 2.7: A daily-life example of a posterior

We continue the farmer example from Section 2.4.2. Suppose we don't only know the proportion of farmers in each province, but also the population of each province. What is then the posterior probability that the randomly chosen farmer lives in province X? We would simply calculate the number of farmers in each province by multiplying the proportion of

farmers (the likelihood) with the province's population. This number for province X divided by the total number of farmers in the country (obtained by summing over all provinces) gives the posterior probability that our farmer lives in province X.

This example is Bayes' rule but with probabilities represented by counts: we used the *number* of people in a province as the prior and the total *number* of farmers as the normalization. In a strict application of Bayes' rule, the former would be replaced by the *proportion* of the country's population in the given province, and the latter by the *proportion* of farmers in the country. The outcome would be the same, since numerator and denominator would be divided by the same number (the country's population).

We will now compute the posterior under the assumptions we made in the previous section about the stimulus distribution and the measurement distribution. Upon substituting the expressions for  $L(s;x)$  and  $p(s)$  into Eq. (2.7), we see that in order to compute the posterior, we need to compute the product of two Gaussian functions. (We use "functions" instead of distributions, since in general the functions do not need to be normalized; for example, as already mentioned,  $L(s;x)$  does not need to be normalized over  $s$ .) Multiplying two Gaussian functions over the same variable is a common occurrence in Bayesian models of perception. The result is, after normalization, a new Gaussian distribution (see Box 2.8).

### Box 2.8: Multiplying two Gaussian functions

Here, we discuss the product of two Gaussian functions over the same random variable  $Y$ . One has mean  $\mu_1$  and variance  $\sigma_1^2$ , and the other mean  $\mu_2$  and variance  $\sigma_2^2$ :

$$p_1(y) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(y-\mu_1)^2}{2\sigma_1^2}}, \quad p_2(y) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}.$$

We multiply these distributions just as we would multiply regular functions, and normalize the result (since the product is not automatically normalized). The resulting probability

distribution is another normal distribution, now with mean  $\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}$  and variance  $\frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}$ .

Since it is a normal distribution, it should get the standard normalization constant of the normal distribution, namely 1 divided by the square root of  $2\pi$  times the variance. These results are derived in Problem 3 at the end of this Chapter.

Applied to our problem, we find from Eq. (2.7) that the posterior is a new Gaussian distribution

$$p(s|x) = \frac{1}{\sqrt{2\pi\sigma_{\text{posterior}}^2}} e^{-\frac{(s-\mu_{\text{posterior}})^2}{2\sigma_{\text{posterior}}^2}}, \quad (2.8)$$

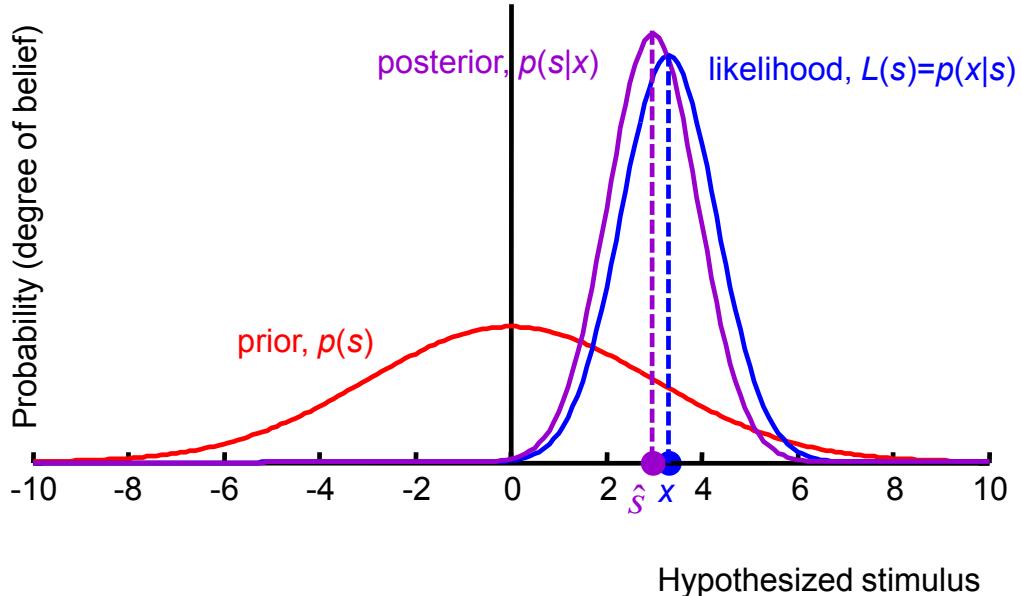
where the mean of the posterior is

$$\mu_{\text{posterior}} = \frac{\frac{x}{\sigma^2} + \frac{\mu}{\sigma_s^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}} \quad (2.9)$$

and its variance is

$$\sigma_{\text{posterior}}^2 = \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}}. \quad (2.10)$$

The posterior is drawn in Fig. 2.5.



**Figure 2.5.** The posterior distribution is obtained by multiplying the prior distribution with the likelihood function and normalizing the resulting function. The hypothesized stimulus value with the highest posterior probability is the observer's maximum-a-posteriori estimate of the stimulus. How would the posterior change if the likelihood were wider?

The interpretation of this posterior (Eqs. (2.8)-(2.10)) is important. The mean of the posterior, Eq. (2.9), is of the form  $ax+b\mu$ , in other words, a linear combination of  $x$  and the mean of the

prior,  $\mu$ . The coefficients  $a$  and  $b$  in this linear combination are  $\frac{1}{\sigma^2}$  and  $\frac{1}{\sigma_s^2}$ ,  
 $\frac{1}{\sigma^2 + \sigma_p^2}$  and  $\frac{1}{\sigma^2 + \sigma_s^2}$ ,

respectively. These sum to 1, and therefore the linear combination is a *weighted average*, where the coefficients act as weights. This weighted average,  $\mu_{\text{posterior}}$ , will always lie somewhere in-between  $x$  and  $\mu$ .

### Box 2.9: Weighted averages

A daily-life example of a weighted average appears when determining a student's overall grade in a class. Suppose the student takes a midterm and a final exam and gets grades  $\mu$  (for midterm) and  $x$  (for exam). However, the midterm is less important than the final. Therefore, the teacher weights the final exam grade by a factor  $a = 0.7$  and the midterm grade by a factor  $b = 0.3$ . Then the student's overall grade in the class is the weighted average  $ax+b\mu$ . It will lie in between  $\mu$  and  $x$ .

Exercise 2.2: Show that when  $z = ax + by$ , with  $x < y$ ,  $a$  and  $b$  both non-negative, and  $a+b=1$ , then  $x < z < y$ .

Where exactly  $\mu_{\text{posterior}}$  lies is determined by the weights. The weights are normalized versions of the inverse variances of the likelihood function and the prior distribution. If the variance of the likelihood is lower than that of the prior distribution, the inverse variance of the likelihood (i.e., the reliability of the measurement) is higher than that of the prior. As a consequence, the weight to  $x$  is higher than to  $\mu$ , causing the mean of the posterior to lie closer to  $x$  than to  $\mu$ . Of course, the reverse also holds: if the variance of the likelihood is larger than that of the prior, the mean of the posterior will lie closer to the mean of the prior than to the measurement.

As you can see in Eq. (2.7), the prior acts in exactly the same way as the likelihood. Effectively, to Bayes' rule priors and likelihoods are just two pieces of information that need to be combined. Each piece of information has an influence that corresponds to the quality of the information.

Exercise 2.3: In the special case that  $\sigma = \sigma_s$ , compute the mean of the posterior.

The intuition behind the weighted average in Eq. (2.9) is that the prior “pulls the posterior away” from the measurement and towards its own mean, but its ability to pull depends on how narrow it is compared to the likelihood function. If the likelihood function is narrow – which happens

when the noise level is low – then the posterior won't budge much: it will be centered close to the mean of the likelihood function. This intuition is still valid if the likelihood function and the prior are not Gaussian but are roughly bell-shaped.

So far, we have considered the mean of the posterior. The variance of the posterior is given by Eq. (2.10). It is interpreted as the overall level of uncertainty the observer has about the stimulus after combining the measurement with the prior. It is different from both the variance of the likelihood function and the variance of the prior distribution.

Exercise 2.4:

- a) Show that the variance of the posterior can also be written as  $\sigma_{\text{posterior}}^2 = \frac{\sigma^2 \sigma_s^2}{\sigma^2 + \sigma_s^2}$ .
- b) Show that the variance of the posterior is smaller than both the variance of the likelihood function and the variance of the prior distribution.

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

Exercise 2.5: What is the variance of the posterior in the special case that  $\sigma = \sigma_s$ ?

Exercise 2.6: What are the mean and the variance of the posterior when  $\frac{\sigma}{\sigma_s}$  is very large or very small? Interpret.

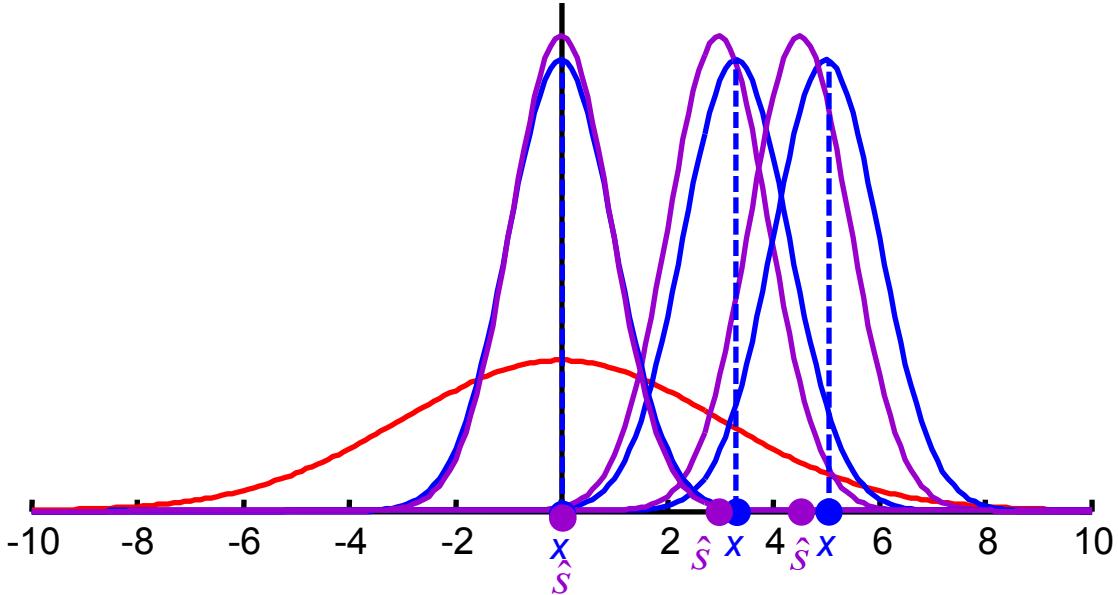
An intuitive way to understand Eq. (2.10) is by rewriting it as

$$\frac{1}{\sigma_{\text{posterior}}^2} = \frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}.$$

Inverse variance of a subjective probability distribution is interpreted as level of certainty, so this equality expresses that the posterior certainty is the sum of the certainties from the likelihood function and the prior. This way of writing also makes it obvious that combining a measurement with a prior can only increase certainty.

Just like the likelihood function, discussed in Section 2.4.2 (and Fig X), the posterior distribution “wiggles around” from trial to trial because it depends on  $x$ , which varies across trials. This is shown in Fig. 2.6. The blue dots represent several random draws of  $x$ , each on a separate trial. For each draw  $x$ , the posterior distribution  $p(s|x)$  is shown. Mean and variance of

the posterior distribution are different from those of the likelihood function, as we have seen above.



**Figure 2.6.** Likelihood functions (blue) and corresponding posterior distributions (purple) on three example trials on which the true stimulus was identical. The prior distribution is shown in red. The likelihood function, the posterior distribution, and the maximum-a-posteriori estimate are not fixed objects: they move around from trial to trial because the measurement  $x$  does.

#### 2.4.4 The maximum-a-posteriori (MAP) estimate

Now that we have computed the posterior distribution on a given trial, Eqs. (2.8)-(2.10), the next step is to use this distribution to obtain an estimate of the world state of interest, here the stimulus  $s$ . The most common way of obtaining this estimate, or read-out, is to choose the value of the state-of-the-world variable that has the highest probability under the posterior distribution. This is called *maximum-a-posteriori* (MAP) estimation. In our case, the MAP estimate would be

$$\hat{s}_{\text{MAP}} = \underset{s}{\text{argmax}} p(s | x). \quad (2.11)$$

If the function to be maximized is a probability distribution, as here, then the argmax is also called the *mode* of the distribution.

MAP estimation is an intuitive way of reading out a posterior, since this method picks the most probable option. MAP estimation is also the read-out method that maximizes expected reward if reward is obtained only when the estimate is exactly correct, that is, when  $\hat{s}$  is equal to the true stimulus on a given trial. We will discuss the connection between MAP estimation and

reward in Chapter 6?. We will then see that under other plausible reward scenarios, expected reward is maximized by estimates different from the MAP estimate.

In our example of combining a measurement with a Gaussian prior, the MAP estimate is the mode of the posterior distribution in Eq. (2.8). Since this distribution is Gaussian, the mode is equal to the mean, which is given by Eq. (2.9). Thus, we have

$$\hat{s}_{\text{MAP}} = \frac{\frac{x}{\sigma^2} + \frac{\mu}{\sigma_s^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}}. \quad (2.12)$$

The remarks we made above about the mean of the posterior therefore also apply to the MAP estimate. In particular, the MAP estimate is a weighted average of the measurement  $x$  and the prior mean  $\mu$ , weighted by the inverse variances of likelihood function and prior, respectively. Recall that the observer's task is to report an estimate of the stimulus on a continuum. According to the model of the task adopted here, the MAP estimate is what the observer should perceive. Thus, in this Bayesian model, the observer's *percept* is the MAP estimate. We could say that the observer sees, hears, etc. the MAP estimate of the stimulus.

## 2.5 Step 3: The distribution of the MAP estimate

The MAP estimate of the world state of interest, Eq. (2.11) in general and Eq. (2.12) in our example, is the observer's best guess of the world state when given a noisy measurement  $x$ . This noisy measurement is a reflection of the internal state of the observer on a given trial, and not directly accessible to the experimenter. The experimenter only knows – or more precisely, assumes – the distribution of  $x$  when the true stimulus is  $s$ : it is given by the measurement distribution, Eq. (2.3). Since  $x$  is a random variable, so is the MAP estimate (Fig. 2.6). Hence, in response to repeated presentations of the same stimulus, the MAP estimate will be a random variable with a probability distribution.

### Take-away: Variability of the MAP estimate

The MAP estimate will vary from trial to trial (i.e. will be stochastic) even when the true stimulus is held fixed, because the measurement varies from trial to trial.

Nowhere in our model have we added extra noise beyond the Gaussian noise that we started out with, Eq. (2.3). The mapping from measurement to MAP estimate, Eq. (2.12), is completely deterministic. The stochasticity in the MAP estimate is *inherited from* the stochasticity in the measurement  $x$ .

The experimenter can only control the stimulus. To compare our Bayesian model with an observer's behavior in a psychophysical task, we need to specify what the Bayesian model predicts for the observer's responses when the true stimulus is  $s$ . In other words, we need to

know the distribution of the MAP estimate when the true stimulus is  $s$ . We denote this distribution by  $p(\hat{s}_{\text{MAP}} | s)$ .

From Eq. (2.3), we know that when the true stimulus is  $s$ ,  $x$  follows a Gaussian distribution with mean  $s$  and variance  $\sigma^2$ . From Eq. (2.12), we know that the random variable  $\hat{s}_{\text{MAP}}$  is linearly related to the random variable  $x$ . We now make use of useful properties of random variables with a Gaussian distribution (see Box 2.10).

**Box 2.10: Linear combinations of normally distributed random variables**

Normally distributed variables have some convenient properties that we use frequently:

1. If a random variable  $X$  has a normal distribution with mean  $\mu$  and variance  $\sigma^2$ , and  $a$  and  $b$  are constants, then the random variable  $aX+b$  has a normal distribution with mean  $a\mu+b$  and variance  $a^2\sigma^2$ .
2. If random variables  $X$  and  $Y$  are independent and have normal distributions with means  $\mu_X$  and  $\mu_Y$ , and variances  $\sigma_X^2$  and  $\sigma_Y^2$ , then the random variable  $X+Y$  has a normal distribution with mean  $\mu_X+\mu_Y$ , and variance  $\sigma_X^2+\sigma_Y^2$ .

These properties can be proven using the rules for functions of random variables; see Appendix A, Section 11.3.

Exercise 2.7: Combine both properties to show that the random variable  $aX+bY$  has mean  $a\mu_X+b\mu_Y$ , and variance  $a^2\sigma_X^2+b^2\sigma_Y^2$ .

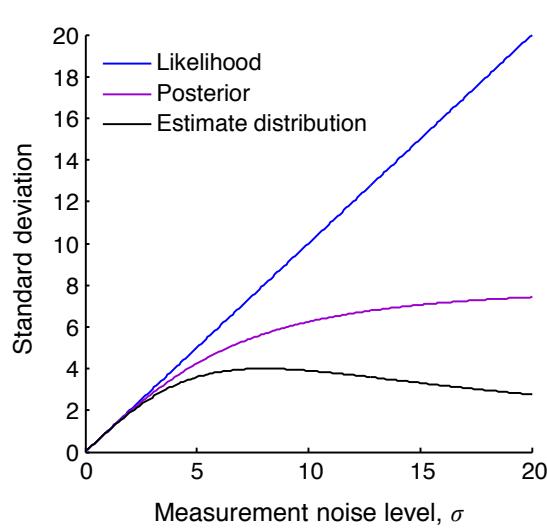
Exercise 2.8: Using the properties in the Box, show that when the true stimulus is  $s$ , the

MAP estimate follows a Gaussian distribution with mean  $\frac{\frac{s}{\sigma^2} + \frac{\mu}{\sigma_s^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}}$  and variance  $\frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}}$

$$\frac{\frac{1}{\sigma^2}}{\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}\right)^2}:$$

$$p(\hat{s}_{\text{MAP}} | s) = \text{Normal}\left(\hat{s}_{\text{MAP}}, \frac{\frac{s}{\sigma^2} + \frac{\mu}{\sigma_s^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}}, \frac{1}{\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}\right)^2}\right). \quad (2.13)$$

Note that the variance of the MAP estimate,  $\sigma_{\text{MAP}}^2 = \frac{1}{\sigma^2} \left( \frac{1}{\sigma^2} + \frac{1}{\sigma_s^2} \right)^{-1}$ , is different from the variance of the posterior (from Eq. (2.10)),  $\sigma_{\text{posterior}}^2 = \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}}$ . This might seem like a subtle difference, but the difference becomes clear if we plot it:



**Figure 2.7.** The standard deviation of the likelihood function, the posterior, and the estimate distribution as a function of the measurement noise level, when the stimulus distribution has a standard deviation of  $\sigma_s=8$ . As  $\sigma$  grows very large, the standard deviation of the posterior will converge to the standard deviation of the prior,  $\sigma_s$ , while the standard deviation of the estimate distribution will go to 0.

As the sensory noise level increases, the posterior gets wider and wider. Initially, the same holds true for the distribution of the MAP estimate. However, when  $\sigma$  grows large enough, the standard deviation of the MAP estimate will decrease again.

Exercise: Why does this make intuitive sense?

The variance of the MAP estimate distribution is different from any variance we have encountered before: it is different from the variance of the measurement distribution, from the variance of the likelihood function, and from the variance of the posterior distribution. This might help to distinguish the various probabilistic concepts (see Fig. 2.10).

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

	Mean	Variance
STEP 1 Generative model	Stimulus distribution $p(s)$ $\mu$	$\sigma_s^2$
	Measurement distribution $p(x s)$ $s$	$\sigma^2$
STEP 2 Inference	Prior distribution $p(s)$ $\mu$	$\sigma_s^2$
	Likelihood function $L(s;x) = p(x s)$ $x$	$\sigma^2$
STEP 3 Estimate distribution	Posterior distribution $p(s x)$ $\frac{\frac{x}{\sigma^2} + \frac{\mu}{\sigma_s^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}} = \hat{s}$	$\frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}}$
	MAP estimate distribution $p(\hat{s} s)$ $\langle \hat{s} \rangle = \frac{\frac{s}{\sigma^2} + \frac{\mu}{\sigma_s^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}}$	$\frac{1}{\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}\right)^2}$

**Figure 2.8:** Means and variances of all distributions discussed in this chapter.

## 2.6 Sanity checks

With any model or theory, it is wise to check if the equations one has derived make intuitive sense. After all, something could have gone wrong in the derivation. One type of sanity check is to substitute special values for the variables in the equation. Which special values are useful

depends on the problem, but 0 and infinity are often useful. In our problem at hand, suppose we want to perform a sanity check on Eq. (2.12) for the MAP estimate. Our intuition tells us that if the prior distribution is completely flat, it does not express any particular beliefs and therefore should not have any effect on the observer's estimate. The prior distribution is given by Eq. (2.1). Although a Gaussian distribution like this can never be completely flat, one can approximate this situation by choosing it to be extremely wide, in other words, by taking the limit of  $\sigma_s$  going to infinity. Substituting infinity for  $\sigma_s$  in Eq. (2.12) gives  $\hat{s}_{\text{MAP}} = x$ , that is, the MAP estimate reduces to the measurement. This is intuitive: in the absence of any other information, the most probable value of the stimulus is equal to the measurement.

Exercise 2.12: Following a similar logic, perform a sanity check on Eq. (2.12) in the following cases: a) sensory noise,  $\sigma$ , is very large; b) sensory noise is very small; c) the measurement  $x$  happens to be equal to the mean of the prior.

Exercise 2.13: Perform all sanity checks you can think of on the mean and variance of the MAP estimate (see Eq. (2.13)).

## 2.7 Concluding remarks

We described how Bayesian modeling consists of three steps: defining the generative model, deriving an expression for the observer's MAP estimate, and deriving the distribution of the MAP estimate over many trials. If the generative model is known to and used by the observer, probability calculus specifies the inference process exactly. Although many of the functions and distributions encountered look similar (in this Chapter, they are all Gaussian), they must be distinguished carefully. All relevant quantities are shown in Fig. 2.8. We will dwell more on their interpretations in the next chapter.

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

The essence of Bayesian observers is that they consider all possible values of the state-of-the world variable, and computes their respective probabilities. In other words, the Bayesian observer does not commit to a limited set of hypotheses (or to a single hypothesis) unless so directed by the evidence.

## 2.8 Problems

**Problem 2.1.** Match the following functions that play a role in Bayesian modeling with the descriptions:

### FUNCTIONS

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

### DESCRIPTIONS

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

**Problem 2.2.** Imagine you have collected data about reported sightings of the dodo throughout history. We will call these data  $F$ . Suppose you are interested in the time the dodo went extinct, denoted  $Q$ . Then the likelihood function of interest to you is

- a)  $p(Q|F)$  as a function of  $F$
- b)  $p(Q|F)$  as a function of  $Q$
- c)  $p(F|Q)$  as a function of  $F$
- d)  $p(F|Q)$  as a function of  $Q$

Incidentally, a paper has calculated this likelihood: Roberts DL, Solow AR (2003), *Flightless birds: when did the dodo become extinct?* Nature, 426 (6964): 245.



Dodo (extinct)

**Problem 2.3.** Let  $s$  be the stimulus of interest,  $x$  the measurement of  $s$ ,  $p(x|s)$  the measurement distribution, and  $p(s)$  the prior distribution of the stimulus.

- a) Write down the posterior distribution over  $s$ , given a measurement  $x$ .

- b) Which of the terms in your expression is called the likelihood function?
- c) What is the difference between the likelihood function and the measurement distribution?

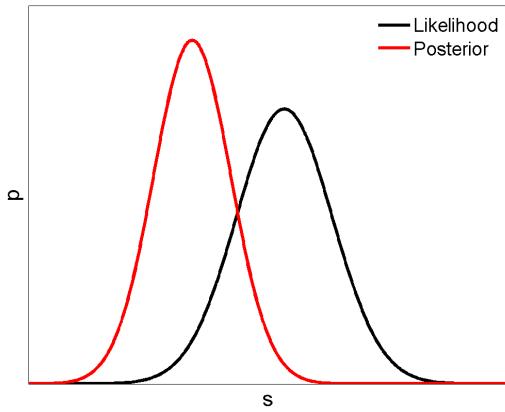
**Problem 2.4.** Many Bayesian inference problems involve a product of two or more Gaussians. A convenient property of Gaussians is that their product is also Gaussian. In this problem, we will lead you through an example to derive this property yourself. Consider an observer who infers a stimulus  $s$  from a measurement  $x$ . Suppose that the measurement distribution  $p(x|s)$  is a Gaussian distribution with standard deviation  $\sigma$  and the prior distribution is a Gaussian with mean  $\mu$  and standard deviation  $\sigma_s$ . If you get stuck, consult Section 7.3 of the Appendix.

- a) Write down the equations for  $p(x|s)$  and  $p(s)$ .
- b) Use Bayes' rule to write down the equation for the posterior,  $p(s|x)$ . Substitute  $p(x|s)$  and  $p(s)$ , but do not simplify.

The numerator is a product of two Gaussians. The denominator,  $p(x)$ , is a normalization factor that ensures that the integral equals 1. For now, we will ignore it and focus on the numerator.

- c) Apply the rule  $e^A e^B = e^{A+B}$  to simplify the numerator.
- d) Expand the two quadratic terms in the exponent.
- e) Rewrite the exponent to the form  $as^2 + bs + c$ .
- f) Show that any quadratic function of the form  $as^2 + bs + c$  can be written as  $a\left(s + \frac{b}{2a}\right)^2 + c - \frac{b^2}{4a}$ . This operation is known as ‘completing the square’.
- g) Rewrite your expression obtained in (e) by completing the square.
- h) Apply the rule  $e^A e^B = e^{A+B}$  to rewrite this into the form  $e^Z e^{-\frac{(s-\mu_{\text{combined}})^2}{2\sigma_{\text{combined}}^2}}$ . Express  $\mu_{\text{combined}}$  and  $\sigma_{\text{combined}}$  in terms of  $x$ ,  $\sigma$ ,  $\mu$ , and  $\sigma_s$ .
- i) Why is  $\mu_{\text{combined}}$  the same as the maximum-a-posteriori (MAP) estimate of the stimulus (i.e., the  $s$  that maximizes the posterior distribution,  $p(s|x)$ )?
- j) Recall that  $p(s|x)$  is a distribution and that its integral should therefore be equal to 1. However, the expression that you obtained in (e) is not properly normalized because we ignored  $p(x)$ . Modify the expression such that it is properly normalized, without using  $p(x)$  (Hint: does  $e^Z$  depend on  $s$ ?)

**Problem 2.5.** The figure below shows a likelihood function and a posterior distribution. Both are Gaussian, with  $\sigma_{\text{posterior}}=1.2$  and  $\sigma_{\text{likelihood}}=1.5$ .



Assume that the prior is also Gaussian. Which of the following statements is true? Explain.

- a) The prior is centered to the left of the likelihood function and is narrower.
- b) The prior is centered to the left of the likelihood function and is broader.
- c) The prior is centered to the right of the likelihood function and is narrower.
- d) The prior is centered to the right of the likelihood function and is broader.

**Problem 2.6.** In the model discussed in this chapter, show that the variance of the MAP estimate is smaller than or equal to the variance of the measurement. What happens when the variance of the prior is much larger than the variance of the measurement? Interpret this result.

**Problem 2.8.** True or false? Explain.

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

**Problem 2.9.** Show mathematically that the maximum of the red curve in Fig. 2. X is at  $\sigma=\sigma_s$ . (Hint: recall from calculus how to find the maximum of a function.)

**Problem 2.10.** An observer infers a stimulus  $s$  from a measurement  $x$ . The measurement distribution  $p(x|s)$  is a Gaussian distribution with mean  $s$  and variance  $\sigma^2$ . The stimulus distribution  $p(s)$  is 0 for  $s < 0$  and an exponential distribution,  $p(s) \propto \exp(-\lambda s)$ , for  $s \geq 0$ .

- a) Compute the MAP estimate.
- b) (\*) Compute the probability density function of the MAP estimate for given  $s$ .

**Problem 2.X\*.**

Calculate the expected squared error of the MAP estimator when the stimulus is  $s$ . This expected squared error is defined as  $\langle (F(x) - s)^2 \rangle_{p(x|s)}$ . Hint: use the bias-variance trade-off from Appendix X.

**LAB PROBLEMS****Problem 2.12.**

- Reproduce Fig. 2.7 (the one with the different standard deviations) using the equations in Table 2.1.
- What happens to the three curves if you change the standard deviation of the prior to 4? Explain why the changes make sense.
- Plot the same three curves as a function of the standard deviation of the prior, for  $\sigma=8$ . Explain their shapes.

**Problem 2.13**

An observer infers a stimulus  $s$  from a measurement  $x$ . Let's say that on a particular trial, the measurement is  $x = 30$ . The measurement distribution  $p(x|s)$  is Gaussian with standard deviation  $\sigma=5$ . Assume a Gaussian stimulus distribution  $p(s)$  with mean 20 and standard deviation 4; this also serves as the prior distribution. We are now going to calculate the posterior pdf using Matlab.

- Define a vector of possible  $s$ -values: 0, 0.2, 0.4, ..., 40.
- Compute the likelihood function and the prior on this vector of values of  $s$ .
- Multiply the likelihood and the prior using the “.\*” command.
- Divide this product by its sum over all  $s$  (normalization step).
- Convert this posterior probability mass function into a probability density function by dividing by the step size you used in your vector of  $s$ -values (e.g., 0.2).
- Plot the likelihood, prior, and posterior in the same plot. Is the posterior wider or narrower than likelihood and prior? Do you expect this based on the equations we discussed?
- Change the standard deviation of the measurement distribution to a very large value. What happens to the posterior? Can you explain this?
- Change the standard deviation of the measurement distribution to a very small value. What happens to the posterior? Can you explain this?

**Problem 2.14**

Repeat Problem 2.12, but instead of using a single value of the measurement  $x$ , start with a fixed value of  $s=10$ . From this value of  $s$ , draw ten values of  $x$  from the measurement distribution. You should observe that, from trial to trial, the likelihood function and posterior probability density

function “jump around”. Observe how the posterior shifts under the influence of the “jumping” likelihood function and stationary prior. Explain.

### Problem 2.15

Generate a distribution of MAP and ML estimates by

1. drawing an  $s$  from the stimulus distribution;
2. drawing a single  $x$  from the measurement distribution, and calculating the posterior distribution.
3. For each of 1,000 repetitions of 1 and 2, plot the MAP estimate (y-axis) against the true stimulus (x-axis). On a separate graph, plot the MLE (i.e., measurement,  $x$ ) against the true stimulus.
4. Repeat 1, 2 and 3 using different values of the noise standard deviation relative to prior standard deviation. When the noise standard deviation is very small, the MAP and MLE plots should look the same. Why? When the noise standard deviation is very large, the MAP plot looks flat, whereas the MLE plot looks very scattered. Why?

## Contents

3	Understanding Bayesian models.....	3-2
3.1	Linking mathematical to psychological quantities.....	3-2
3.1.1	Uncertainty.....	3-3
3.1.2	Confidence .....	3-5
3.1.3	Bias .....	3-7
3.2	Common mistakes in Bayesian modeling .....	3-8
3.2.1	Constancy of the likelihood function .....	3-8
3.2.2	Estimate distribution = likelihood function .....	3-9
3.2.3	Estimate distribution = measurement distribution x prior .....	3-9
3.2.4	When measurement = stimulus, then estimate distribution = posterior.....	3-10
3.2.5	Prior probability = overall probability of responding .....	3-11
3.2.6	Biased estimators cannot be optimal.....	3-12
3.3	The optimality of MAP estimation.....	3-12
3.3.1	Discrete variables.....	3-12
3.3.2	Continuous variables.....	3-14
3.3.3	Expected squared error .....	3-15
3.4	Suboptimal inference.....	3-17
3.4.1	Prior mismatch .....	3-18
3.4.2	The case of a flat prior .....	3-19
3.4.3	Suboptimal estimators.....	3-20
3.5	Generalizing the model .....	3-20
3.5.1	Different stimulus domains.....	3-21
3.5.2	Stimulus-dependent noise .....	3-22
3.5.3	Higher-dimensional stimulus spaces.....	3-23
3.5.4	Response noise.....	3-24
3.6	Experimental tests of the model .....	3-24
3.6.1	Why would the brain be optimal? .....	3-24
3.6.2	Empirical evidence.....	3-25
3.7	Problems.....	3-27

### 3 Understanding Bayesian models

In Chapter 2, we went through the detailed mechanics of a simple optimal Bayesian model, calculating everything there was to calculate and ending up with predictions for an experiment. Here, we will reflect on the calculations we did there. This consists of several components:

- Associating the mathematical constructs with psychological quantities such as uncertainty, confidence, and bias
- Discussing easily made technical mistakes arising from confusing different parts of the calculation. We strongly recommend that before you read/teach this part, you go through the exercises of Chapter 2.
- Discussing the limitations of the model, and possible ways to address those limitations
- Discussing the strength of empirical evidence for this model
- Extending the model to *suboptimal* Bayesian inference

#### 3.1 Linking mathematical to psychological quantities

Our starting point is the summary of Chapter 2:

	Mean	Variance
STEP 1 Generative model	Stimulus distribution $p(s)$	
	$\mu$	$\sigma_s^2$
STEP 2 Inference	Measurement distribution $p(x s)$	
	$s$	$\sigma^2$
STEP 3 Estimate distribution	Prior distribution $p(s)$	
	$\mu$	$\sigma_s^2$
	Likelihood function $L(s;x) = p(x s)$	
	$x$	$\sigma^2$
	Posterior distribution $p(s x)$	
	$\frac{\frac{x}{\sigma^2} + \frac{\mu}{\sigma_s^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}} = \hat{s}$	$\frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}}$
	MAP estimate distribution $p(\hat{s} s)$	
	$\langle \hat{s} \rangle = \frac{\frac{s}{\sigma^2} + \frac{\mu}{\sigma_s^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}}$	$\frac{1}{\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}\right)^2}$

**Table 3.1**

There are several common psychological quantities associated with Steps 2 and 3: uncertainty, confidence, and bias.

### 3.1.1 Uncertainty

Uncertainty can be associated with every belief function in the inference stage. Prior uncertainty reflects the quality of the observer's knowledge about the state of the world before making any observations. Posterior uncertainty reflects the quality of the observer's knowledge about the state of the world after making the observations. Likelihood uncertainty or sensory uncertainty reflects the quality of the observer's knowledge about the state of the world obtained solely from the observations. In all cases, uncertainty is tied to the observer and therefore a *subjective* quantity (see Box 2.3). It would not make sense to talk about uncertainty in the generative model (Step 1).

Mathematically, uncertainty is a number extracted from a probability distribution or likelihood function. There is no unique agreed-upon definition, but the intuitive use of the term seems fully captured if we define uncertainty as the standard deviation of the probability distribution or a likelihood function. For example, a narrow posterior distribution  $p(s|x)$  means low posterior uncertainty, and a wide posterior distribution means high posterior uncertainty.

**Uncertainty (about a state of a world):** Standard deviation of a probability distribution or likelihood function over that state of the world, or a monotonic transformation of that quantity.

The part about “monotonic transformation” refers to the fact that any monotonically increasing function of standard deviation (for example, variance  $\sigma^2$ , or  $\log \sigma$ ) would also be a legitimate definition of uncertainty, but you have to choose your function and then stick with it.

In the treatment of Chapter 2, all distributions are Gaussian and we can directly read off the different types of uncertainty from Table 3.1:

- Prior uncertainty:  $\sigma_s$
- Likelihood uncertainty:  $\sigma$
- Posterior uncertainty:  $\sigma_{\text{posterior}} = \frac{1}{\sqrt{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}}}$ , which can also be written as  $\frac{\sigma\sigma_s}{\sqrt{\sigma^2 + \sigma_s^2}}$

We see that likelihood uncertainty happens to have the same numerical value as sensory noise level, but in more complicated examples, that is not necessarily the case (see Box 3.1)

For Gaussian priors and likelihoods, posterior uncertainty is always smaller than both prior uncertainty and likelihood uncertainty (see Exercise 2.4b). The definition of uncertainty extends to non-Gaussian distributions (see Fig. 3.1). However, for non-Gaussian priors and/or likelihoods, it is not necessarily the case that posterior uncertainty is always smaller than both prior uncertainty and likelihood uncertainty (see Problem 3.1).

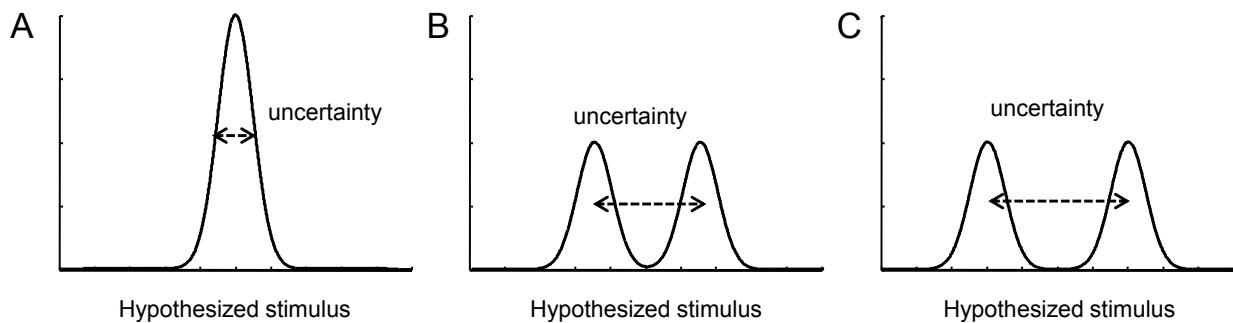
An alternative definition of uncertainty is the *entropy* of a distribution or likelihood function, a measure of disorder. This is explored in Problem 3.2.

### Box 3.1: Terminology: noise, uncertainty, variability

In this book, the term “noise” is reserved for the process by which the observations are generated, i.e. it describes the trial-to-trial variability of observations or measurements. Noise is thus part of the generative model. “Uncertainty”, on the other hand, reflects the observer’s knowledge, or lack of knowledge, about variables in the world. The width of the posterior distribution is a measure of uncertainty, not of noise. Uncertainty is part of the inference process. Noise is one possible cause of uncertainty, but not the only one. When stimuli are ambiguous (e.g. an object’s position that is occluded), the observer has uncertainty without having noise.

Variability is an encompassing term for anything that varies from trial to trial. Noise is

a form of variability; this can be called the variability of the measurement. The MAP estimate is also variable from trial to trial. This can be called “behavioral variability”. Uncertainty is *not* a form of variability.



**Fig. 3.1. Uncertainty.** The curves shown could be likelihood, prior, or posterior.

### 3.1.2 Confidence

In daily life, decisions are made with greater or lesser confidence. You might be confident that you can cross the road before a car reaches you, that it's your friend who is approaching you, or that someone's accent is Italian. Confidence naturally fits into a Bayesian framework and is related to the posterior distribution. Thinking back to Chapter 2, the decision is the estimate that the observer makes of the stimulus. Once the observer makes an estimate of the stimulus, we can define *confidence* (also *decision confidence*, or *judgment confidence*) as the posterior probability density of that estimate:

$$\text{Confidence} = p_{s|x}(\hat{s} | x). \quad (3.1)$$

We use a subscript “ $s|x$ ” here to denote the random variable, as well as the random variable it is conditioned on, because we are *evaluating* the conditional probability distribution at a specific value, namely  $\hat{s}$ . If we had written “confidence =  $p(\hat{s} | x)$ ”, this could have meant the distribution of the estimate for given  $x$  (which would have been a delta function).

**Confidence (about an estimate of a world state):** Posterior probability distribution evaluated at that estimate, or a monotonic transformation of that quantity.

In this definition, confidence can be a priori or a posteriori (since a prior distribution can be thought of as a special case of a posterior distribution), but there is no notion of confidence directly tied to the likelihood function by itself.

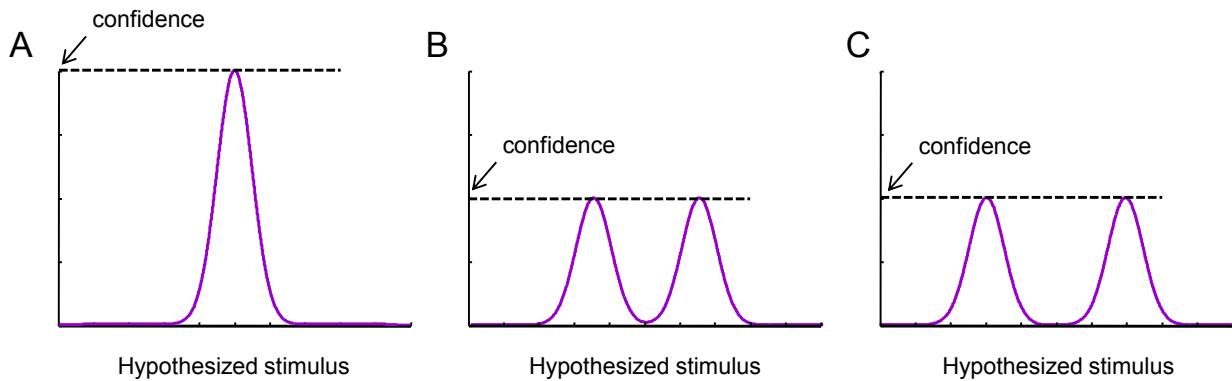
The estimate in the definition of confidence might or might not be the MAP estimate. When the estimate is the MAP estimate, confidence is equal to the maximum value of the posterior distribution (Fig. 3.2):

$$\text{Confidence}_{\text{MAP}} = \max_s p(s|x). \quad (3.2)$$

The confidence associated with the MAP estimate is closely related to uncertainty. The maximum of a Gaussian probability density function is always equal to the factor preceding the exponential (1 divided by the square root of  $2\pi$  times the variance). Thus, for a Gaussian

posterior, confidence from Eq. (3.2) would be  $\frac{1}{\sigma_{\text{posterior}}\sqrt{2\pi}}$ . This is inversely proportional to

uncertainty: when the standard deviation of the posterior is larger, its maximum value will be lower. This is not true for posteriors of any shape. For example, if the posterior is bimodal (=has two peaks), then uncertainty will depend on the separation between the peaks (Fig. 3.1B-C), while estimation confidence will not (Fig. 3.2B-C).



**Fig. 3.2. Confidence.** The curves shown could be prior or posterior.

From Table 3.1, we can read off the posterior confidence associated with the MAP estimate:

$$\text{Confidence}_{\text{MAP}} = \sqrt{\frac{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}}{2\pi}} \quad (3.3)$$

To illustrate the monotonic transformation in the definition of confidence: we could alternatively define MAP confidence as  $2\pi$  times variance, because the operation “square and multiply by  $2\pi$ ” is a monotonically increasing function. Then, MAP confidence would be  $\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}$ . Either way, confidence is higher when the prior is sharper ( $\sigma_s$  smaller).

The study of confidence is sometimes called metacognition (see Box 3.1).

### Box 3.1: Metacognition

A key aspect of Bayesian inference in perception is that observers know and utilize both their uncertainty about measurements and their confidence about their responses. Does this extend to more abstract cognitive levels?

Knowing what you know is called metacognition. Research in education has shown that to become an effective learner, students have to have strong metacognitive skills. Students can acquire such skills by reflecting on what they know and how confident they are about their knowledge. In some places, new forms of testing are designed to incorporate such confidence judgments. In a “Variable Weight” testing format, a test could consist of fifteen questions. Students choose ten about which they are confident (indicated by placing their response on the left side of the answer sheet). The questions for which the students indicate they are confident of their answer and in fact answer correctly are weighted more heavily than questions they answer correctly but for which they indicate less confidence. In another innovation,

“Students are asked a question that requires them to indicate whether they are absolutely sure, fairly sure, or just guessing at their answer. The points for their response are dependent on the correctness of their answer and the confidence they state. For example, if students indicate they are absolutely sure, they earn nine points if they are correct, but no points if they are wrong. However, if they indicate they are unsure or just guessing, they earn three points if they are correct and two points if they are wrong. Students quickly learn that carefully reflecting on their confidence improves their grade.”

Source: Randy M. Isaacson, “Building a Metacognitive Curriculum: An Educational Psychology to Teach Metacognition”, Tomorrow’s Professor Mailing List

### 3.1.3 Bias

There is an apparent contradiction associated with Bayesian MAP estimation. On the one hand, we have asserted that MAP estimation is optimal. On the other hand, the MAP estimate is *biased* in the statistical sense of being on average different from the true stimulus (see Table 3.1). This seems contradictory: wouldn’t it always be better to have an unbiased estimate?

We have seen that the MAP estimate is a weighted average between the measurement and the mean of the Gaussian stimulus distribution (Eq. (2.12)). Therefore, the mean MAP estimate is a weighted average between the true stimulus and the mean of the stimulus distribution (Eq. (2.13)). This has the paradoxical consequence that the mean of the optimal estimate is not equal to the true stimulus: the MAP estimate is *biased* from the stimulus towards the mean of the prior.

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

Formally, the bias of an estimator is defined as the difference between the mean estimate and the true stimulus. It is a function of the true stimulus  $s$ :

$$\text{Bias}(s) = \langle \hat{s} \rangle_{p(\hat{s}|s)} - s . \quad (3.4)$$

The notation  $\langle f(X) \rangle_{p(x)}$  indicates the expected value of the random variable  $f(X)$  under the distribution  $p(x)$  (see Appendix X.X if you need background). The bias is the average difference between the estimate and the true stimulus.

Exercise 3.1:

- a) Verify that the bias of the maximum-likelihood estimate (i.e., the measurement) is zero.
- b) Calculate the bias of the MAP estimate from Table 3.1.

## 3.2 Common mistakes in Bayesian modeling

To build a Bayesian model, we must formulate our generative model and do proper inference on it. Some aspects of this inference are counterintuitive, and scientists doing research on Bayesian models occasionally become confused about the resulting relations between variables. When we personally started building Bayesian models, we got confused about a great number of different issues. It literally took us years to develop a cleaner understanding of the issues and to deal with misconceptions we were holding. In fact, much time during our book-writing retreats was spent refining our own understanding about elementary concepts. It thus seems important to address these issues.

Here we address several misconceptions that frequently arise during Bayesian modeling. In our experience, mistakes about the formalism tend to be variations on a few themes, and we consider it useful to understand why they are mistakes. Although the Gaussian prior and Gaussian likelihood example we went through in Chapter 2 is basic, it allows us to describe different types of mistakes. As you are reading or teaching this section, we strongly recommend that you first go through the problems of Chapter 2, and after that, try to first understand each mistake without reading the explanation.

### 3.2.1 Constancy of the likelihood function

One misconception is that a Bayesian observer's likelihood function is determined by the true stimulus, and therefore it is always the same function as long as the stimulus is held fixed. In reality, as we saw in Chapter 2 (Fig. 2.6) and Chapter 10 (Fig. 10.X), that the likelihood function varies from repetition to repetition because it is based on the measurement, so a single value of the stimulus gives rise to a different likelihood functions whenever the stimulus is given. The misconception arises from confusing the likelihood distribution with the measurement distribution. The measurement distribution *is* indeed constant as it generally is assumed to depend only on the stimulus. Even the measurement distribution can change if there are nuisance parameters. However, the likelihood function always varies from trial to trial.

A related misconception is that the posterior is one given function for a given value of the stimulus. In reality, the posterior also varies from trial to trial even when the stimulus is fixed.

### 3.2.2 Estimate distribution = likelihood function

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

### 3.2.3 Estimate distribution = measurement distribution x prior

The fallacy:

*“To obtain the probability density of an observer’s MAP estimate for a given stimulus, I multiply the measurement distribution  $p(x|s)$  with the prior probability density. This is correct because it will give me a density function that is centered in between the prior mean and the true stimulus.”*

This misconception frequently gets combined with the first one, confusing the measurement distribution with the likelihood function. Then the multiplication statement seems even more correct, even though it is not.

One might think that it is possible to take a shortcut by immediately calculating the estimate distribution instead of going through steps 2 and 3 of the modeling process in sequence. This argument, which is a more sophisticated version of the argument in the previous section, could go as follows: “The ML estimate of the stimulus is equal to the measurement  $x$  [true]. Thus, the distribution of the ML estimate for a given true stimulus  $s$  is equal to the measurement distribution, with  $s$  [true]. Assume that the prior distribution over the estimate  $\hat{s}$  is Gaussian with mean  $\mu$  and variance  $\sigma_s^2$ . Therefore the MAP estimate distribution given a stimulus  $s$  can be obtained by multiplying the ML estimate distribution with the prior [false].” This view suggests that:

$$p_{\hat{s}|s}(\hat{s}|s) \propto p_{x|s}(\hat{s}|s) p_s(\hat{s}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\hat{s}-s)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma_s^2}} e^{-\frac{(\hat{s}-\mu)^2}{2\sigma_s^2}} \quad (3.5)$$

After normalization, this would give a Gaussian distribution with a variance of  $\frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}}$  which

differs from the correct one,  $\frac{1}{\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}\right)^2}$ . An intuitive reason why the answer must be wrong

can be obtained in the zero-reliability limit ( $\sigma \rightarrow \infty$ ). Then, the observer will always estimate the stimulus to be the mean of the prior, and thus the variance of the estimate distribution should be 0. The wrong argument would give a variance of  $\sigma_s^2$ .

The principal mistake made here is that Eq. (3.5) is not a correct application of Bayes' rule. First of all, it does not have the mathematically correct form  $p(y|x) \propto (x|y)p(y)$ . Moreover, in Bayesian modeling of perception, Bayes' rule does not act at the level of estimates over many trials, but at the level of a single trial. Therefore, both components on the right-hand side of Eq. (3.5) are incorrect. The likelihood function should not be a function of the ML estimate or the measurement, but of the hypothesized value of the stimulus  $s$ , again on a single trial. The argument of the prior distribution is not the stimulus estimate but the hypothesized stimulus, again on a single trial.

### 3.2.4 When measurement = stimulus, then estimate distribution = posterior

The fallacy:

*“For the distribution of the observer’s MAP estimate for a given stimulus, I can simply use the ‘average’ posterior, the one obtained when the measurement is equal to the true stimulus  $s$ . This works because it gives me a distribution that is centered in between the prior mean and the true stimulus.”*

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

Either of these mistakes leads to the conclusion that the variance of the estimate distribution is  $\frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}}$ , while in reality this is the variance of the posterior. Equating the

variance of the estimate distribution and the posterior is a frequent mistake. To correctly describe

the relationship of prior, measurement distribution and distributions of MAP estimates, there is no alternative to going through the three steps in Chapter 2.

### 3.2.5 Prior probability = overall probability of responding

- *What is wrong with: “The prior probability of responding rightward is 0.5”?*
- *I have a simple way to derive the prior distribution used by the observer directly from the data. I can simply tally up the observer’s estimates across all trials in the experiment. The higher the prior probability of a stimulus value, the more often the observer will report that stimulus value.*

Over the course of an experiment, an observer’s responses will follow “overall estimate distribution”  $p(\hat{s})$ : certain estimates are more common than others. This differs from the estimate distribution  $p(\hat{s}|s)$  in Table 3.1 because that was the distribution of the stimulus estimate as the true stimulus is fixed at  $s$ . Over the course of an experiment,  $s$  will vary according to the stimulus distribution  $p(s)$ .

It is tempting but incorrect to regard  $p(\hat{s})$  as the observer’s prior distribution. Stated differently, if one world state is twice as common as another world state, it doesn’t necessarily mean I will report it twice as often. Stated yet differently, estimation rate is not equal to base rate. To understand this, consider the task of Chapter 2. Intuitively, if the observer uses the correct prior distribution, than all their responses will be biased towards the mean of the prior and the response distribution will thus be narrower than the prior distribution.

We can also formally calculate the overall distribution of the MAP estimate. The MAP

estimate on a single trial when the measurement is  $x$  is  $\frac{\frac{x}{\sigma^2} + \frac{\mu}{\sigma_s^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}}$ . This is a random variable

because  $x$  is a random variable. To calculate the distribution of the MAP estimate conditioned on the true stimulus  $s$  we used that  $x$  given  $s$  was Gaussian with mean  $s$  and variance  $\sigma^2$ . Similarly, we now consider the distribution of the MAP estimate across all  $s$ . The distribution of  $x$  across all  $s$  is Gaussian with mean  $\mu$  and variance  $\sigma^2 + \sigma_s^2$ .

Exercise: Why? (Hint: Use Box 2.10 on linear combinations of normally distributed variables.

Then, the MAP estimate will have mean  $\mu$  and variance  $\frac{\sigma_s^4}{\sigma^2 + \sigma_s^2}$  (see Problem 3.4). This shows

that the overall estimate distribution is not identical to the prior distribution. The overall estimate

distribution is not frequently used in Bayesian models, since the more informative quantity is the estimate distribution given the true stimulus (or stimuli).

This fallacy also highlights an incorrect use of the term “prior”. A phrase like “ $p(s)$  is the prior probability of responding  $s$ ” is incorrect. A correct way of using the term “prior” is “ $p(s)$  is the prior probability (observer’s belief) that the state of the world is  $s$ ”.

### 3.2.6 Biased estimators cannot be optimal

As we saw in Section 3.1.3, the MAP estimate is biased, which means that on average it is not equal to the true stimulus. On the other hand, MAP estimation is supposed to be the best possible strategy of solving the inference problem. This apparent contradiction highlights the fact that we have so far not carefully defined what it means for an inference strategy to be optimal. Once we do that, it will become obvious that this strategy will produce biases. This is so central to inference that we devote the entire next section to it.

## 3.3 The optimality of MAP estimation

In what sense is MAP estimation optimal? The answer is very simple: when the observer knows and uses the correct generative model MAP estimation is the optimal strategy when the observer’s goal is to maximize *proportion correct* (also called *accuracy*, although some authors use that term for other performance metrics). We will first see, for discrete variables, why that is the case and how this strategy induces biases towards frequently occurring world states. Then we will remark on the complication of dealing with continuous variables.

### 3.3.1 Discrete variables

MAP estimation maximizes the probability of being right on any given trial:

$$\text{Probability correct (for a discrete variable)} = \Pr(\hat{s} = s)$$

This does involve being “biased” towards more frequent values. To see this, we consider an example we already used in chapter 2: Three states  $s$  each have equal populations. The likelihood of a state  $s$  being the home state of a randomly met farmer is equal to the proportion of farmers in that state or more generally  $p(x=\text{occupation}|s)$ . These proportions are given in the table below for three U.S. states, along with the proportions of farmers, proportions of retail workers, and proportions of other occupations. If we know that our subject is from one of the three states, maximum-likelihood estimation would lead you to report that the teacher you just met is from California, the farmer from Kansas, and the retail worker from Wisconsin (highlighted).

	$s=\text{California}$	$s=\text{Wisconsin}$	$s=\text{Kansas}$	%correct ML estimator
Total population	19771 (79.0%)	3446 (13.8%)	1805 (7.2%)	
$x=\text{teacher}$	1.5%	0.9%	1.1%	86%
$x=\text{farmer}$	0.4%	2.6%	3.6%	28%
$x=\text{retail}$	8.4%	9.3%	8.4%	15%
$x=\text{other}$	89.6%	87.3%	87.0%	79%

It is intuitive to see what is wrong with this reasoning. The three states have different total populations (see second row in the table; numbers are in thousands). To maximize the probability of being correct, you do not care about the *proportion* of farmers in each state, but about the *absolute number*, i.e. the proportion multiplied by the total population. These numbers are computed in Table 2.

	$s=\text{California}$	$s=\text{Wisconsin}$	$s=\text{Kansas}$	%correct MAP estimator
$x=\text{teacher}$	304	31	19	86%
$x=\text{farmer}$	81	88	65	38%
$x=\text{retail}$	1669	319	151	78%
$x=\text{other}$	17717	3008	1570	79%

This gives a completely different set of decisions (highlighted). Whereas the best guess of the teacher's home state is still California, the best guess of the farmer's home state is now Wisconsin, and the best guess of the retail worker's homestate is again California. This is because California has a larger population than Wisconsin, which has a larger population than Kansas. Guessing the state that is overall most frequent in combination with a given occupation is MAP estimation, and you see in the right column of the table that for the farmer and the retail worker, the MAP estimator is much more accurate than the ML estimator.

To recognize that what we just did is MAP estimation, it is easiest to express the numbers in Table 2 as proportions of the total population across our three-state country (25022).

	s=California	s=Wisconsin	s=Kansas	%correct MAP estimator
x=teacher	1.2%	0.1%	0.1%	86%
x=farmer	0.3%	0.4%	0.3%	38%
x=retail	6.7%	1.3%	0.6%	78%
x=other	70.8%	12.0%	6.3%	79%

Of course, dividing by a common number does not change the decisions. These percentages could also be obtained by multiplying the percentages for every row with an  $x$  in Table 1 by the proportions of the total populations resident in each state, in the second row of Table 1. In other words,  $p(x,s)$ , the frequency of an occupation-state pair, is obtained by multiplying the state-specific probability of the occupation,  $p(x|s)$ , by the prior probability that a random worker lives in that state,  $p(s)$ :  $p(x,s) = p(x|s) p(s)$ . Since this is, up to a normalization, equal to the posterior distribution  $p(s|x)$ , finding the highest proportion per row in Table 3 is MAP estimation. Although MAP estimation is thus “biased” towards states with higher populations, this is optimal. Note also that the MAP estimator does not *always* go with the state with the highest population. In the case of the farmer, the proportion of farmers in Wisconsin is large enough to overcome the overall population disadvantage. This indicates that the evidence (knowing that someone is a farmer) can be strong enough to overcome the prior (that California has a much higher population), and it does so through the likelihood function.

### 3.3.2 Continuous variables

For continuous variables, the notion of being exactly correct is a bit strange. For example, if you are throwing darts, the probability of hitting any one point is 0, no matter how good you are at darts. As explained in Appendix X, probabilities for continuous variables are only well-defined as probabilities of the variable taking a value in an interval or set of intervals. Nevertheless, it is possible to define correctness as the *probability density* evaluated at the true stimulus. To be specific, if the true stimulus is  $s$ , and the estimate distribution is the density  $p_{\hat{s}|s}(\hat{s}|s)$ , then the probability of correctness is this density evaluated at  $\hat{s} = s$  :

Probability correct (for a continuous variable) when the stimulus is  $s = p_{\hat{s}|s}(\hat{s}|s)$

The slightly funny notation  $p_{\hat{s}|s}(\hat{s}|s)$  is not a typo: it refers to the estimate distribution evaluated at  $\hat{s} = s$ . Of course, we are interested in probability correct overall, not at a given  $s$ , so we have to average over all  $s$ . We will discuss this in Problem 3.X.

### 3.3.3 Expected squared error

Although the notion of correctness is strange for continuous variables, we can make a more intuitive argument for the optimality of the MAP estimator when the posterior is Gaussian. For a Gaussian distribution, the mode (used as the MAP estimate) and the mean are identical. Thus, the MAP estimator is identical to the mean estimator. It turns out that the mean estimator is optimal in a very intuitive sense: it minimizes the expected squared error between the estimate and the true stimulus,  $\langle (\hat{s} - s)^2 \rangle_{p(x,s)}$ , where the expected value  $\langle \rangle$  is over both  $x$  and  $s$ . We will discuss the reason for this mathematical fact in detail in Chapter 9 (Reward), but it does help us understand why the MAP estimate is superior, in two ways.

Intuitively, when the measurements are very noisy, you will be off by least if you pick the mean of the stimulus distribution,  $\mu$ . If the measurement  $x$  is noiseless, you should estimate the stimulus at  $x$ . It makes sense that for intermediate noise levels, a strategy of picking an estimate in between  $\mu$  and  $x$  will cause you to be off by the smallest possible amount on average. Thus, a bias towards the mean of a stimulus distribution is a sign of an optimal strategy.

More mathematically, we can find a simple expression for the expected squared error. We will compute the expected value in two steps: first over  $x$  drawn from a given  $s$ , then over  $s$ :

$$\langle (\hat{s} - s)^2 \rangle_{p(x,s)} = \left\langle \left\langle (\hat{s} - s)^2 \right\rangle_{p(x|s)} \right\rangle_{p(s)} \quad (3.6)$$

We first compute the inside expected value. Using the *bias-variance tradeoff* (see Appendix X for proof), this part can be written as

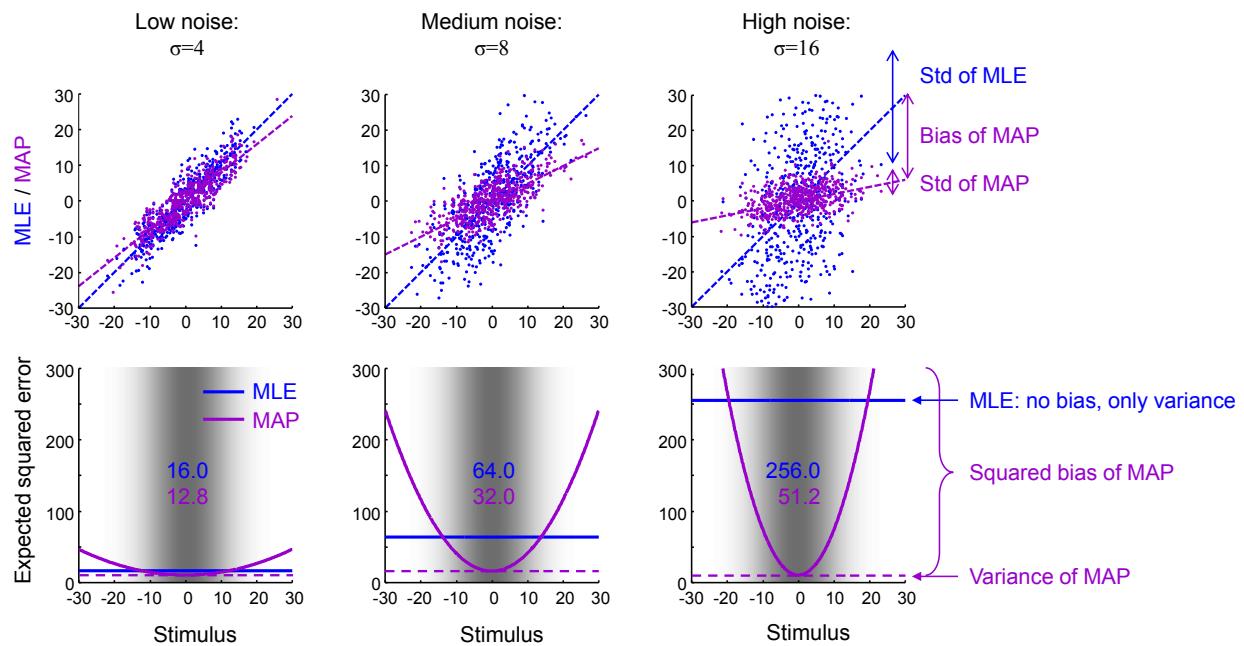
$$\langle (\hat{s} - s)^2 \rangle_{p(x|s)} = \left( \langle \hat{s} \rangle_{p(x|s)} - s \right)^2 + \text{Var}_{p(x|s)} \hat{s} .$$

If  $\hat{s}$  is the MAP estimator, we can simply substitute  $\langle \hat{s} \rangle$  and  $\text{Var}_{p(x|s)} \hat{s}$  from Table 3.1:

$$\begin{aligned} \langle (\hat{s} - s)^2 \rangle_{p(x|s)} &= \left( \frac{\frac{s}{\sigma^2} + \frac{\mu}{\sigma_s^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}} - s \right)^2 + \frac{\frac{1}{\sigma^2}}{\left( \frac{1}{\sigma^2} + \frac{1}{\sigma_s^2} \right)^2} \\ &= \left( \frac{\frac{\mu - s}{\sigma_s^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}} \right)^2 + \frac{\frac{1}{\sigma^2}}{\left( \frac{1}{\sigma^2} + \frac{1}{\sigma_s^2} \right)^2} \end{aligned} \quad (3.7)$$

This can be further simplified, but we keep this form because now the first term is still the square of the bias of the MAP estimator, and the second the variance of the MAP estimator. We have plotted the resulting stimulus-dependent expected squared error as a function of  $s$ . The variance is independent of  $s$ , and the bias grows linearly with the distance between  $s$  and the mean of the stimulus distribution (this mean is 0 in the plot), so the squared bias grows quadratically with that distance. For comparison, we also plotted the expected squared error of the maximum-likelihood estimator,  $\hat{s} = x$ . This estimator is unbiased, since  $\langle \hat{s} \rangle = s$ , and the variance is constant at  $\sigma^2$ . As it turns out, the maximum-likelihood estimator is the estimator that generally has the lowest variance of all unbiased estimators.

The plot allows us to understand why the MAP estimator is optimal even though it is biased. The variance of the MAP is lower than of the MLE. The squared bias gets added to the variance, but their sum still stays below the variance of the MLE as long as the stimulus is close enough to the mean of the stimulus distribution. Of course, that is exactly where most stimuli will be distributed, since the stimulus distribution is Gaussian (shown as a gray shading). This is why when we evaluate the outside expected value in Eq. (3.6) to calculate the overall expected squared error, we end up with a lower number for MAP than for MLE (in the example from the figure, it is 15.6 versus 16.0). See also Problem 3.X.



**Figure 3.3 Comparison between MAP and MLE estimators in the Gaussian Bayesian integration model.** In this example, the width of the stimulus distribution,  $\sigma_s$ , was 8. **Top:** Scatterplots of MLE and MAP estimators against the true stimulus. Dashed lines indicate the expected values. The larger the noise, the lower the slope of the expected value of the MAP. **Bottom:** Expected squared error as a function of the stimulus for the MLE and MAP estimators. Expected squared error (solid lines) is the sum of squared bias and variance. Although the MAP

is biased, the variance of the MAP (dashed purple line) is lower than that of the MLE (blue line). The stimuli that occur often according to the stimulus distribution (shading indicates probability) are such that the stimulus-averaged expected squared error of the MAP (purple number) is always lower than that of the MLE (blue number).

### 3.4 Suboptimal inference

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

- Learning of the generative model has not completed.
- The generative model is too complex to learn and the observer approximates it.
- The generative model of an experiment is different from that in the natural environment, and the observer uses the latter for inference.

**Model mismatch:** For inference, the observer assumes a generative model that is incorrect for the task.

Terminology: when we say MAP estimator, we by default mean MAP estimation under the correct generative model. Otherwise, we will say “MAP estimator using an incorrect generative model”

In the first section, we will focus on the most basic case, that the prior distribution the observer uses during inference,  $q(s)$ , is different from the stimulus distribution, which is still  $p(s)$ . This inference is illustrated in Table 3.2, which should be compared to Table 3.1. It is also possible that the likelihood

	Mean	Variance
STEP 1 Generative model	Stimulus distribution $p(s)$	
	$\mu$	$\sigma_s^2$
STEP 2 Inference	Measurement distribution $p(x s)$	
	$s$	$\sigma^2$
STEP 3 Estimate distribution	Prior distribution $q(s)$	
	$\mu_{\text{assumed}}$	$\sigma_{s, \text{assumed}}^2$
	Likelihood function $L(s;x) = p(x s)$	
STEP 3 Estimate distribution	$x$	$\sigma^2$
	Posterior distribution $q(s x)$	
	$\frac{\frac{x}{\sigma^2} + \frac{\mu_{\text{assumed}}}{\sigma_{s, \text{assumed}}^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_{s, \text{assumed}}^2}} = \hat{s}$	$\frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma_{s, \text{assumed}}^2}}$
STEP 3 Estimate distribution	MAP estimate distribution $p(\hat{s} s)$	
	?	?

function does not correspond to the measurement distribution.

When the observer's prior differs from the world state distribution, the observer will still be performing Bayesian inference (in the sense of computing a posterior distribution), but this inference might no longer be optimal. Bayesian observers are optimal when they possess and correctly incorporate, full knowledge of all distributions in the generative model, but an observer with a mismatched likelihood or prior will in many cases be suboptimal.

### 3.4.1 Prior mismatch

The prior distribution at value  $s$ ,  $p(s)$ , can be thought of as our belief that the stimulus was  $s$  before we have received any sensory information. Within the context of a psychophysics experiment, one cannot blindly assume that subjects learn the world state distribution and use it as a prior. In general, the observer's prior might differ from the world state distribution. Subjects might be using a prior that they come into the experiment with, for example one that is based on the world state distribution in the natural world. Well-established examples are a prior favoring low speeds (see Fig. 2.6) and a prior for light coming from above (see Fig. 2.7) or a. Such "natural" priors, acquired over a lifetime of sensory experience, might be hard to override during the relatively short duration of an experiment. Extensive training might be needed to override a natural prior.

An example of an experiment in which this happened is the sensorimotor task described in Section 2.7, where a target is, unbeknown to the subject, displaced by 1 cm. In that experiment, extensive training was necessary because the experimental prior over displacement was almost certainly very different from the default prior – in the real world, the targets we attempt to reach are not on average displaced by 1 cm! The less experience or training the observer has with the world state distribution used in the experiment, and the more complex this distribution is (for example, multi-peaked), the less likely the observer is to use this distribution as the prior.

Of course, subjects might be using a prior intermediate between the "natural world state distribution" and the "experimental world state distribution". The prior might also change over time, as the observer is exposed to more stimuli during the experiment.

In Step 2, under prior mismatch, the observer would be computing the MAP estimate using a different generative model than in Chapter 2:

$$q(s|x) \propto q(s)p(x|s).$$

If the prior distribution  $q(s)$  has mean  $\mu_{\text{assumed}}$  and standard deviation  $\sigma_{s,\text{assumed}}$ , then the MAP estimate when the measurement is  $x$  is

$$\hat{s} = \frac{\frac{x}{\sigma^2} + \frac{\mu_{\text{assumed}}}{\sigma_{s, \text{assumed}}^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_{s, \text{assumed}}^2}}.$$

### 3.4.2 The case of a flat prior

A special case of prior mismatch is if the observer uses a constant prior on the real line,  $q(s)=\text{constant}$ . Such a prior can never reflect the true stimulus distribution, since a stimulus distribution is an objective distribution and has to be properly normalized. For inference, however, a prior that is not normalizable (*improper*) is ok as long as the posterior is normalizable (see Box 3.3. ; see Box 3.3). Using a flat prior simplifies inference. If the measurement distribution (and therefore the likelihood function) is Gaussian, then the posterior is:

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

This is the simplest possible posterior distribution in the presence of Gaussian noise. The MAP estimate is equal to the ML estimate, which is equal to the measurement,  $\hat{s}_{\text{MAP}} = \hat{s}_{\text{ML}} = x$ , and the distribution of the MAP estimate is equal to the measurement distribution. The ML estimator is an unbiased but suboptimal estimator in the sense that it will not maximize the proportion of correct estimates (see Section **Error! Reference source not found.**).

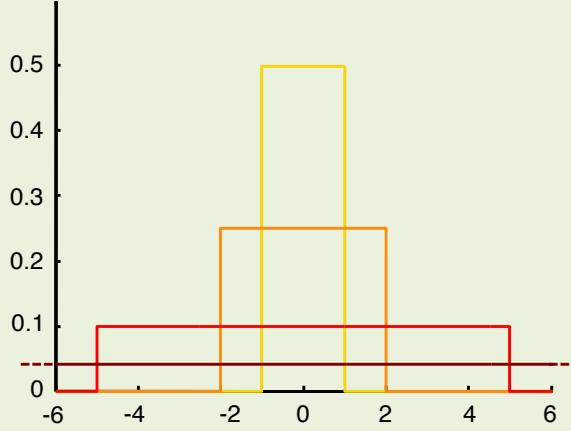
#### Box 3.3: The improper prior

We considered the case of a flat prior,  $p(s)=\text{constant}$ . An immediate question is then what value this constant takes. If  $s$  were limited to an interval, say  $[a,b]$ , the answer would be clear:  $p(s)=1/(b-a)$ , so that the area under the prior is 1 (Fig. 3.X). In this chapter, however, the domain of  $s$  is the entire real line. A uniform distribution is not properly defined on the entire real line, since the line has infinite length, so the uniform distribution would have value 0. In any practical task,  $s$  can of course not grow arbitrarily large in either direction. Therefore, it would be reasonable to cut off its domain at some large value. Choosing this value – and with it, the value of the constant prior – would, however, be arbitrary. Fortunately, this conundrum does not need to be solved, since it turns out that in the inference, the value of the constant prior does not play a role. Namely, if this value is  $c$ , then the posterior distribution is

$$p(s|x) \propto cp(x|s) \propto p(x|s). \quad (3.8)$$

In other words, since the prior is constant, it gets absorbed into the proportionality constant.

After normalizing the posterior distribution,  $c$  would drop out. We could even pretend that  $p(s)=c$  on the entire real line (Fig. 2.X, cyan), and the posterior would still be well-defined in spite of the prior distribution not being normalized. Such an unnormalized prior distribution (i.e., one whose integral is infinite rather than 1) is called an *improper prior*.



**Figure 3.4.** Uniform priors on interval  $[-1, 1]$  (yellow),  $[-2, 2]$  (orange),  $[-5, 5]$  (red), and an improper prior on the entire real line (dark red).

### 3.4.3 Suboptimal estimators

There are other ways of constructing suboptimal observers, such as by simply postulating a different estimator than the MAP estimate. If we had not heard about Bayesian inference, we might have thought that a reasonable estimate of the stimulus would be the average of the mean of the stimulus distribution,  $\mu$ , and the measurement  $x$ :

$$\hat{s} = \frac{x + \mu}{2}$$

Another way to view this estimate is that it is the MAP estimate under a wrong generative model in which  $\sigma$  happens to be equal to  $\sigma_s$ . In general, suboptimal estimators can often be interpreted as MAP estimates under a wrong generative model.

## 3.5 Generalizing the model

Suboptimal inference is an important way in which we can generalize the model of Chapter 2; this discussion followed from relaxing the assumption of perfect observer knowledge of the generative model that we made when describing Step 2 (inference) in Chapter 2. However, we made many other assumptions in Chapter 2 that can be relaxed or generalized. We will discuss different stimulus domains, stimulus-dependent measurement noise, and high-dimensional stimulus spaces, which have to do with Step 1 (the generative model). Finally, we will discuss response noise, which affects Step 3 (estimate distribution).

### 3.5.1 Different stimulus domains

We have assumed that the domain of the stimulus is the entire real line, from  $-\infty$  to  $\infty$ . There are many environmental variables that have a different domain. For example, orientation and motion direction have a circular domain (even location on a line becomes a circular variable if you consider the fact that space “wraps around” the observer’s body like a sphere). Color has a bounded three-dimensional domain, or a circular domain if one restricts oneself to a color wheel. There are also many variables with a domain from 0 to  $\infty$ , including line length, depth, weight, speed, loudness, duration, and surface lightness. These can be called magnitude variables, and none of them can ever take negative values. For none of these variables do Gaussian distributions for either the stimulus distribution or the measurement distribution make sense.

For circular variables, one solution is to choose a *Von Mises* distribution (see Fig. 3.X). This can be regarded as the circular analog of a Gaussian distribution. It has two parameters: a *circular mean*, and a *concentration parameter*, which is similar to the reciprocal of the variance of a Gaussian. A Von Mises distribution over a circular variable  $s$  with domain  $[0, 2\pi)$ , circular mean  $\mu$ , and concentration parameter  $\kappa_s$  is

$$p(s) \propto e^{-\kappa_s \cos(s-\mu)} . \quad (3.9)$$

This is not yet a normalized distribution, which is why we use a proportionality sign (see Problem 3.5). Similarly, the distribution of a measurement given a stimulus can be Von Mises, with circular mean  $s$  and concentration parameter  $\kappa$ :

$$p(x|s) \propto e^{-\kappa \cos(x-s)} . \quad (3.10)$$

For magnitude variables, a common choice is a log-normal distribution. Since the domain of  $s$  is  $[0, \infty)$ , the domain of the logarithm of  $s$  is the entire real line. Thus, it is possible to define a Gaussian distribution on  $\log s$ :

$$p(\log s) = \frac{1}{\sqrt{2\pi\sigma_{\log}^2}} e^{-\frac{(\log s - \log \mu)^2}{2\sigma_{\log}^2}} .$$

#### Study Tip

Read Appendix X if you don’t know how to transform probability distributions

Transforming this to the original variable  $s$ , we obtain

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

This is called the log-normal distribution. Examples are shown in Fig. X).

The log-normal distribution can also be used for the measurement distribution:

$$p(x|s) = \frac{1}{\sqrt{2\pi\sigma_{\log}^2}} e^{-\frac{(\log x - \log s)^2}{2\sigma_{\log}^2}}. \quad (3.11)$$

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

### 3.5.2 Stimulus-dependent noise

Even when the stimulus domain is the real line and the measurement distributions are Gaussian, the model of Chapter 2 makes a strong assumption: that the level of measurement noise is independent of the stimulus itself. There are many reasons why this would not be the case. For example, the level of noise in the measurement of horizontal stimulus position will depend on the density of photoreceptors in the retina that encode that position, as well as on the size of their receptive fields. At greater distance from the fovea (this distance is called retinal eccentricity), the density decreases and the receptive field size increases, causing an increase in the level of measurement noise. Since eccentricity is equivalent to horizontal position if the subject is foveating a point on the line, the parameter  $\sigma$  of the Gaussian distribution will depend on the horizontal position (or eccentricity)  $s$ :

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

While at first glance, this does not seem to be a big deal, this has a major consequence for the calculations we did in Chapter 2. The reason is that even though the measurement distribution is Gaussian, the likelihood function,  $p(x|s)$  as a function of  $s$ , is not! For example, suppose that variance increases quadratically with  $s$  as  $\sigma(s)^2 = a + bs^2$ , where  $a$  and  $b$  are constants, then

$$L(s; x) = p(x|s) = \frac{1}{\sqrt{2\pi(a+bs^2)}} e^{-\frac{(x-s)^2}{2(a+bs^2)}}.$$

Suddenly, the stimulus dependence appears in multiple places, and this is by no means a Gaussian function anymore. Examples of this likelihood function are shown in Fig. 3.X. If the likelihood is no longer Gaussian, then the analytical expressions in Chapter 2 for the posterior, the MAP estimate, and the MAP estimate distribution no longer hold. What is worse, analytical expressions for the MAP estimate and the MAP estimate distribution do not even exist anymore! This makes it necessary to perform simulations. While obtaining results from simulations is considerably more cumbersome than plotting an equation, it is also the reality of almost every practical Bayesian model.

Another example of stimulus-dependent measurement noise is orientation of a line. Horizontal and vertical orientations in a visual scene are corrupted by less measurement noise than other orientations.

Exercise: what would this change in Eq. (3.10)?

In touch,  $\sigma$  is smaller for stimuli applied to the fingertip than to the forearm.

### 3.5.3 Higher-dimensional stimulus spaces

We discussed localization on a line; thus, the stimulus variable, horizontal position, was a one-dimensional stimulus variable. However, an event could in principle take place anywhere in three-dimensional space, in which case the world state variable of interest would be a three-dimensional vector.

Another important example of a high-dimensional stimulus space is a visual image. If we represent an image through its pixel intensities, then it is represented by a vector with as many dimensions as there are pixels. For example, a small 10 pixel by 10 pixel image would correspond to a 100-dimensional vector of intensity values (300-dimensional if it were a color image, since each pixel would have an Red, Green, and Blue channel).

In high-dimensional spaces, one can still define stimulus and measurement distributions. For example, we represent an image by a vector  $\mathbf{s} = (s_1, s_2, s_3, \dots, s_N)$ , where  $N$  is the number of pixels. Then the stimulus distribution would be denoted by  $p(\mathbf{s})$ . Similarly, the observations would consist of a measurement vector  $\mathbf{x} = (x_1, x_2, x_3, \dots, x_N)$ , and its conditional distribution would be denoted by  $p(\mathbf{x}|\mathbf{s})$ .

**Vector notation:** We use boldface for vectors. For example,  $\mathbf{s}$  could be a vector consisting of components  $(s_1, s_2, s_3)$ .

A simple, and grossly oversimplified generative model would be one in which every pixel intensity is independently drawn from its own Gaussian distribution: the first pixel intensity from  $p(s_1)$ , the second from  $p(s_2)$ . If the pixels are independent, then the probability of the entire vector  $\mathbf{s}$  is the product of the probabilities of the individual pixel intensities:

$$p(\mathbf{s}) = p(s_1) \cdot p(s_2) \cdots p(s_N).$$

We also say that the multidimensional distribution *factorizes*. Similarly, if the noise corrupting the measurements of the individual pixel intensities is independent (uncorrelated) between pixels, then the measurement distribution will factorize:

$$p(\mathbf{x} | \mathbf{s}) = p(x_1 | s_1) \cdot p(x_2 | s_2) \cdots p(x_N | s_N).$$

Each of the factors could be a Gaussian distribution similar to Chapter 2.

### 3.5.4 Response noise

In principle, the observer's response could be the MAP estimate. In practice, every response on a continuum will be subject to some response (e.g., motor) noise, reflecting for instance the accuracy with which the observer is able to position the computer cursor. Moreover, the observer's memory of the MAP estimate might decay slightly between the moment the estimate is made and the moment that the response is submitted. Thus, the observer's response distribution is not necessarily the same as the estimate distribution. A complete model of the task would include the response noise. For example, the observer's response could be drawn from a Gaussian distribution with mean  $\hat{s}_{\text{MAP}}$  and some variance  $\sigma_{\text{motor}}^2$ . We will treat this case in Problem 8. However, response noise is not central to the Bayesian formalism.

## 3.6 Experimental tests of the model

### 3.6.1 Why would the brain be Bayesian?

So far, we have discussed that MAP estimation maximizes accuracy and in some cases minimizes squared error. Why would we believe this formalism to be relevant for behavior? The justification is twofold. The first is evolutionary: basic perception is essential for survival across the animal kingdom, and it would not be surprising that over the course of hundreds of millions of years, the perceptual system would have learned to deal well with uncertain sensory input. This argument is normative: it postulates that the brain has the *goal* of optimizing performance. The second justification of testing Bayesian models is practical: although eventually it is important to compare multiple models, it is convenient to start with one model that can be formulated without even having done the experiment!

We will now look at some experiments that lend support to the hypothesis that the brain is Bayesian.

### Box: David Marr

David Marr was a vision scientist who conceptualized the study of the brain at three different levels: computational, algorithmic, and implementational. At the computational level, the brain is seen as part of an organism trying to maximize its chances of survival. Evolutionary pressure thus creates a goal that the system is pursuing. This is the fundamental motivation for the notion that perception might in many cases be Bayesian.

### 3.6.2 Empirical evidence

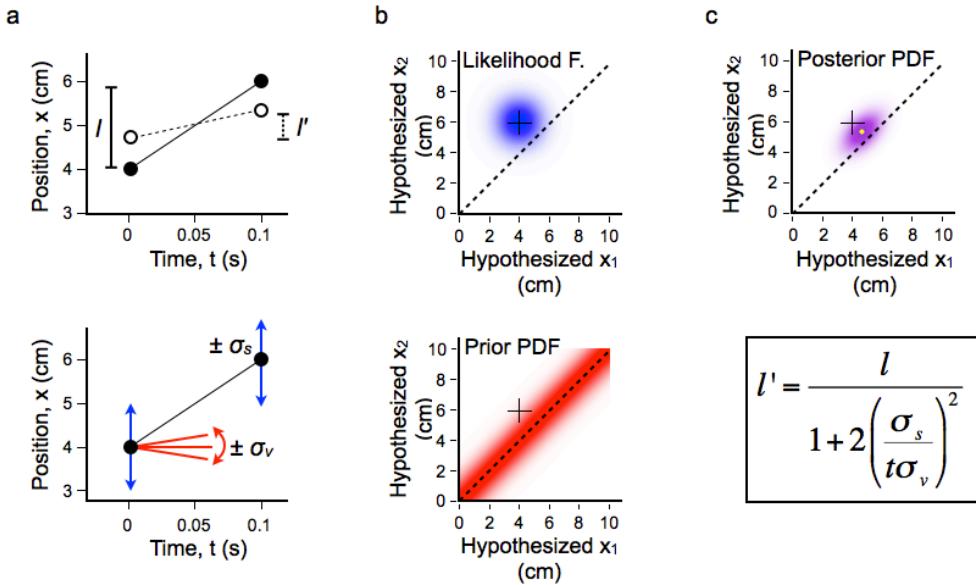
Here, we describe empirical evidence that observers combine a likelihood with a prior, beyond only the Gaussian example from Chapter 2.

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

Weiss, Simoncelli, and Adelson (2002) studied the perception of visual motion, in particular a large set of motion illusions. For example, Thompson (1980 and 1982) found that low-contrast stimuli (e.g., a gray grating on a background of a different grayscale) appear to move slower than high-contrast stimuli (for example, a white grating on the same background), even though in reality they move at the same speed. You can observe this yourself in the demo that we put on the book's website. Weiss and colleagues explain this puzzling phenomenon and other illusions using a prior distribution over speed that is higher for lower speeds. This is explained using the equivalent of Eq. **Error! Reference source not found.**: a lower-contrast grating provides less reliable information, which means that the likelihood function is wider, which in turn means that the MAP estimate is shifted more towards the mean of the prior. If the prior peaks at 0, then the perceived speed will be shifted towards a lower value when contrast is lower. This paper stands out by providing strong evidence against previously highly regarded but ad hoc models of motion perception. This illustrates that Bayesian models can often successfully challenge long-held beliefs in a field. In later work, Stocker and Simoncelli (2006) refined the Bayesian model for motion perception by estimating the shape of the prior. They did this by obtaining a predicted distribution of MAP estimates in a way analogous to what we did in this

chapter, but using a much more flexible prior, and subsequently fitting the parameters of that prior.

In the realm of somatosensory perception, Goldreich (2007) showed how tactile spatiotemporal illusions – the perceived contraction of the distance between two taps delivered in rapid succession to the skin – can be explained using the model in this chapter, again with a prior expectation for low-speed movement (Fig. 2.7). The proposed model provides a unified explanation for several tactile illusions (Fig. 2.8).



**Figure 2.7.** A Bayesian model for tactile spatiotemporal illusions (adapted from Goldreich, 2007). **(a)** Upper panel: when the skin is tapped in rapid temporal succession at two locations separated by distance  $l$  (filled circles), humans perceive the taps (open circles) to be closer together than they really are. The perceived distance,  $l$ -prime, is less than the true distance,  $l$ , a phenomenon known as *perceptual length contraction*. Lower panel: a Bayesian model in which the observer expects successive skin contacts to jump slowly, if at all (Gaussian prior centered at zero velocity with standard deviation  $\sigma_v$ ), and in which the measured location of each contact is noisy, resulting in a Gaussian likelihood function with standard deviation  $\sigma_s$ . **(b)** The likelihood function (upper) and prior PDF (lower) over tap locations. Each pixel represents a particular hypothesized first tap and second tap location. Darker color represents higher probability. Dotted line indicates hypotheses for which  $\text{tap1} = \text{tap2}$  (i.e., no movement along the skin, the most probable outcome according to the prior PDF). Although the likelihood function will vary from trial to trial, on average it is centered on the true tap location, indicated by the crosshairs in both panels. **(c)** Upper panel: the posterior PDF is a compromise between the likelihood function and the prior PDF. The MAP estimate (yellow dot) is taken as the perception. Lower panel: the perceptual length contraction formula, derived from the posterior PDF, predicts that perceived

distance will shrink progressively as the time,  $t$ , between taps is shortened; for a given  $t$ , perceptual length contraction is more pronounced when the ratio  $\sigma_s / \sigma_v$  is larger.

Add more: Bayesian integration model of sound localization by the barn owl ([Fischer & Pena 2011](#))

<http://www.sciencedirect.com/science/article/pii/S0960982213003321>

### 3.7 Problems

**Problem 3.1.** If we do not limit ourselves to Gaussian priors, the posterior uncertainty (defined as in Section 3.1.1) can be larger than the likelihood uncertainty. Work out an example.

**Problem 3.2.** In Section 3.1.1, we defined uncertainty as the standard deviation of a belief distribution. An alternative is as the *entropy* of the distribution. Look up the definition of entropy.

- a) Show that for a Gaussian distribution, the definitions are equivalent.
- b) Construct an example of two belief distributions with the same entropy but different standard deviation.
- c) What are the advantages and disadvantages of either definition?

**Problem 3.3.** In Eq. (3.3), we saw that when all distribution are Gaussian, having a non-flat prior increases confidence in this case. This is, however, not generally true. Construct a prior such that for a given Gaussian likelihood function (i.e. a single trial), confidence is lower with than without the prior.

### Problem 3.4. Overall variance of the MAP estimate

Refer back to Section 3.2.5 about the overall estimate distribution.

- a) Show that across all trials in an experiment, the MAP estimate will have mean  $\mu$  and variance  $\frac{\sigma_s^4}{\sigma^2 + \sigma_s^2}$  (Hint: Use Box 2.10 on linear combinations of normally distributed variables.
- b) Show that the overall variance of the MAP estimator from Part (a) is always smaller than the overall variance of the ML estimator.
- c) Perform the sanity check of  $\sigma=0$  on the result of part (a): in the absence of sensory noise, the observer's estimate will always be identical to \_\_\_\_\_, which in turn will be identical to \_\_\_\_\_, which is distributed according to the \_\_\_\_\_ distribution. Thus, the overall estimate distribution should then be equal to the \_\_\_\_\_ distribution, so its variance will be \_\_\_\_\_, in accordance with the result from part (a).
- d) Perform a similar sanity check corresponding to  $\sigma \rightarrow \infty$ . (Hint: if sensory noise is extremely large, what can be said about the observer's estimate?)

**Problem 3.5. Overall expected squared error of the MAP estimate**

In Section 3.3.2, we calculated the expected squared error of the MAP estimator for given  $s$ , Eq. (3.7). We now calculate the average of this quantity over all  $s$  (i.e. not only averaged over  $x$ , but over both  $x$  and  $s$ ). This is the mean squared error one would expect in a sufficiently long experiment. Keep in mind that the distribution of  $s$  is  $p(s)$ .

- For the Gaussian Bayesian integration case, calculate the overall expected squared error.

Hint:  $\langle (s - \mu)^2 \rangle = \text{Var } s$ .

- Any estimate  $\hat{s}$  can be thought of as a function of the measurement, denoted  $F(x)$ . Consider all estimators for which this function is a weighted average of the measurement and the mean of the stimulus distribution:  $\hat{s} = F(x) \equiv ax + (1-a)\mu$ . Show that among these estimators, the MAP estimate is the one that has the lowest expected squared error across all trials in the experiment.
- Repeat for all linear combinations of the measurement:  $\hat{s} = ax + b$ .

**Problem 3.6**

For the Von Mises distribution in Eq. (3.9),

- Look up the appropriate normalization.
- Plot it as a function of  $\kappa$ .
- How does this compare to the normalization of a Gaussian as a function of  $\frac{1}{\sigma^2}$ ?

**Problem 3.7.** In discussing magnitude variables in Section 3.5.1, we defined a distribution over the log of the measurement.

- Making use of the rules for transforming random variables (see Appendix Section 11), to derive Eq. (3.11).
- Show that the standard deviation of this distribution is proportional to the mean. This relates the log-normal distribution to Weber's law.

**Problem 3.8.** At the end of Section 3.5.3, we mentioned that an observer's response might be corrupted by response or motor noise. Assume the model discussed in this chapter. Assume that motor noise is present and can be modeled as zero-mean Gaussian noise with standard deviation  $\sigma_m$ .

- What is the distribution of the observer's response when the true stimulus is  $s_{\text{true}}$ ?
- How can one experimentally distinguish motor noise from noise in the measurement?  
(There are multiple correct answers to this question.)

**Problem 3.9.** We define *relative bias* within the context of Chapter 2 as the difference between the mean MAP estimate and the true stimulus divided by the difference between the mean of the

prior and the true stimulus. Calculate relative bias as a function of the ratio  $R = \frac{\sigma}{\sigma_s}$ . Plot relative bias as a function of  $R$  and interpret the plot.

**Problem 3.10.** In Section 3.5.1, we discussed stimulus variables that take values on the circle, such as motion direction, which takes values between (for instance)  $-\pi$  and  $\pi$ . Suppose that in a laboratory experiment, motion direction is drawn from a Von Mises distribution given by Eq. (3.9). Assume that the measurement distribution is also Von Mises, given by Eq. (3.10). Show that the posterior over  $s$ ,  $p(s|x)$ , is a Von Mises distribution and find expressions for the cosine of its mean, the sine of its mean, and its concentration parameter.

**Problem 3.11.** In this chapter, the prior was Gaussian and so was the posterior. If the prior is not Gaussian, the posterior might not be Gaussian either. Can you construct a prior that would give rise to a posterior with two local maxima?

**Problem 3.12\*.** In this chapter, the posterior distribution and the distribution of the MAP estimate were both Gaussian, i.e. they belonged to the same family of functions. Construct an analytically solvable example in which all distributions and functions are continuous but the estimate distribution for given  $s$  does not belong to the same family of functions as the posterior distribution. (“Analytically solvable” means that you can write down a closed-form expression for every relevant distribution and function.)

**Problem 3.13\*.** The joint distribution of stimulus and measurement,  $p(s,x)$ , can be calculated as  $p(x|s)p(s)$ .

- a) In our example, show that this distribution is a bivariate Gaussian and compute its mean vector and covariance matrix.
- b) Show that the correlation coefficient between  $s$  and  $x$  is....
- c) Perform sanity checks on the expression in (b), by taking appropriate limits of  $\sigma$  and  $\sigma_s$ .
- d) How can the correlation between  $s$  and  $x$  be seen in Fig. 3.XA?

**Problem 3.14\*. Overall expected accuracy of the MAP estimate**

MAP estimation is the read-out method that maximizes expected accuracy across a large number of trials. Here, we verify this claim for the model of Chapter 2 and for linear estimators. Referring to Section 3.3.2, the accuracy of an estimator  $\hat{s}$  when the stimulus is  $s$  should be defined as the estimate distribution evaluated at  $\hat{s} = s$ , denoted by  $p_{\hat{s}|s}(s|s)$ .

- a) Using the model of Chapter 2 (Table 3.1), show that the accuracy of the MAP estimator

$$\text{when the stimulus is } s \text{ is } p_{\hat{s}|s}(s|s) = \frac{\sigma}{\sqrt{2\pi}} \left( \frac{1}{\sigma^2} + \frac{1}{\sigma_s^2} \right) e^{-\frac{(s-\mu)^2}{2\sigma^2}}.$$

- b) To calculate overall expected accuracy across many trials where stimuli are drawn from  $p(s)$ , we have to average  $p_{\hat{s}|s}(s|s)$  over  $s$  using  $p(s)$ :

$$\text{overall expected accuracy} = \langle p_{\hat{s}|s}(s|s) \rangle_{p(s)}. \text{ Show that this is equal to } \sqrt{\frac{1}{2\pi} \left( \frac{1}{\sigma^2} + \frac{1}{\sigma_s^2} \right)}.$$

- c) Expected accuracy must have a higher value when the observer performs MAP estimation than when they use any other method of estimation. Consider all estimators that are weighted averages of the measurement and the mean of the stimulus distribution:  $\hat{s} = ax + (1-a)\mu$ . Show that among these estimators, the MAP estimate is the one that maximizes expected accuracy.
- a) Repeat for all linear combinations of the measurement and the mean of the stimulus distribution:  $\hat{s} = ax + b\mu$ .

**Problem 3.15.** Assume the stimulus and measurement distributions of Chapter 2, but an observer who uses a Gaussian prior with standard deviation  $\sigma_p \neq \sigma_s$  (but the correct  $\mu$ ). This observer's MAP estimate can be thought of as being based on an incorrect generative model.

- a) When the measurement is  $x$ , what is the MAP estimate?  
 b) What are the mean and variance of the MAP estimate for given  $s$ ?  
 c) What are the mean and variance of the MAP estimate across all trials in the experiment?  
 (Hint: the expression should contain both  $\sigma_p$  and  $\sigma_s$ .)

**Problem 3.16.** Repeat the previous problem but now under the scenario that the observer's assumed prior has a mean  $\mu_p$ , which can be different from the correct  $\mu$ .

**Problem 3.17.** When the observer uses a wrong prior to compute the posterior, can decision confidence be higher than when using the correct prior? If so, give an example. If not, prove it.

**Problem 3.18** Suppose a stimulus  $s$  takes discrete values, so given the measurement  $x$ , the posterior distribution  $p(s|x)$  is a discrete-valued distribution in  $s$ . Show that the MAP estimate maximizes proportion correct.

**Problem 3.19. The “blind” estimator.** On the other extreme of the MLE, there is the estimator  $\hat{s} = \mu$ . One of the following questions is a trick question.

- a) Derive an expression for expected squared error of this estimator.  
 b) How large must  $\sigma$  be for this estimator to have a lower overall expected squared error (see Problem 3.5?) than the MLE?  
 c) How large must  $\sigma$  be for this estimator to have a lower overall expected squared error than the MAP?

## Contents

4	Cue combination.....	4-1
4.1	What is cue combination and what purpose does it serve? .....	4-2
4.1.1	Why combine cues .....	4-2
4.1.2	Combining cues with different levels of noise.....	4-3
4.2	Formulation of the Bayesian model .....	4-5
4.2.1	Step 1: Generative model .....	4-5
4.2.2	Step 2: Inference.....	4-8
4.2.3	Step 3: Estimate distribution.....	4-11
4.3	Artificial cue conflicts.....	4-12
4.4	Distinguishing the distributions.....	4-14
4.5	Cue combination under ambiguity.....	4-15
4.6	Tests of model predictions.....	4-15
4.7	Concluding remarks.....	4-18
4.8	Further reading .....	4-19
4.9	Problems .....	4-20

## 4 Chapter 3: Cue combination

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

In Chapter 2, we studied the simplest possible generative model, with a single stimulus,  $s$ , and a single measurement,  $x$ . Here, we study an extension in which there are two measurements, say  $x_A$  and  $x_V$ . These are based on sensory inputs that are also called cues. The measurements could correspond to an auditory and a visual measurement of a stimulus, such as the location at which a ball drops on the ground. The observer estimates the stimulus value based on both cues.

There are three important reasons to study this generative model. First, cue combination occurs very commonly in daily life. Second, it is an early and still prominent domain of application of Bayesian modeling. Third, this generative model is our first example in which the Bayesian observer computes the likelihood function over the world state of interest from two simpler likelihoods.

Plan of the chapter: We will start this chapter by discussing the intuitions behind cue combination. We will then develop the mathematical background of optimal cue

combination. Specifically, we will ask how the nervous system could combine cues in a way that is close to optimal. Next, we will discuss the scientific literature that addresses how well people actually combine cues. We will discuss how the Bayesian cue combination framework relates to illusions that have been described.

#### **4.1 What is cue combination and what purpose does it serve?**

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

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

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

##### **4.1.1 Why combine cues**

Why should the brain combine cues? To answer this question, let's explore the consequences of an obvious alternate strategy: the brain could simply use the single most

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

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

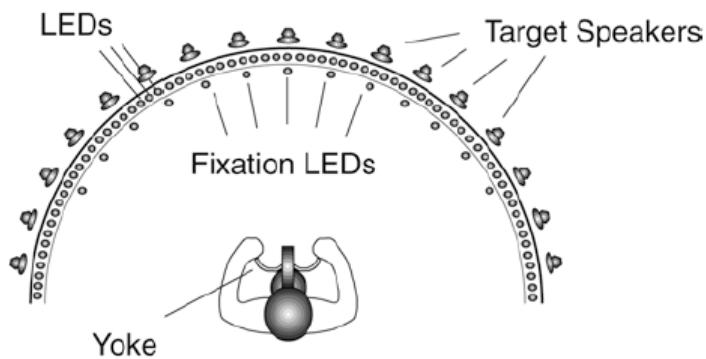
Second, let's suppose we wish to infer the identity of a spherical object, placed on a tabletop, in the dark. As we rest our hand on the object, our proprioceptors inform us about its size, but we cannot unambiguously identify an object from its size alone. In this case, our uncertainty about the object's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

#### 4.1.2 Combining cues with different levels of noise

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

rely on vision. When audition is more informative than vision, we expect the reverse – and indeed at night we often rely primarily on our auditory system.

Over the last couple decades many scientists have studied cue combination in the laboratory. In a typical experiment, the subject is surrounded by an array of loudspeakers and light-emitting diodes (LEDs, see Fig. 4.1). An audiovisual stimulus is produced by the simultaneous occurrence of a brief auditory beep and a light flash. The subject is instructed to indicate the perceived location of the beep. With this apparatus, scientists can probe how visual and auditory cues are combined by the nervous system.



**Figure 4.1.** Experimental set-up for testing auditory localization or multisensory perception. A speaker produces a brief tone. At the same time, an LED might produce a brief flash. The subject points a laser pointer at the perceived location of the tone. Figure from Wallace et al. (2004).

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

Related to these finding is the famous ventriloquist illusion, in which a puppeteer moves the head and mouth of a puppet in sync with his own (visually concealed) speech (Fig. 4.2). Even though the actual sound source – the ventriloquist – is displaced from the puppet, when performed well this procedure produces a powerful illusion of the puppet talking,. This illusion was used in the late Middle Ages to disprove the claim that perception has access to the actual outside truth in the world. It has been used recently to investigate how the brain combines cues. For example, we might want to know which

criteria are necessary for the brain to fall for the illusion: How closely timed must the speech and the movements of the puppet be for the illusion to occur? How far apart spatially can the puppeteer be from the puppet? By investigating these and related questions, we can more fully characterize the audiovisual cue combination process.

Interestingly, we do not have to attend a ventriloquist show in order to experience the ventriloquist illusion. We experience the illusion whenever we watch television. We see the action, such as people talking, on the TV screen, while the sound comes from speakers in a different location – to the sides, or underneath the screen. Yet this displacement of the audio and visual cues poses no problem for perception; we still have the strong impression that the sound of an actor speaking is coming from the screen. Cue combination enables us to enjoy movies.



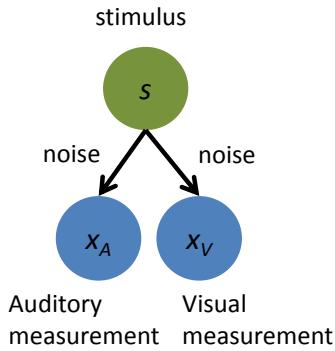
**Figure 4.2.** Gareth Oliver, a British ventriloquist, with his puppet Pava, performing during Britain's Got Talent in 2009. Skilled performers like Oliver can make the observer believe that the speech is originating from the puppet.

## 4.2 Formulation of the Bayesian model

Here we formulate a Bayesian model for optimal cue combination. In developing our formulation, we use the example of auditory-visual location estimation, but the same approach can be applied to other cue combination scenarios. Our approach follows the same three steps of Bayesian modeling outlined in Chapter 2.

### 4.2.1 Step 1: Generative model

The generative model is shown in Figure 4.3.



**Figure 4.3.** Generative model of cue combination.

It consists of three nodes: the stimulus  $s$ , a visual measurement  $x_V$ , and an auditory measurement,  $x_A$ . Associated with the stimulus is a stimulus distribution  $p(s)$ . Unlike in Chapter 2, here we assume that the prior distribution is flat (see also Section 3.2.2). We do this not only because it is the most common assumption in cue combination studies, but also because it illustrates nicely that a nonuniform prior is not necessary for a Bayesian model to be interesting and important.

**Myth:** Bayesian inference is all about priors.

**Truth:** In many interesting Bayesian models, the prior is flat. Cue combination is the best-known example. What makes the Bayesian model for cue combination interesting is that the MAP estimate is not just determined by the measurements but also by their precisions / reliabilities.

Associated with the measurements are noise distributions  $p(x_A|s)$  and  $p(x_V|s)$ . The separate arrows pointing to  $x_A$  and  $x_V$  reflect a key assumption, namely that these measurements are conditionally independent. Conditional independence of two random variables (see Box) means that they are independent of each other when conditioned on another variable, in this case the actual stimulus  $s$ . Specifically, this means that while vision is noisy and audition is noisy, the noise corrupting the two streams is uncorrelated: there is zero noise covariation between the two modalities.

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

Mathematically, the conditional independence of  $x_A$  and  $x_V$  given  $s$  is expressed as

$$p(x_A, x_V | s) = p(x_A | s) p(x_V | s). \quad (3.1)$$

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

One might wonder what the generative model would look like if the two measurements were *not* assumed conditionally independent.

Exercise 4.1: What do you think?

In that case, the two measurements would have to be considered as a pair, so instead of two bottom-level nodes, each containing one measurement, there would be one bottom-level node containing both measurements. It might then be possible to specify the distribution  $p(x_A, x_V | s)$  in some way, but it would not be a product of a distribution over  $x_A$  and another over  $x_V$ .

**Box: Conditional independence.**

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

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

$$p(x, y | z) = p(x | z) p(y | z) \quad (3.2)$$

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

Exercise 4.2: Think of another real-world example of conditional independence.

Another way to interpret Eq. (3.2) is by starting from a rule that is generally true, the product rule for probabilities:

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

Then to get to Eq. (3.2), we have to make the assumption that  $p(x|y,z) = p(x|z)$ . In other words, knowledge of  $z$  fully specifies our understanding of the probability of  $x$ . When we know  $z$ , also knowing  $y$  does not contribute anything to our assessment of the probability of  $x$ .

For the distribution of an individual measurement: we choose a Gaussian distribution, just as we did in Chapter 2:

$$p(x_A | s) = \frac{1}{\sqrt{2\pi\sigma_A^2}} e^{-\frac{(x_A - s)^2}{2\sigma_A^2}}$$

$$p(x_V | s) = \frac{1}{\sqrt{2\pi\sigma_V^2}} e^{-\frac{(x_V - s)^2}{2\sigma_V^2}}$$

Thus, the generative model contains two variances: one for the auditory measurement ( $\sigma_A^2$ ) and one for the visual measurement ( $\sigma_V^2$ ). Note that, on any given trial, the measurements will generally be unequal, because they are produced by independent noise processes.

#### 4.2.2 Step 2: Inference

The observer infers the stimulus  $s$  from the measurements  $x_A$  and  $x_V$ . The auditory and visual likelihoods over the stimulus are the same as the noise distributions but regarded as a function of  $s$ :

$$L_A(s) = p(x_A | s) = \frac{1}{\sqrt{2\pi\sigma_A^2}} e^{-\frac{(x_A - s)^2}{2\sigma_A^2}}$$

$$L_V(s) = p(x_V | s) = \frac{1}{\sqrt{2\pi\sigma_V^2}} e^{-\frac{(x_V - s)^2}{2\sigma_V^2}}$$

We call each of these an *elementary* likelihood function, defined as a likelihood function over a stimulus feature associated with an individual measurement. Similarly, the *elementary ML estimate* is equal to the measurement.

The posterior distribution over the stimulus is computed from Bayes' rule:

$$p(s|x_A, x_V) \propto p(x_A, x_V | s) p(s).$$

We have left out the factor  $1/p(x_A, x_V)$  for the same reason as in Chapter 2: it only acts as a normalization, so if we normalize the distribution in the end, we automatically take this factor into account. Since the stimulus distribution is flat, the prior is flat as well, and the posterior is determined by the likelihood  $p(x_A, x_V | s)$  only. To make progress, we use Eq. (3.1), the assumption of conditional independence of the measurements. Then, the posterior becomes

$$p(s|x_A, x_V) \propto p(x_A | s) p(x_V | s) p(s).$$

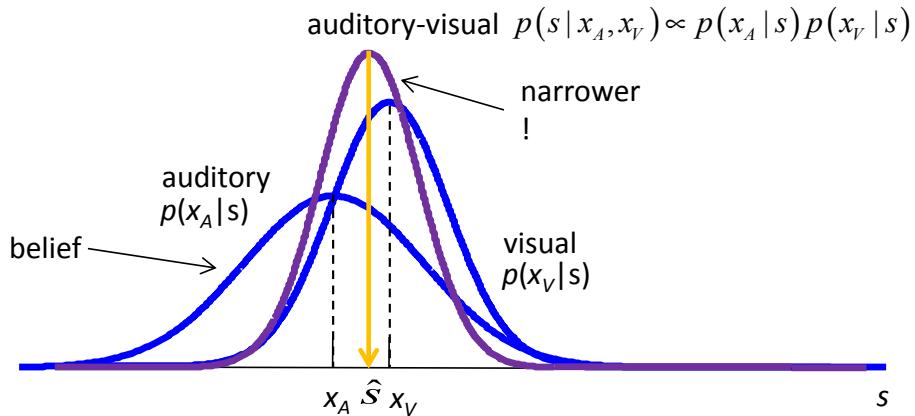
This step – making use of the structure of the generative model to express the likelihood in terms of elementary likelihoods – is the only conceptually new element in this chapter compared to Chapter 2. Expressing the likelihood over the state-of-the-world variable in terms of elementary likelihoods is at the core of Bayesian inference models for many tasks.

We can substitute the two Gaussian distributions into this equation and simplify in the same way we did in Chapter 2. The result is that the posterior is another Gaussian distribution (Fig. 4.4),

$$p(s|x_A, x_V) = \frac{1}{\sqrt{2\pi\sigma_{\text{combined}}^2}} e^{-\frac{(s-\mu_{\text{combined}})^2}{2\sigma_{\text{combined}}^2}},$$

$$\text{where the mean is } \mu_{\text{combined}} = \frac{\frac{x_A}{\sigma_A^2} + \frac{x_V}{\sigma_V^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2}} \text{ and the variance is } \sigma_{\text{combined}}^2 = \frac{1}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2}}.$$

Exercise 4.3: Show this.



**Figure 4.4.** The computation of the posterior in cue combination. Note that both blue curves are likelihood functions. The prior is flat and not shown.

You might have noticed that the observer's computation of the posterior is exactly analogous to the computation in Chapter 2, where we combined a single cue with a prior. The second cue has now taken the role of the mean of the prior distribution. In the present example, the prior is flat, and therefore the posterior is equal to the normalized likelihood function.

Since the posterior is Gaussian, the MAP estimate is equal to the mean:

$$\hat{s}_{\text{MAP}} = \frac{\frac{x_A}{\sigma_A^2} + \frac{x_V}{\sigma_V^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2}}. \quad (3.3)$$

We can rewrite this as a linear weighting

$$\hat{s}_{\text{MAP}} = w_A x_A + w_V x_V$$

where the weights are proportional to the inverse variances:

$$w_A = \frac{\frac{1}{\sigma_A^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2}}, \quad w_V = \frac{\frac{1}{\sigma_V^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2}} \quad (3.4)$$

These weights sum to 1, indicating that the MAP estimate is a weighted average of the two measurements. Averaging with weights proportional to the inverse of the variance is by far the most frequently used model in cue combination. An analogy is that of a police

investigator trying to reconstruct a crime based on the testimonies of two witnesses (Fig. 4.5). One witness was drunk at the time of the crime, the other wasn't. The testimony of the sober witness would be the result of a low-noise process; of the inebriated witness, the result of a high-noise process. A good investigator would take both testimonies into account, but would more heavily weight the testimony of the less noisy witness.



**Figure 4.5.** Weighting measurements by reliability: a good crime investigator would rely more on the testimony of the “less noisy” witness.

Note that because the prior is flat, the MAP estimate of the stimulus is equal to the ML estimate based on both cues. It is in general not equal to either of the *elementary* ML estimates, because those are the measurements  $x_A$  and  $x_V$ .

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

Exercise 4.4: Prove this.

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

#### 4.2.3 Step 3: Estimate distribution

As the third step in our Bayesian model, we are interested in the distribution of the MAP estimate across many trials. The MAP estimate is given as a function of the measurements  $x_A$  and  $x_V$  in Eq. (3.3), but the measurements are themselves random variables – their values vary from trial to trial. As a consequence, the MAP estimate varies from trial to trial as well. Since in a behavioral experiment, we never know the measurements on a single trial (they are in the observer's head), we have to compare

behavior with the distribution of the MAP estimate over many trials. To find the mean and variance of the MAP estimate, we apply the rules for linear combinations of normally distributed variables from the Box in Section 2.4. The means of  $x_A$  and  $x_V$  are both  $s$ . Therefore, the model predicts that the mean MAP estimate will be  $w_A s + w_V s = s$ . In other words, for cue combination with a flat prior, the MAP estimator is unbiased. (Recall from Section 3.5 that bias is defined as the difference between the mean estimate and the true stimulus.) The variance of the MAP estimate will be  $\frac{1}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2}}$

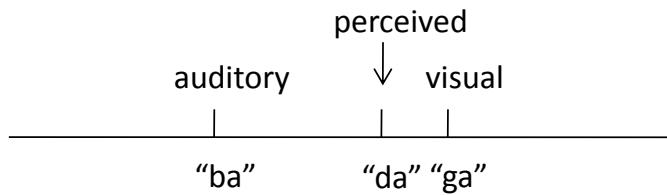
Exercise 4.5: Prove this.

Thus, in this cue combination problem, the variance of the MAP estimate distribution happens to be identical to the variance of the posterior. This stands in contrast to Chapter 2, where we found that the variance of the estimate distribution was different from the variance of the posterior. This difference arises because the prior was chosen flat in the current chapter. When the prior is Gaussian, then the variance of the MAP estimate will differ from the variance of the posterior also for cue combination (see Section 4.4).

### 4.3 Artificial cue conflicts

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

The ventriloquist effect as performed by a professional is an example of an artificial cue conflict: the visual information originates from the puppet, the auditory information from the performer. Another famous example is the McGurk effect (McGurk and MacDonald, 1976). When we hear the sound of someone saying baba while we play the video of the same person saying gaga, we perceive the person saying dada (Fig. 4.6). Live demos of the McGurk effect can be found online.



**Figure 4.6.** The McGurk effect.

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

At first consideration, phenomena such as the ventriloquist illusion and the McGurk effect, which commonly arise in artificial cue-conflict experiments, appear to reveal a form of *suboptimal* inference, because although the investigator has deliberately used two discrepant stimuli, the observer nonetheless incorrectly infers a single common stimulus. Another way to think about this behavior, however, is to consider that the observer is applying a prior based in natural statistics (see Chapter 3, section 3); in the real-world, when the observer simultaneously sees a ball hit the ground and hears a thud, the visual and auditory stimuli nearly always result from the same event, and therefore originate from the same location. In the laboratory experiment, the investigator has contrived a situation that would rarely occur in the world, and therefore is easily misinterpreted by the observer. Keep in mind that even when there is truly a single stimulus, the auditory and visual measurements will differ from each other on each trial because they have uncorrelated noise. Thus, the mere fact that the two cues differ does not imply that they resulted from two stimuli at different locations. The observer's inference may still be optimal, then, under a real-world prior.

Of course, the observer will only believe in a single stimulus if the discrepancies introduced are small. Otherwise, the observer will notice a conflict. For example, if the sound of bouncing ball originates at a sufficiently large distance from the visual image of the ball, the observer will realize that two separate stimuli were presented. Similarly, if a movie is poorly dubbed, the discrepancy in time between the speaker's mouth movement and voice will be too large to go unnoticed. When the observer does not necessary believe that there is a single common cause, the observer's inference process changes. This interesting situation will be discussed in Chapter 7 (causal inference).

If the observer indeed believes that there is a single common cause, then Step 2 above is unchanged. In Step 3, however, the means of  $x_A$  and  $x_V$  are no longer both  $s$ , but  $s_A$  and  $s_V$ , respectively. As a consequence, the mean MAP estimate will be

$$\langle \hat{s} \rangle = w_A s_A + w_V s_V \quad (3.5)$$

This estimator is biased. For example, the bias with respect to the true auditory stimulus is

$$\text{bias} = \langle \hat{s}_{\text{MAP}} \rangle_{p(\hat{s}|s)} - s = w_V (s_V - s_A) \quad (3.6)$$

Eq. (3.5) provides an additional prediction of the Bayesian model.

#### 4.4 Distinguishing the distributions

As in Chapter 2, it is important to distinguish between the prior, the likelihoods, the posterior, and the estimate distribution. In cue combination with a Gaussian prior with standard deviation  $\sigma_p$ , these functions are all Gaussian, but they have different means and variances, as shown in the table below. (You will derive these expressions in a problem at the end of this chapter.) It just happens to be the case that when the prior is flat ( $\sigma_p \rightarrow \infty$ ), the estimate distribution has the same variance as the posterior, but this is not the case in general.

**Table 3.1**

Distribution	Argument	Mean	Variance
Auditory noise distribution	auditory measurement, $x_A$	auditory stimulus $s_A$	$\sigma_A^2$
visual noise distribution	visual measurement, $x_V$	visual stimulus $s_V$	$\sigma_V^2$
prior distribution	hypothesized stimulus $s$	$\mu$	$\sigma_p^2$
auditory likelihood function	hypothesized stimulus $s$	measurement, $x_A$	$\sigma_A^2$
visual likelihood function	hypothesized stimulus $s$	measurement, $x_V$	$\sigma_V^2$
combined likelihood	hypothesized stimulus $s$	$\frac{x_A}{\sigma_A^2} + \frac{x_V}{\sigma_V^2}$ $\frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2}$	$\left( \frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2} \right)^{-1}$
posterior distribution	hypothesized stimulus $s$	$\frac{x_A}{\sigma_A^2} + \frac{x_V}{\sigma_V^2} + \frac{\mu}{\sigma_p^2}$ $\frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2} + \frac{1}{\sigma_p^2}$	$\left( \frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2} + \frac{1}{\sigma_p^2} \right)^{-1}$

MAP estimate distribution	stimulus estimate $\hat{s}$	$\frac{\frac{s_A}{\sigma_A^2} + \frac{s_V}{\sigma_V^2} + \frac{\mu}{\sigma_p^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2} + \frac{1}{\sigma_p^2}}$	$\frac{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2}}{\left(\frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2} + \frac{1}{\sigma_p^2}\right)^2}$
---------------------------	-----------------------------	---	---

#### 4.5 Cue combination under ambiguity

So far, we have considered cue combination under sensory noise. However, as we have seen, uncertainty sometimes arises not from sensory noise but because of inherent ambiguities in the inputs. For instance, in attempting to identify an object in the hand based on its size and weight, we may experience uncertainty not because of sensory noise (given sufficient time to obtain reliable measurements of these variables) but rather because multiple objects can have the same size and weight. Nevertheless, the logic of the inference process is the same in this scenario: the different possible objects are hypotheses (typically discrete ones: apple, orange, etc.), and the measurement of each feature (size, weight, etc.) has a certain probability under each hypothesis. As a function of the hypothesis, this is an individual-feature likelihood function. When the features are independent conditioned on object identity, the likelihood of a particular hypothesis is the product of the individual-feature likelihoods. What complicates matters is that the features do not tend to be conditionally independent. For example, even when restricted to oranges, weight and size tend to correlate strongly. Formally, this is a classification task, not an estimation task. Classification tasks under ambiguity will be discussed in Chapter 6.

#### 4.6 Tests of model predictions

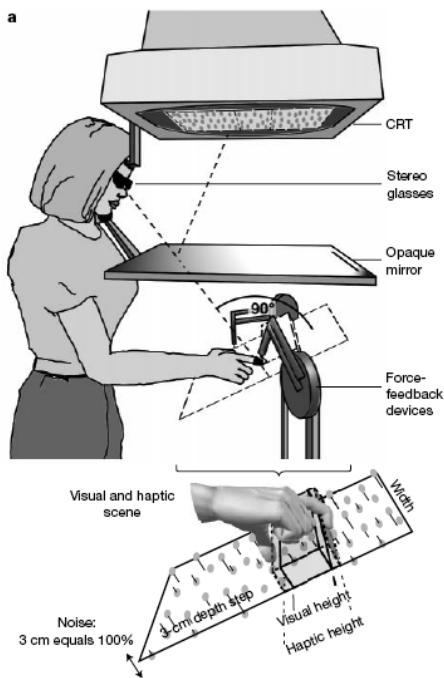
There is a long tradition of probing cue combination by analyzing how humans integrate position information from vision and audition. In many cases, vision is very precise (precision  $\sim$ min arc), while audition is relatively imprecise (precision  $\sim$ 10 degrees). This has the effect that when reliable visual information is available, people generally rely primarily on vision, a behavior predicted from Eq. (3.5).

To test the more subtle predictions of the Bayesian model, it was necessary to create situations where vision and audition are similarly precise. In a seminal study, Alais and Burr (2004) accomplished this by blurring visual inputs. The authors used several levels of blur so that visual precision would change unpredictably from trial to trial. They estimated visual precision by presenting visual stimuli alone, making use of the fact that the prior is flat so the variance of the MAP estimate for vision alone will equal the variance of the likelihood function (and the noise distribution). Similarly, they presented auditory stimuli alone to estimate the variance of the auditory likelihood function. Using the resulting estimates for  $\sigma_A$  and  $\sigma_V$ , the authors predicted the weights human subjects would place on vision and on audition when combining these cues Eq. (3.4). They found

that human behavior was well predicted by Eq. (3.5) with those weights. Moreover, the same model successfully predicted the variance of the MAP estimate.

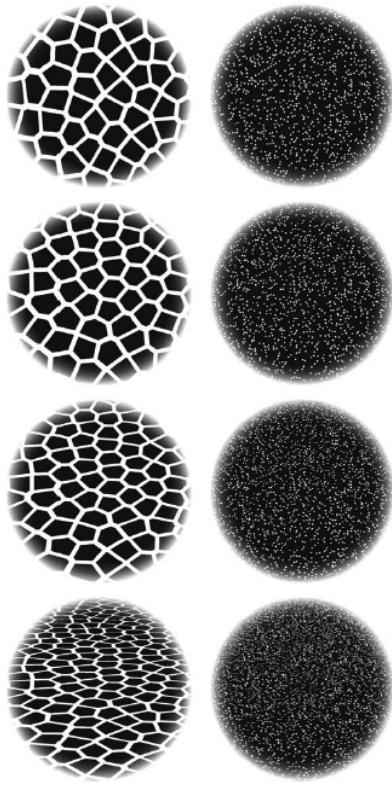
An important technical detail in many cue combination studies is that the experiments usually do not ask for estimates on a continuum (as in Fig. 4.1) but instead use a so-called two-alternative forced choice paradigm, in which the subject is presented with stimuli in two intervals and required to make a choice between them. For example, subjects are presented with two sets of auditory-visual stimuli, and asked in which of the two the auditory stimulus was more to the left. This allows the investigator to estimate the variances (precisions) of the cues in a way that is unaffected by the subject's prior. We examine the details of this procedure in Chapters 5 and 6.

The Bayesian model of cue combination has been tested in other sensory modalities as well. A classical study is that of Ernst and Banks (Fig. 4.7). The authors studied subjects' estimation of the size of an object that could be both seen and felt. Under normal viewing conditions, vision is often more precise than touch; the authors blurred the visual feedback in order to reduce its precision. They found that across different visual feedback precisions, the weight subjects placed on vision was very close to the value predicted by Eq. (3.4).



**Figure 4.7.** Experimental set-up in the study by Ernst and Banks (Nature, 2002) on combination of visual and tactile cues. Subjects viewed a virtual-reality image and received pressure from a force-feedback device to simulate a visual and a tactile stimulus.

Many experiments have probed the integration of two cues originating from a single sensory modality (for example, Jacobs 1999; Knill and Saunders 2003). One example is the estimation of slant (orientation of a plane) based on visual texture and visual disparity. Texture provides information about slant (Fig. 4.8). Disparity also provides information about slant; the part of the plane that is closer to the viewer will have a smaller binocular disparity (Fig. 4.8).



**Fig. 4.8.** Estimating surface slant using texture and stereo cues. Stimuli used in Knill and Saunders (2003). Surface slant was defined using texture (left), stereo (random dot patterns, as right), or both. The four rows shown correspond to 0°, 30°, 50°, and 70° slant. The random dot patterns have to be presented stereoscopically for the observer to perceive slant.

In a typical study, subjects would be shown a surface with a texture that indicated a given slant. The texture information can be made more or less informative. For example, circles provide a highly informative cue, whereas random white noise provides a very uninformative cue. The disparity cue can also be manipulated, and importantly changed independently of texture. By independently varying the texture and disparity cues, the authors of such studies have generally found that subjects integrate these cues in accordance with the predictions of the Bayesian model. For example, as the texture cue is varied to indicate different slants, it exerts a roughly linear influence on the estimated slant. The slope of that influence, the weight on texture (see Eq. (3.5)), fits well with the prediction of the Bayesian model.

When we want to estimate the position of our hand in a two-dimensional plane, such as a tabletop, we have to solve a two-dimensional estimation problem. To make this estimate, we can use proprioceptors that signal body posture. We can also use vision. The

proprioceptive and visual cues to hand position have different properties. Proprioception is generally noisy, but good at estimating changes in the direction of our smaller joints. Vision is quite good in terms of direction but rather poor at estimating depth. In a seminal study, van Beers and collaborators (1996) probed how the nervous system combines visual and proprioceptive cues in this task. It was found that the cue combination proceeds almost exactly as predicted by the Bayesian model.

Several authors have also studied speech perception from the point of view of the Bayesian model, both in McGurk-like settings with simple syllables, and with monosyllabic words (e.g. Bejjanki et al. 2011; Ma et al. 2009).

#### 4.7 Cue combination for technical problems: Naïve Bayes

There are many technically relevant cue combination problems in a field called machine learning. A nice example of machine learning is Spam Filtering. There are certain words that are correlated with an email being spam. For example, real emails addressed at me do not tend to contain the word Viagra very often. We are given a database of emails, one of which are emails known to be genuine, and another database of emails which are known to be spam. How can we estimate if a new email is genuine? Let us say we only want to use the counts of each possible word for this process. What we really want to do is to estimate  $p(\text{spam} | \text{word counts})$ .

We can assume that, given knowledge of emails being genuine or spam that word counts are independent from one another and that there is Gaussian noise on the word counts. This assumption is what is called *naïve Bayes*: the approach is naïve in the sense that it's independence assumptions are generally not met. For example, the word "Bayesian" is correlated with the word "statistics" in my emails.

For each word we can calculate the mean count for spam and genuine emails and also the associated standard deviation. We thus know for word  $i$ :

$$p(\text{word}_i | \text{spam}) \propto e^{(\text{word}_i - \mu_{\text{spam}})^2 / (2\sigma_{\text{spam}}^2)}$$

and for the overall probability of spam:

$$p(\text{spam} | \text{words}) \propto \frac{1}{Z} p(\text{spam}) \prod \frac{1}{\sigma_{\text{spam},i}} e^{(\text{word}_i - \mu_{\text{spam},i})^2 / (2\sigma_{\text{spam},i}^2)}$$

This same approach of assuming that all cues are independent, even when they are not used in many domains of machine learning. Naïve Bayes is often used to solve real classification problems and is, for certain problems a competitive machine learning technique. It is particularly strong when there is very little available data.

#### 4.8 Concluding remarks

We have seen that cue combination is a frequent and important perceptual activity that often happens automatically and outside of our conscious control. A simple Bayesian model can explain how humans combine cues in a wide variety of settings. Unlike the

winner-take-all strategy, the optimal Bayesian solution, which is followed by humans in many instances, is to weight each cue according to its reliability.

We will now discuss a few directions in which the Bayesian model could be extended. Most cues are directly related to one of the variables that we care about. However, there may be some cues that only obtain their meaning through other cues. For example, shadows are not normal cues. If we do not know the direction of the sun then shadows will not actually help us estimate the size of an object. These cues are called pseudo-cues. Recent studies have probed how subjects make use of these non-standard cues.

Cue combination can also take place over time, in which case it is sometimes called evidence accumulation, evidence integration, or decision-making. Suppose you are outdoors and want to determine the wind direction. You stick a finger in the air to obtain a measurement. You could do this not once, but multiple times. (We assume that the wind direction doesn't change in the meantime.) Each measurement comes with an elementary likelihood function. These likelihoods will not be equally wide, because each measurement comes with its own particular amount of noise. For instance, if the wind weakens momentarily, you will have more noise in your measurement than when the wind is strong. As in auditory-visual cue combination, the MAP estimate is a linear combination of the individual measurements, weighted by their precisions. We will work this case out in a problem.

Sometimes, cues are not combined in a linear fashion. In certain cases, it can be shown that nonlinear cue combination is optimal. We will discuss such situations in Chapter 7. Finally, in Chapter 12, we will discuss how optimal cue combination can be realized in neural populations.

## 4.9 Further reading

- Ernst, M. O. and M. S. Banks (2002). "Humans integrate visual and haptic information in a statistically optimal fashion." *Nature* 415(6870): 429-433.
- Alais, D. and D. Burr (2004). "The ventriloquist effect results from near-optimal bimodal integration." *Curr Biol* 14(3): 257-262.
- Yuille, A. L. and H. H. Bulthoff (1996). Bayesian decision theory and psychophysics. *Perception as Bayesian Inference*. D. C. Knill and W. Richards. New York: Cambridge, University Press.
- Jacobs, R.A. (1999) Optimal integration of texture and motion cues to depth. *Vision Res* 39, 3621-3629.
- Knill, D. C. and J. A. Saunders (2003). "Do humans optimally integrate stereo and texture information for judgments of surface slant?" *Vision Research* 43(24): 2539-2558.
- Brouwer, A.-M. and D. C. Knill (2007). "The role of memory in visually guided reaching." *Journal of Vision* 7(5): 1-12.

- van Beers, R. J., A. C. Sittig, et al. (1996). "How humans combine simultaneous proprioceptive and visual position information." *Exp Brain Res* 111(2): 253-261.
- Bejjanki V.R., Clayards M., Knill D.C., Aslin R.N. (2011). Cue integration in categorical tasks: insights from audio-visual speech perception. *PLoS ONE* 6(5), e19812
- Ma, W.J., Zhou, X., Ross, L.A., Foxe, J.J. & Parra, L.C. (2009) Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. *PLoS ONE* 4, e4638.
- Wallace, MT, Roberson, G.E, Hairston, W. D., Stein, B. E., Vaughan, J. W., Schirillo, J. A. (2004). *Unifying multisensory signals across time and space*. *Exp Brain Res* 158(2): 252-258.
- McGurk, H. and MacDonald, J. (1976) Hearing lips and seeing voices. *Nature* 264: 746–748

#### 4.10 Problems

**Problem 3.1.** An observer combines cues  $A$  and  $B$ . When  $B$  becomes more reliable, the observer's estimate will

- a) shift towards A
- b) shift towards B
- c) stay the same
- d) Insufficient information

**Problem 3.2.** True or false?

- In modeling speech recognition, the sound and the image can be assumed to be independent of each other.
- Conflicts between two measurements generated by a single source do not normally occur in the natural world.

**Problem 3.3.** Suppose  $p_1(x), p_2(x), \dots, p_N(x)$  are Gaussian functions in  $x$  given by

$$p_i(x) = \exp\left[-\frac{(x - \mu_i)^2}{2\sigma_i^2}\right] \text{ for } i \in \{1, 2, \dots, N\}$$

If  $q(x) \propto p_1(x)p_2(x)\dots p_N(x)$  is a probability distribution, show that  $q(x)$  is a Gaussian distribution with mean and variance

$$\mu_q = \frac{\sum_{i=1}^N \frac{\mu_i}{\sigma_i^2}}{\sum_{i=1}^N \frac{1}{\sigma_i^2}}$$

$$\sigma_q^2 = \frac{1}{\sum_{i=1}^N \frac{1}{\sigma_i^2}}$$

respectively. (**Hint 1:** Show by induction; for  $N=1$ , the result is trivial. Assume it is true for some  $N = k$ , and show that then it would also be true for  $N = k+1$ . Alternatively, direct computation is also possible. **Hint 2:** Leave  $1/\sigma_i^2$  factors as such everywhere, don't rewrite in terms of  $\sigma_i^2$ .)

**Problem 3.4.** An observer infers a stimulus  $s$  from conditionally independent measurements  $x$  and  $y$ . The noise distributions  $p(x|s)$  and  $p(y|s)$  are Gaussian distributions with variances  $\sigma_x^2$  and  $\sigma_y^2$ , respectively. The stimulus distribution is Gaussian with mean  $\mu$  and variance  $\sigma_s^2$ . Compute the posterior distribution and the distribution of the MAP estimate. You can use the equations given in Problem 3.3 and verify your answers in Table 3.1.

**Problem 3.5.** Organisms accumulate information over time. This is an important factor to consider whenever sensory information is available for hundreds of milliseconds or more. We now model the process of combining different measurements over time. An observer infers a stimulus  $s$  from a sequence of conditionally independent measurements  $x_1, x_2, \dots, x_N$  (the index refers to time points on a single trial). The conditional distribution of the  $i^{\text{th}}$  measurement,  $p(x_i|s)$ , is Gaussian with mean  $s$  and variance  $\sigma_i^2$  (the variance of the measurements could be different). The stimulus distribution is Gaussian with mean  $\mu$  and variance  $\sigma_s^2$ .

- What are the mean and variance of the posterior? You can use the equations given in Problem 3.3.
- For a given true stimulus value  $s$ , we define *relative bias* as the difference between the mean MAP estimate of  $s$  (mean over many trials, not over  $i$ ) and  $s$  itself, divided by the difference between  $\mu$  and  $s$ . Calculate relative bias.
- Compute the variance of the MAP estimate when the true stimulus value is  $s$ .
- If all noise variances are identical, that is  $\sigma_i = \sigma$  for all  $i$ , simplify and interpret the expression for relative bias. (Interpretation means explaining how the dependencies on the variables make sense.)

**Problem 3.6\***

In this problem, we examine *suboptimal* estimation in the context of cue combination. Suppose an observer estimates a stimulus  $s$  from two conditionally independent, Gaussian-distributed measurements,  $x_A$  and  $x_V$ .

- a) What is the expression for the MAP estimate in terms of the measurements? What is the variance of this MAP estimate? The prior is flat.
- b) Now suppose the observer uses an estimator of the form  $\hat{s} = wx_A + (1-w)x_V$ . Show that this estimate is unbiased (just like the MAP estimate); this means that the mean (expected value) of the estimate is equal to  $s$ .
- c) What is the variance of this estimate as a function of  $w$ ? Plot this function. At which value of  $w$  is it minimal, and does this value make sense? State your final conclusion in words.
- d) Which of the conclusions in (b) and (c) break down when we consider estimates that are general linear combinations of measurements,  $\hat{s} = w_A x_A + w_V x_V$ ? Explain.
- e) What quantity does the MAP estimate minimize in this more general setting?

## LAB PROBLEMS

### Problem 3.7.

In Chapters 2 and 3, we were able to derive analytical expressions for the posterior distribution. For more complex psychophysical tasks, however, analytical solutions often do not exist. In such a case, we can use numerical methods to approximate the distribution of interest. To get some familiarity with this method, we will reconsider the cue combination experiment described in this chapter, but we will now compute the distribution of MAP estimates using numerical methods. We assume that the experimenter introduces a cue conflict between the auditory and the visual stimulus:  $s_A=5$  and  $s_V=10$ . The standard deviation of the auditory and of the visual noise is  $\sigma_A=2$  and  $\sigma_V=1$ , respectively. We assume a flat prior over  $s$ .

- a) Randomly draw an auditory measurement  $x_A$  and a visual measurement  $x_V$  from their respective distributions.
- b) Plot the corresponding elementary likelihood functions,  $p(x_A|s)$  and  $p(x_V|s)$ , in one figure.
- c) Calculate the combined likelihood function,  $p(x_A, x_V|s)$ , by numerically multiplying the elementary likelihood functions in Matlab. Plot this function.
- d) Calculate the posterior distribution by normalizing the combined likelihood function. Plot this distribution in the same figure as the likelihood functions.
- e) Use Matlab to find the MAP estimate of  $s$ , i.e. the value of  $s$  at which the posterior distribution is maximal.
- f) Compare with the MAP estimate of  $s$  computed from Eq. (3.3) using the measurements drawn in (a).

- g) In the above, we simulated a single trial and computed the observer's MAP estimate of  $s$ , given the noisy measurements on that trial. If an analytical solution does not exist for the distribution of the MAP estimates, we can repeat the above procedure many times to approximate this distribution. Here, we practice this method even though an analytical solution is available in this case. Draw 100 pairs  $(x_A, x_V)$  and numerically compute the observer's MAP estimate for each pair as in (e).
- h) Compute the mean of the MAP estimates obtained in (g) and compare with the mean estimate predicted using Eq. (3.5).
- i) Make a histogram of the MAP estimate (use the "hist" function).
- j) *Relative auditory bias* is defined as the mean MAP estimate minus the true auditory stimulus, divided by the true visual stimulus minus the true auditory stimulus. Compute relative auditory bias for your set of estimates.

### Problem 3.8.

Lets say we are the NSA and we are reading the emails of all Americans. Some Americans are dangerous and are critical of the government. As you, unfortunately, ran out of agents to read all the emails, you need to write software to estimate who is good and who is critical. Based on the prior work by agents, you do have a dataset of emails of critical Americans and a dataset of good Americans (download datasetCritical).

You want to build a naïve Bayes system that estimates if a given email is critical using only the counts of each word.

- a) Calculate for each word the average frequency for normal and critical americans and also the aossociated standard deviations.
- b) Do both groups of citizens have the same average word frequencies? Which of the words exhibit significant differences between the two groups?
- c) What would be a good measure for the informativeness of a word? Which word shows the strongest difference?
- d) If you estimated the group based on just this word, how good would you do on average?
- e) If you combine data from all words, using a naïve Bayes approach how good can you be at the problem?
- f) Is this a difficult problem? Could it have real-world relevance? Can you think of an application of naïve Bayes decoding that is more exciting?

## Contents

4 Chapter 4: Binary estimates .....	4-2
4.1 Two typical binary estimation tasks .....	4-3
4.1.1 Yes-no task.....	4-3
4.1.2 Two-alternative forced-choice paradigm.....	4-4
4.2 Yes/no discrimination: Bayesian model .....	4-5
4.2.1 Generative model .....	4-6
4.2.2 Inference model .....	4-6
4.2.3 Distribution of the MAP estimate .....	4-13
4.3 Yes/no classification .....	4-15
4.3.1 Generative model .....	4-15
4.3.2 Inference model .....	4-17
4.3.3 Distribution of the MAP estimate .....	4-20
4.4 Detection .....	4-21
4.4.1 Hit and false-alarm rates .....	4-22
4.4.2 Confidence .....	4-23
4.4.3 Receiver operating characteristic .....	4-25
4.4.4 Bias .....	4-27
4.4.5 Sensitivity .....	4-29
4.4.6 Applications .....	4-30
4.5 Where does the optimal criterion lie? .....	4-31
4.6 Concluding remarks .....	4-31
4.7 Further reading.....	4-33
4.8 Problems .....	4-34

## 4 Chapter 4: Binary estimates

*How can we make decisions between two possibilities?*

Binary decisions estimation tasks are central to perception research. In Chapters 2 to 4, we discussed the basics of Bayesian modeling, often using the example of a spatial localization task. In such a task, the variable of interest, e.g. the location on a line, takes on a continuum of values. The task was to estimate the location on this continuum. In such a task, the subject has in principle an infinite number of possible responses.

Many if not most tasks in the lab ask for a choice between only two alternatives. This is called a binary choice or a binary decision task. Binary choices and is based on binary estimates. Such tasks are also common in the real world, for example: will it rain today, can I trust this person, can I make it to the bus stop in time when I run, is this email spam or not? Each of these questions has a yes/no answer, and the corresponding random variable (whether it will rain today, etc.) is therefore binary. Many examples we encountered in Chapter 1, such as determining whether a bag on the baggage carousel is yours or not, also featured binary estimations.

Two types of binary estimates tasks are particularly important. One class of tasks deal with *yes/no tasks* where a subject is shown a single stimulus and is asked if it has a certain property. As an example of a detection task, imagine you are a radiologist trying to determine whether a tumor is present on a noisy X-ray image. A second class of tasks deal with *comparisons* where a subject is asked to estimate if one stimulus is, on some axis, larger or smaller than another stimulus. Usually this is done in a way called *discrimination tasks*. Detection and discrimination two-alternative-forced-choice (2AFC) where the subject can only choose one of the comparison outcomes. These two approaches characterize most popular psychophysical paradigms.

These two tasks appear in many laboratory experiments: was the motion to the left or to the right was a vertical line present in the display, did you feel a stimulus on your finger, are you touched more frequently on one finger or another. Even when the underlying stimulus variable is continuous (e.g. duration), it is common to phrase the task in terms of a choice between two options (e.g. which stimulus lasted longer), while continuously manipulating the stimuli from trial to trial. As we will see below, they allow these laboratory experiments to be interpreted in a straightforward way.

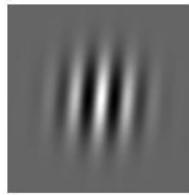
Modeling binary estimation, is arguably easier and more straightforward than continuous tasks. Prior, posterior, and the estimate distribution (but not the likelihood function) are characterized by a single number, since the probabilities of both possible world states have to sum up to 1. This facilitates plotting, analysis and quantification and allows a particularly meaningful characterization of the decision process.

Plan of the chapter: This chapter is structured around the same three-step process as Chapters 2 and 4: the generative model, the inference process, and the estimate distribution. The chapter also introduces special approaches for binary decisions. We will use the same basic task

as in Chapter 2, combining a measurement with a prior. A note on what we will *not* discuss in this chapter: Many binary choices involve not only perceptual information, but also considerations of cost and reward. Shall I go by car or by bus? Shall I buy my vegetables here or check out another store first? Should I propose to my sweetheart now or later? In these situations there is not a universally correct answer, because the “best possible” choice depends on cost functions other than just the correctness of the response. We consider such reward-based choices in Chapter 8.

#### 4.1 Two typical binary estimation tasks

Before we delve into the derivations of optimal behavior for the different kinds of behavioral situations we want to first discuss two examples of rather typical experiments. One of the yes-no flavor and one of the comparison (2AFC) flavor.

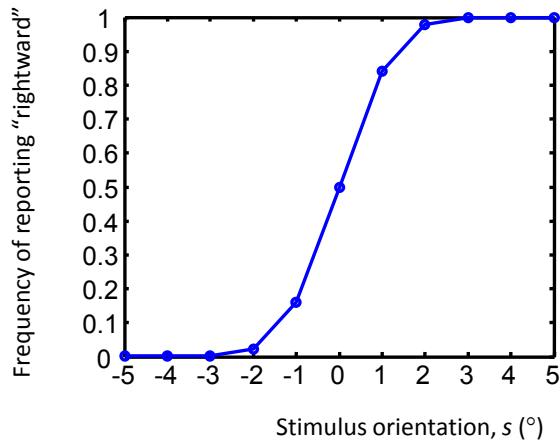


**Figure 5.2.** Yes/no orientation discrimination. Is the stimulus tilted to the right or to the left with respect to the vertical?

##### 4.1.1 Yes-no task

Imagine you are shown for just 10 ms an oriented pattern (as in Fig. 5.2) with an orientation that is drawn from a Gaussian distribution with mean  $\mu$  (for example, vertical) and standard deviation  $\sigma_s$ . Instead of reporting the stimulus on a continuum, as in Chapter 2, you have to decide whether the orientation was to the right (clockwise) or left (counterclockwise) of  $\mu$ . You report for example by pressing one of two keys. In the present discrimination task example, the observer is in essence responding “yes” or “no” to the question: was this stimulus rightward of  $\mu$ ?

It stands to reason that the more rightward the stimulus is, the greater the probability will be that the observer responds “rightward” (i.e., “yes”). The psychometric curve plots the proportion of trials that the observer reports “rightward” as a function of stimulus orientation. Psychometric curves have been measured in many experiments, and they generally look sigmoidal, like the curve shown in Fig. 5.8, with y-axis values extending from 0 (for far-left stimulus presentations) to 1 (far-right presentations), passing through approximately 0.5 when the stimulus is equal to  $\mu$ . In virtually all experiments the psychometric curve of an optimal observer looks like this.



**Figure 5.3.** Example of a psychometric curve in a yes-no paradigm

Why would human behavior show such a behavior? Intuitively, the further to the right the stimulus is the more obvious it gets that it is to the right and vice versa. Unavoidable noise in perception is converting a response curve that would be perfect (switching at zero) into a somewhat smooth sigmoidal shape.

Exercise 4.1: What do you think determines the steepness of the sigmoid around the center?

When people plot psychometric functions they generally boil them down to two relevant measures. The first is the Point of Subjective equality (PSE), the point where the probability is equal to choose left or right. The second is the Just noticeable difference (JND), the range that it takes the curve to get from 25% probability to 75% probability. These two variables characterize bias in choice (PSE) and how noisy the choice is (JND). We will see more below about how to interpret these variables.

#### 4.1.2 Two-alternative forced-choice paradigm

In the yes-no task, the observer was asked to judge the stimulus against a reference orientation,  $\mu$ , that the observer was assumed to know accurately from experience. While convenient in some cases – humans have a rather accurate internal representation (memory) of what a vertical orientation looks like – this procedure is impractical in others. For example, a subject in a psychophysics experiment might have difficulty judging whether an object is farther away than a specified reference distance (e.g., 15 m), because the subject does not have an accurate internal representation of the reference. Similarly, a subject might be unable to judge accurately whether a stimulus lasts longer than a specified reference duration (e.g., 0.25 s) for which he has only a vague internal representation.

A psychophysical paradigm that solves this conundrum is to compare two stimuli against each other. This is called the two-alternative forced-choice (2AFC) paradigm. This terminology can be confusing because every binary decision has two alternatives, including the ones discussed so far. However, the terminology “two-alternative forced-choice” is used to refer specifically to a task in which two stimuli are presented on a given trial and the observer has to *compare* them to each other. When the two alternatives are presented sequentially, the paradigm is also called *two-interval forced choice* or 2IFC. To avoid confusion, some researchers avoid the term 2AFC altogether and use 2IFC instead.

We now consider the 2AFC version of the experiment in Section 4.1.1. An observer views two orientations,  $s_1$  and  $s_2$ , each drawn from the same Gaussian stimulus distribution  $p(s)$ . Measurements are  $x_1$  and  $x_2$ , drawn from Gaussian distributions with means  $s_1$  and  $s_2$ , and the same variance  $\sigma^2$ . The observer reports which of the two stimuli was more clockwise (in our nomenclature, “greater”). In contrast to the task in the previous section, this is now a relative rather than an absolute judgment. However, the logic is the same. The observer now judges whether  $s_1$  is greater or smaller than  $s_2$ . The MAP estimates of  $s_1$  and  $s_2$  are

the observer will stay behave the same way as the message

As in the yes-no case, in a 2AFC paradigm, the point where the psychometric curve crosses 0.5 is also called the *point of subjective equality* (PSE), because when the difference between the stimuli has that magnitude, the observer perceived  $s_1$  and  $s_2$  to be equal. In the task discussed here, the PSE is equal to 0, but we will see later an example where it is different from 0.

Another common version of the 2AFC paradigm goes as follows: on each trial, you are shown two stimuli, say at two locations on the screen. One of them is to the right and one to the left of vertical. Your task is to report which of the two is further to the right. This task differs from the one above in that the stimuli always have opposite sign. All else being equal, performance will be higher than in the one above. We will examine this task in the Problems.

## 4.2 Yes/no discrimination: Bayesian model

In the previous section, we considered binary versions of the continuous estimation task in Chapter 2. You may have noticed, however, that we did not follow the same structure of Bayesian modeling as in previous chapters. Instead, we used the MAP estimates from the continuous estimation task in Chapter 2, and then postulated the optimal decision rule in the binary task to be  $\hat{s} > \mu$  (in the yes/no task) or  $\hat{s}_1 > \hat{s}_2$  (in the 2AFC task). While correct, these rules were not obtained using a systematic procedure of maximizing the posterior distribution over the world state of interest. The only posteriors we considered were those over the continuous variable  $s$  (in the yes/no task) and over  $s_1$  and  $s_2$  (in the 2AFC task). Those variables were the world states of interest in Chapter 2, but in the binary tasks, the world state of interest is the *sign of  $s - \mu$*  (in the yes/no task) or the *sign of  $s_1 - s_2$*  (in the 2AFC task). To derive the optimal

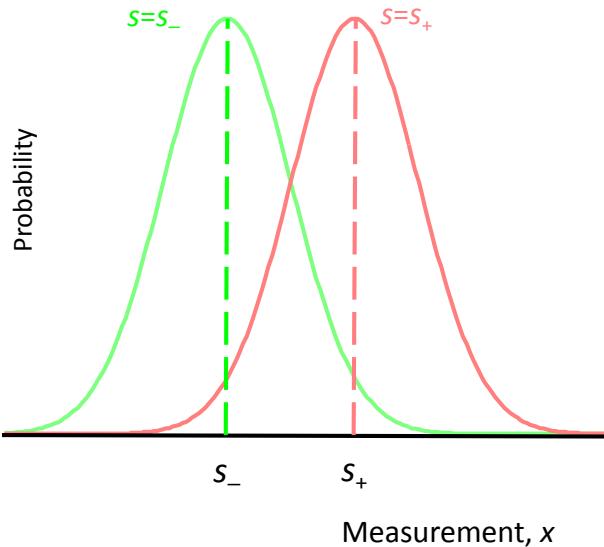
decision rule systematically we will have to compute and maximize the posterior over these signs. In fact, from the point of view of the generative model, the sign is the world state variable to start with, and the stimulus follows a certain distribution once the sign is given. This formal set-up, which we now work out, will allow us to examine the effect of a non-flat prior over the world state variable, for example a larger frequency of rightward than leftward stimuli in the yes/no task. Having a posterior distribution over the world state of interest is also important if knowledge of the world state is used for subsequent computation, for example to compute expected cost of an action.

A proper Bayesian treatment of the yes/no task in Section 4.1.1 is by no means trivial. We will start with a simpler version of the task, in which a stimulus  $s$  can take just two values, which we call  $s_+$  and  $s_-$ , and the observer chooses between them. For example, the observer reports whether an oriented pattern is tilted  $1^\circ$  to the right or  $1^\circ$  to the left of vertical. The difference with the task in Section 4.1.1 is that the choice is between two specific stimulus values rather than classes of stimulus values (all rightward versus all leftward tilted). We now go one by one through the steps in the Bayesian model.

#### 4.2.1 Generative model

The generative model is  $s \rightarrow x$ , as in Chapter 2. The stimulus  $s$  follows a world state distribution  $p(s)$ . The world state distribution is a discrete probability distribution, with values  $p(s=s_+)$  and  $p(s=s_-)$ , which sum to 1 and reflect the frequencies with which the stimulus values are presented. The measurement  $x$  follows the usual Gaussian noise distribution  $p(x|s)$  with mean  $s$  and standard deviation  $\sigma$ . The measurement distribution is shown in Fig. 5.5 for both possible values of  $s$ .

#### 4.2.2 Inference model



**Figure 5.5.** Noise distributions for the yes/no task of discriminating between  $s_-$  and  $s_+$ .

Suppose  $s_+=1^\circ$  and  $s_-=-1^\circ$ , and that on a given trial, your measurement is  $+0.1^\circ$ . Would you report that the stimulus was  $s_+$  or  $s_-$ ? You probably would say  $s_+$  simply because the measurement is closer to  $s_+$  than to  $s_-$ . But now imagine that you knew that  $s_-$  was far more probable a priori than  $s_+$ . In that case, a measurement that is only slightly closer to  $s_+$  would likely have been produced by  $s_-$ . In this subsection, we will work out how the optimal observer would decide.

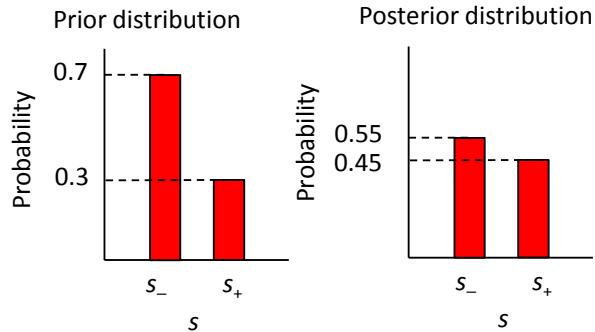
Describing the optimal observer requires calculating the posterior over  $s$ ,  $p(s|x)$ . Since  $s$  takes on two values, the posterior distribution is a discrete probability distribution, with values  $p(s=s_+|x)$  and  $p(s=s_-|x)$ , which have to sum to 1. Bayes' rule tells us that

$$p(s|x) = \frac{p(x|s)p(s)}{p(x)}.$$

For a binary variable, the posterior distribution is uniquely determined by the ratio of the posterior probabilities of the two alternatives. We can calculate this ratio using Bayes' rule:

$$\frac{p(s=s_+|x)}{p(s=s_-|x)} = \frac{\frac{p(x|s=s_+)p(s=s_+)}{p(x)}}{\frac{p(x|s=s_-)p(s=s_-)}{p(x)}} = \frac{p(x|s=s_+)p(s=s_+)}{p(x|s=s_-)p(s=s_-)}. \quad (5.1)$$

This ratio is called the *posterior ratio* or *posterior odds*. Its interpretation is that of the probability of one alternative *relative to* that of the other. We see that the normalization  $p(x)$  drops out and is thus irrelevant to this ratio. For example, if your posterior probability that the stimulus was  $s_+$  is 80%, then your posterior ratio is  $0.80/0.20 = 4$ . The posterior ratio is always positive but can grow arbitrarily large. For example, if the posterior probability of  $s_+$  is 99%, then the posterior ratio is  $0.99/0.01 = 99$ . Knowing the posterior ratio one can calculate the probability of each of the alternatives and vice versa.



**Figure 5.6.** Example prior and posterior distribution over a binary variable.

**Exercise 5.6:** Suppose the prior distribution and the posterior distribution are as in Fig. 5.6. Calculate the likelihood ratio. Does the sensory evidence indicate that the stimulus was  $s_+$  or  $s_-$ ?

You will often see Eq. (5.1) with the natural logarithm taken of both sides ( $\equiv$  means “is defined as”):

$$d(x) \equiv \log \frac{p(s = s_+ | x)}{p(s = s_- | x)} = \log \frac{p(x | s = s_+) p(s = s_+)}{p(x | s = s_-) p(s = s_-)}. \quad (5.2)$$

This simplifies a good number of mathematical derivations, as we will see below. The quantity  $d(x)$  is called the *log posterior ratio* (also: *log posterior odds*). If the posterior probability of  $s_+$  is 80%, then the log posterior ratio is  $\log(0.80/0.20) = \log(4) = 1.39$ . The log posterior ratio contains the same information as the posterior distribution itself; after all, we can exponentiate it and calculate the probability of each alternative from it as we did above.

**Exercise 5.7:** In Matlab, create a vector of 99 possible posterior probabilities of  $s_+$ , from 0.01 to 0.99 in steps of 0.01. For each value, calculate the log posterior ratio  $d$ . Then plot the log posterior ratio as a function of the posterior probability of  $s_+$ . This should show that every posterior probability corresponds to exactly one log posterior ratio and the other way round, so that knowing one is as good as knowing the other.

**Exercise 5.8:** In the previous exercise, why did we not include the posterior probabilities 0 and 1?

Exercise 5.9: This exercise goes in the other direction than Exercise 2. Suppose you know the log posterior ratio  $d$ . Express the posterior probability of  $s=s_+$ ,  $p(s=s_+|x)$ , as a function of the log posterior ratio  $d$  only. Do the same for  $p(s=s_-|x)$ .

The log posterior ratio is symmetric in the following sense: flipping the posterior probabilities of the two alternatives is equivalent to flipping the sign of the log posterior ratio. For example, if the posterior probability of  $s_+$  is 20% and that of  $s_-$  80%, then the log posterior ratio is  $\log(0.20/0.80)=\log(0.25)=-1.39$ . When two alternatives have the same posterior probability, the log posterior ratio is zero. If the log posterior ratio is positive then  $p(s=s_+|x)$  is larger than  $p(s=s_-|x)$ . Therefore, the MAP estimate is  $\hat{s}=s_+$  when  $d(x)>0$ , and  $\hat{s}=s_-$  when  $d(x)<0$ . Thus, binary decision-making is concisely described in terms of log posterior ratios.

Important: For binary decisions, the MAP estimate is determined by the *sign of the log posterior ratio*.

Exercise 5.10: Why does the case  $d(x)=0$  usually not have to be considered? What would the observer do when  $d(x)=0$ ?

We now point out some common terminology. The inequality used to determine the MAP estimate,  $d(x)>0$ , is also called the *decision rule* of the Bayesian MAP observer, and  $d(x)$  is called the *decision variable* (hence the symbol  $d$ ). The scalar value to which the decision variable is compared in order to make a decision, here 0, is also called the *decision criterion*, or simply the *criterion*. The terminology of decision rule, decision variable, and decision criterion is not specific to Bayesian models. Any inequality of the form  $f(x)>k$ , with  $f$  a function and  $k$  a constant, can serve as a model for how the observer turns a measurement into a decision.

Log posterior ratios also simplify other aspects of the derivations. Since the logarithm of a product is the sum of the logarithms, the right-hand side of Eq. (5.2) can be rewritten as a sum:

$$d(x) \equiv \log \underbrace{\frac{p(s=s_+|x)}{p(s=s_-|x)}}_{\text{Posterior ratio}} = \log \underbrace{\frac{p(x|s=s_+)}{p(x|s=s_-)}}_{\text{Likelihood ratio}} + \log \underbrace{\frac{p(s=s_+)}{p(s=s_-)}}_{\text{Prior ratio}}. \quad (5.3)$$

Each of these terms has an intuitive meaning. The first term on the right-hand side is called the *log likelihood ratio* and reflects the amount of evidence provided by the measurement  $x$ . The second term on the right-hand side is the log prior ratio and reflects our relative prior beliefs in the two alternatives. The sum of these terms is the log posterior ratio.

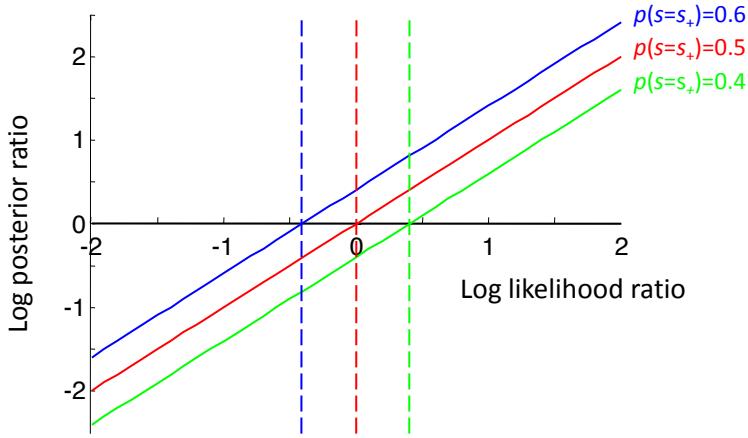
Whenever  $d(x)$  is greater than zero,  $s_+$  is most probable. The MAP decision rule is thus:

$$\log \frac{p(x|s=s_+)}{p(x|s=s_-)} > -\log \frac{p(s=s_+)}{p(s=s_-)}. \quad (5.4)$$

This expresses that the log likelihood ratio must be greater than the negative log prior ratio for the MAP observer to report  $\hat{s} = s_+$ . We thus have a compact description for the optimal decision rule.

Exercise 5.11: Before reading on, can you give an intuition for Eq. (5.4)?

Thus, when  $s_-$  is expected to occur with higher probability than  $s_+$  (right hand term is positive), the optimal decision rule is to report  $\hat{s} = s_+$  only when the measurement  $x$  provides sufficiently strong evidence in favor of  $s_+$  to overcome the prior bias in favor of  $s_-$ . Imagine you are standing in front of a classroom of people, and you are asked whether or not a given person in the audience has blue eyes ( $s_+$ ). The quality of information is affected by your distance to the person. Suppose that in your region of the world, brown eyes ( $s_-$ ) are more common than eyes of other colors. If you are asked about someone nearby, your sensory information will be of high quality and you will be able to base your decision purely on this sensory information. If you are asked about someone farther away, the quality of the sensory information is worse or even uninformative. The lower the quality of the visual information, the more you will rely upon your knowledge of the prevalence of brown eyes in the general population. When no information is available at all, your best bet is to always respond that the person has brown eyes. This increasing effect of the prior as the quality of sensory information decreases is exactly expressed by Eq. (5.4). The term on the left-hand side – the log-likelihood ratio – will tend to be smaller in magnitude (either positive or negative) when the sensory information is of lower quality (“tend to” because this term is a random variable that inherits its distribution from the distribution of  $x$ ). The relative quality of prior and likelihood information determines which information dominates.



**Figure 5.7.** The log posterior ratio is the sum of the log likelihood ratio and the log prior ratio. As the prior more strongly favors  $s_+$  (going from green to red to blue), the decision criterion on the log likelihood ratio shifts towards smaller values, indicating that the subject has a stronger tendency to report  $s_+$ .

The probability of each stimulus based on priors, likelihoods and posteriors can be visualized (Fig. 5.7). If the prior distribution were uniform, i.e.  $p(s=s_+)=p(s=s_-)=0.5$  (red line), then the log prior ratio would be zero, and the decision would be determined by whether the measurement,  $x$ , is more probable when  $s=s_+$  or when  $s=s_-$ . If the prior favors  $s=s_+$ , for example  $p(s=s_+)=0.6$  (blue line), then the negative log prior ratio would be a negative number ( $-0.405$ ). As a consequence, the measurement  $x$  could produce a negative log likelihood ratio and yet the observer would respond  $\hat{s}=s_+$ . In other words, the prior *biases* the observer towards reporting  $\hat{s}=s_+$ . The opposite happens when the prior favors  $s=s_-$  (green line). The bias seen here is similar to how a Gaussian prior biases the observer towards its mean in the continuous estimation task discussed in Chapters 2 and 3 (Section 3.5). In general, priors and likelihoods can be equally important for any given task.

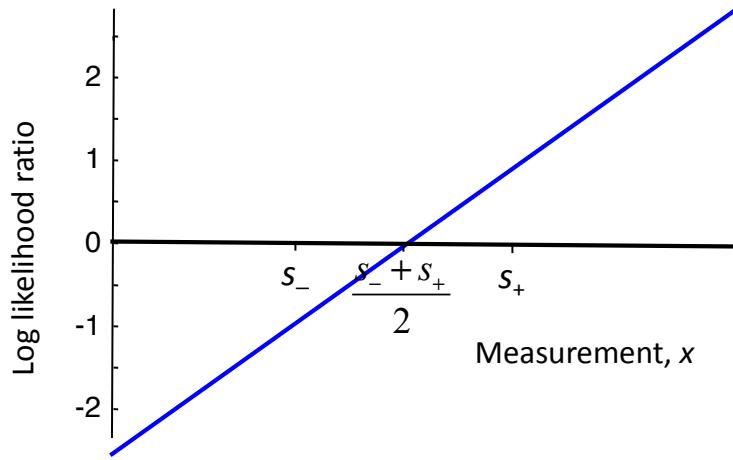
Important message: In a yes/no discrimination task, the prior has the effect of shifting the decision criterion.

When the measurement  $x$  follows a Gaussian distribution, we can further evaluate the log likelihood ratio by substituting the expression for  $p(x|s)$ . This gives

$$\log \frac{p(x|s=s_+)}{p(x|s=s_-)} = \frac{s_+ - s_-}{\sigma^2} \left( x - \frac{s_+ + s_-}{2} \right) \quad (5.5)$$

Exercise 5.12: Show this.

Exercise 5.13: Interpret this expression.



**Figure 5.8.** When noise is Gaussian with  $\sigma$  the same for both stimuli (as in Figure 5.2), the log likelihood ratio in yes/no discrimination is a linear function of the measurement with slope  $(s_+ - s_-)/\sigma^2$ .

In this situation, our decision will depend on our measurement (Fig. 5.6). The log likelihood ratio is positive whenever  $x$  is greater than the mean of  $s_+$  and  $s_-$ , and negative otherwise. This is intuitive: the measurement provides evidence for a class if it lies closer to the stimulus value corresponding to that class. The multiplicative factor  $(s_+ - s_-)/\sigma^2$  scales the magnitude of the log likelihood ratio. This factor tells us that for the same  $x$ , the strength of the evidence is larger when the two stimuli to be discriminated are farther apart ( $s_+ - s_-$  bigger) or when the noise in the measurement ( $\sigma$ ) is smaller.

Substituting Eq. (5.5) for the log likelihood ratio into Eq. (5.4) for the decision rule, we finally arrive at the optimal decision rule for our yes/no discrimination task:

$$\frac{s_+ - s_-}{\sigma^2} \left( x - \frac{s_+ + s_-}{2} \right) > -\log \frac{p(s = s_+)}{p(s = s_-)}. \quad (5.6)$$

The  $1/\sigma^2$  multiplicative factor on the left-hand side acts as a weighting by precision (or reliability), and is reminiscent of Chapters 2 and 4, where evidence was also weighted by its reliability. The weighting matters because the left-hand side is compared to a constant on the right-hand side. Even in binary choice, weighting by precision is a characteristic of the Bayesian observer. Note that the observer must have knowledge of  $s_+$  and  $s_-$  in order to be optimal.

Exercise 5.14: Refer back to the opening example of this subsection. Suppose  $s_+=1^\circ$  and  $s_-=-1^\circ$ , and  $\sigma=0.5^\circ$ . On a given trial, your measurement is  $-0.1^\circ$ , and  $s=s_+$  occurs on 80% of trials. If you are an optimal observer, would you report that the stimulus was  $s_+$  or  $s_-$ ?

In the special case that the prior is flat,  $p(s=s_+)=p(s=s_-)=0.5$ , the decision rule for reporting  $\hat{s} = s_+$ , Eq. (5.6), simplifies to

$$x > \frac{s_+ + s_-}{2}. \quad (5.7)$$

**Important message: Multiple tasks can have the same Bayesian decision rule**

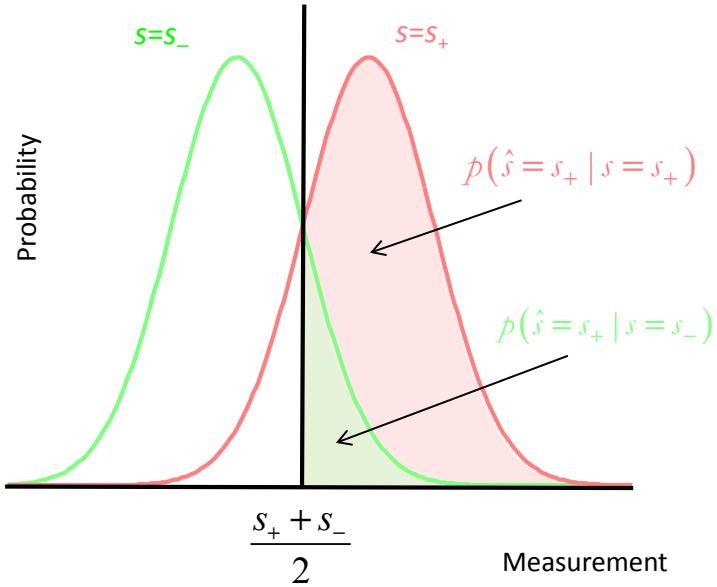
Each binary decision has only one Bayesian decision rule (of course, the same rule can sometimes be written in different forms, for example  $x>0$  would be equivalent to  $e^x>1$ ). However, different tasks can have the same decision rule. As a simple example, consider the decision rule in Eq. (5.7). There are many ways of choosing pairs of stimuli  $(s_+, s_-)$  that have the same mean and therefore the same decision rule. Therefore, it is not possible to reconstruct the task from the decision rule.

#### 4.2.3 Distribution of the MAP estimate

The third step in the Bayesian model is to specify the distribution of the MAP estimate over many trials in which the experimental condition is held fixed. In our task, the experimental condition is completely specified by  $s$ , which can take two values. Therefore, the distribution of the MAP estimate is given by the probability of reporting either  $\hat{s} = s_+$  or  $\hat{s} = s_-$  when  $x$  is drawn from either  $p(x|s=s_+)$  or  $p(x|s=s_-)$ . These four numbers can be reduced to two, since the probability of estimating  $\hat{s} = s_+$  is 1 minus that of estimating  $\hat{s} = s_-$ . Thus, the distribution of the MAP estimate is determined by the following two probabilities:

$$p(\hat{s} = s_+ | s = s_+) = p_{x|s_+}(d(x) > 0) \quad (5.8)$$

and equivalently for  $s_-$ . Here, the notation  $p_{x|s}$  denote the probability of the statement under the distribution  $p(x|s)$ . For convenience, we assume in this subsection that the prior is flat. Evaluating eq (5.8) allows us to calculate predictions for how we expect an observer to behave across multiple trials.



**Figure 5.9.** On which side of  $(s_+ + s_-)/2$  the measurement falls determines the observer's MAP estimate. The probability that the MAP estimate is  $s_+$  is equal to the shaded area when the true stimulus is  $s_-$  (green) or  $s_+$  (red).

We will now compute the probability that Eq. (5.7) is satisfied when  $x$  is drawn from either  $p(x|s=s_+)$  or  $p(x|s=s_-)$ . Eq. (5.7) is satisfied whenever the measurement falls to the right of the vertical line at  $(s_+ + s_-)/2$ . Thus, graphically, the probability that Eq. (5.7) is satisfied is the area under the probability density function to the right of the line. Mathematically, calculating this area corresponds to integrating the density function from  $(s_+ + s_-)/2$  to infinity:

$$\Pr(\hat{s} = s_+ | s = s_+) = \int_{\frac{s_+ + s_-}{2}}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-s_+)^2}{2\sigma^2}} dx \quad (5.9)$$

and equivalently for  $s_-$ . This integral is a cumulative density function of the normal distribution. It cannot be evaluated analytically, but most programming languages have a function that calculates it. This is just a single probability, namely the probability of guessing correctly that the stimulus is  $s_+$ . The probabilities predicted by Eq. (5.9) can be measured in a human psychophysics experiment. The equation predicts how often subjects should estimate  $s_+$  or  $s_-$  as a function of the stimulus really being  $s_+$  or  $s_-$ . As this is what is directly measured in experiments, we now have a straightforward way to compare experiments and theory.

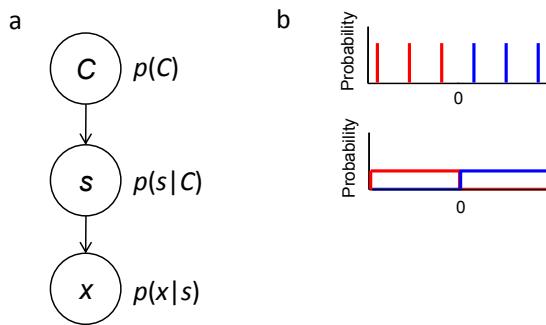
Exercise 5.15: How does Eq. (5.9) change when the prior over  $s$  is not flat?

Yes/no discrimination is a prototype of a binary decision-making task. In the following sections, we explore several generalizations of the formalism: first classification, then detection, and then alternative response paradigms.

### 4.3 Yes/no classification

One unrealistic aspect of the discrimination task discussed so far is that only two stimulus values are allowed. This greatly limits the richness of the data one can collect. In fact, the only quantities that can be measured in an experiment like that are the two probabilities in Eq. (5.9). A much richer and much more common paradigm is one in which many more stimulus values are presented, yet a binary correct response is associated with each. For example, an experimenter could present a random orientation with an integer value between  $-10^\circ$  and  $10^\circ$  with respect to vertical, and ask the subject to report whether it was tilted to the right or left of vertical. This is an example of a *classification* task, defined as a task in which the number of possible response categories is less than the number of possible stimulus values. It is still a yes/no task, since a single stimulus is to be judged – no two stimuli are compared with each other.

#### 4.3.1 Generative model



**Figure 5.9.** (a) Generative model of a classification task in the presence of sensory noise.  $C$  is the class,  $s$  the stimulus, and  $x$  the observation (internal representation) of the stimulus. Associated with these are a prior distribution  $p(C)$ , a class distribution  $p(s|C)$ , and a noise distribution  $p(x|s)$ . (b) Two examples of class-conditioned stimulus distributions: discrete with multiple values and uniform.

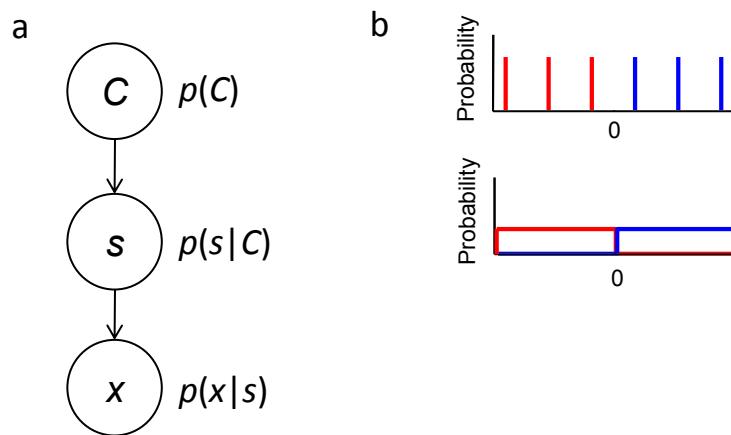
In the yes/no classification task, the observer does not infer the value of a specific stimulus  $s$ , but only the class to which it belongs. Keeping in mind the example of choosing whether an orientation is tilted rightward or leftward relative to vertical (which we define as 0) might be helpful. We denote our state-of-the-world variable by  $C$  for “class”. Inferring the value of a binary state-of-the-world variable is also called *binary classification*. As in 5.1, we are faced with two possible values of  $C$ , which we set to 1 (rightward) and  $-1$  (leftward). We could have

chosen any two values here, and the pair 0 and 1 is more intuitive in other contexts, but here  $C$  is naturally equal to the sign of  $s$ .

### Box: Classification or discrimination?

Classification is one form that either a discrimination or a detection task can take. For instance, a task in which the observer chooses between two sets of orientations is a discrimination task using classification. A task in which the observer chooses between stimulus present (with multiple possible values) or absent, is a detection task involving classification.

Suppose the stimulus values used are  $s_1, s_2, \dots, s_n$  (all positive numbers) when  $C=1$ , and  $-s_1, -s_2, \dots, -s_n$  (all negative numbers) when  $C=-1$ . On each trial, a class is chosen randomly with probability 0.5, and a stimulus value belonging to that class is chosen randomly with probability  $1/n$ . This procedure, called the *method of constant stimuli*, is equivalent to choosing a stimulus value from all  $2n$  possible values, each with probability  $1/(2n)$ , and then post-hoc designating its class. The observer decides whether the stimulus was positive or negative. The observer's behavior is characterized by a plot of the frequency of reporting "positive" (or "negative") as a function of the stimulus value. Any such plot, of a frequency of reports as a function of a physical quantity varied in the experiment, is called a psychometric curve.



**Figure 5.10.** (a) Generative model of a classification task in the presence of sensory noise.  $C$  is the class,  $s$  the stimulus, and  $x$  the observation (internal representation) of the stimulus. Associated with these are a prior distribution  $p(C)$ , a class distribution  $p(s|C)$ , and a noise distribution  $p(x|s)$ . (b) Two examples of class-conditioned stimulus distributions: discrete with multiple values and uniform.

The generative model is shown graphically in Fig. 5.10a. It has three layers: class  $C$ , stimulus  $s$ , and observation  $x$ . The state-of-the-world variable,  $C$ , connects indirectly to the

measurement,  $x$ , by means of the intermediate variable,  $s$ . Associated with  $C$  is a distribution  $p(C)$ , with  $s$  a distribution  $p(s|C)$ , and with  $x$  the usual noise distribution  $p(x|s)$ . The world state distribution again consists of two values,  $p(C=1)$  and  $p(C=-1)$ , which have to sum to 1. These probabilities reflects how often rightward and leftward tilted stimuli occur (in the experiment or in the world). When the orientation is tilted to the left ( $C=-1$ ), it cannot take only a single value, but is randomly drawn from a set of values on the negative line, and similarly for rightward tilts ( $C=1$ ). Thus, each class is defined by a probability distribution over  $s$ : this is the *class-conditioned stimulus distribution*  $p(s|C)$ . The distribution from the example is shown in Fig. 5.10b, top; an alternative distribution (uniform over an interval) is shown in Fig. 5.10b, bottom.

### 4.3.2 Inference model

Since the observer is interested in class,  $C$ , the posterior distribution is now  $p(C|x)$ , not  $p(s|x)$ . This is the first time in the book that the stimulus,  $s$ , does not appear in the posterior: the stimulus is not directly of interest, only the class to which the stimuli belongs to is of interest. Nevertheless, the logic of inference is exactly the same as in previous chapters. In analogy with Eq. (5.3), the Bayesian observer decides based on the log posterior ratio

$$\log \frac{p(C=1|x)}{p(C=-1|x)} = \log \frac{p(x|C=1)}{p(x|C=-1)} + \log \frac{p(C=1)}{p(C=-1)}. \quad (5.10)$$

The question now is how we can write the class likelihood  $p(x|C)$  in terms of the distributions in the generative model. The difficulty is that there is now an intermediate variable,  $s$ . The class likelihood can be obtained by *marginalizing* over the intermediate variable:

$$p(x|C) = \sum_i p(x|s=s_i, C) p(s=s_i|C) \quad (5.11)$$

where the sum is over all values of  $s$ , assumed discrete (as in Fig. 5.10b, top). The marginalization formula, Eq. (5.11), act as a sort of chain rule to link the class,  $C$ , to the observation,  $x$ , by way of the intermediate variable -the stimuli,  $s$ . In words, Eq. (5.11) states that the probability of class  $C$  producing observation  $x$  is the probability that  $C$  would produce stimulus  $s_1$  AND that stimulus  $s_1$  from class  $C$  would produce observation  $x$ , OR that  $C$  would produce stimulus  $s_2$  AND that stimulus  $s_2$  from class  $C$  would produce observation  $x$ , etc. Following the sum and product rules of probability, each “AND” is represented by a product, and each OR by a sum. The result is the overall probability that class  $C$  would produce observation  $x$ .

Exercise 5.16: Prove Eq. (5.11) using only the rules of probability calculus, specifically, the sum rule of probability and the definition of conditional probability.

If  $s$  is distributed continuously (as in Fig. 5.10b, bottom), then the sum is replaced by an integral

$$p(x|C) = \int p(x|s, C) p(s|C) ds. \quad (5.12)$$

So far, we have not made use of the structure of the generative model, and therefore, Eqs. (5.11) and (5.12) are completely general. In the generative model, the distribution of  $x$  only depends on  $s$ , and not directly on  $C$ . In Fig. 5.10a this is graphically understood by the fact that the only arrow pointing to  $x$  comes from  $s$ ; there is no arrow from  $C$  to  $x$ . In other words, when  $s$  is known, knowledge of  $C$  is redundant when one is interested in the distribution of  $x$ . Mathematically, this is expressed as that the conditional distribution  $p(x|s, C)$  is identical to  $p(x|s)$ . Substituting this in Eqs. (5.11) and (5.12), we arrive at the following expressions for the class likelihood (the calculations for the continuous case are perfectly analogous):

$$p(x|C) = \sum_i p(x|s_i) p(s_i|C) \quad (5.13)$$

We have thus achieved our goal of computing the class likelihood in terms of distributions given by the generative model. Eq. (5.13) states that the likelihood of class  $C$  given an observation  $x$  is given by the average of the probability of  $x$  under a stimulus  $s$ , averaged over all possible  $s$  drawn from class  $C$ .

*Daily-life example* of Eq. (5.13): suppose I am interested in the probability  $p(x|C=1)$  that a random Canadian ( $C=1$ ) is a farmer ( $x$ ). I know for each province ( $s$ ) the proportion of farmers ( $p(x|s)$ ). I also know the proportion of all Canadians living in each province,  $p(s|C=1)$ . To obtain my answer, I first multiply those two proportions for every provinces,  $p(x|s)p(s|C=1)$ ; this gives me the proportion of farmers in the province as a function of Canada's population. Finally, I sum over all provinces.

Substituting Eq. (5.13) back into Eq. (5.10),

$$\log \frac{p(C=1|x)}{p(C=-1|x)} = \log \frac{\sum_i p(x|s_i) p(s_i|C=1)}{\sum_i p(x|s_i) p(s_i|C=-1)} + \log \frac{p(C=1)}{p(C=-1)}. \quad (5.14)$$

This is the log posterior ratio  $d$  – the decision variable of the Bayesian MAP observer – for the generative model of Fig. 5.9a. The optimal decision rule is  $d>0$ :

$$\log \frac{\sum_i p(x|s_i)p(s_i|C=1)}{\sum_i p(x|s_i)p(s_i|C=-1)} > -\log \frac{p(C=1)}{p(C=-1)} \quad (5.15)$$

Like in Section 4.2, the prior over  $C$  biases the observer's decision, and its effect is stronger when the sensory evidence is weaker.

*To make further progress, we now assume that the prior is flat.* The optimal decision rule  $d>0$  is then equivalent to  $x>0$ . This is very intuitive, but the proper derivation requires a few steps. We first evaluate the likelihood over  $C$  from Eq. (5.13):

$$\begin{aligned} p(x|C) &= \sum_{i=1}^n p(x|s_i)p(s_i|C) \\ &= \sum_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-Cs_i)^2}{2\sigma^2}} \cdot \frac{1}{n} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{n} \sum_{i=1}^n e^{-\frac{(x-Cs_i)^2}{2\sigma^2}} \end{aligned}$$

Since the prior is flat, MAP estimation is equivalent to reporting class 1 when  $p(x|C=1)$  exceeds  $p(x|C=-1)$ . This inequality can be written as

$$\sum_{i=1}^n \left( e^{-\frac{(x-s_i)^2}{2\sigma^2}} - e^{-\frac{(x+s_i)^2}{2\sigma^2}} \right) > 0. \quad (5.16)$$

At first, it seems that this inequality cannot be simplified any further. However, we can make use of the fact that all  $s_i$  are positive numbers. Since the exponential is a monotonically increasing

function,  $e^{-\frac{(x-s_i)^2}{2\sigma^2}}$  is greater than  $e^{-\frac{(x+s_i)^2}{2\sigma^2}}$  if and only if  $-(x-s_i)^2 > -(x+s_i)^2$ . Evaluating this, we find that this is equivalent to  $xs_i > 0$ , and because  $s_i$  is positive, to  $x>0$ . This shows that the sign of the expression in parentheses is equal to the sign of  $x$  no matter the value of  $i$ . Since a sum of expressions that all have the same sign has the same sign again, the sign of the entire left-hand side is equal to the sign of  $x$ , and Eq. (5.16) is equivalent to the condition  $x>0$ .

It might be amusing that after introducing the notion of marginalization and going through a somewhat involved argument, we end up with a decision rule that is identical to that for discrimination between two stimulus values, say  $s_+=1^\circ$  and  $s_-=-1^\circ$ . In retrospect, there is no other rule that we could have expected, because of the symmetry between left and right in the problem. The simplification to the decision rule  $x>0$  is not limited to the particular class distribution used here (see Problems). However, it is limited to the case of a flat prior. When the prior is not flat (different from 0.5), the decision rule cannot be simplified much beyond Eq.

(5.15). In particular, as this equation shows, the optimal observer will then need to incorporate knowledge of  $p(s|C)$ . This aspect is somewhat problematic for this experimental paradigm and it is often conveniently ignored. We will examine this in greater detail in the Problems.

### 4.3.3 Distribution of the MAP estimate

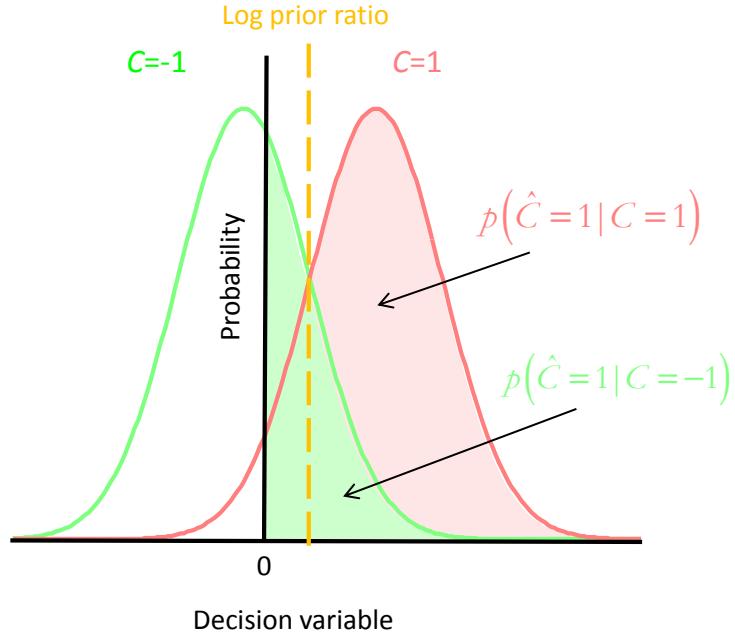
Unlike in Section 4.2, there are now not two, but multiple stimulus conditions in the experiment. Thus, the distribution of the MAP estimate in its greatest possible detail is given by the probability of reporting either  $\hat{C}=1$  or  $\hat{C}=-1$  when  $x$  is drawn from  $p(x|s)$ :

$$\Pr(\hat{C}=1|s) = \Pr_{x|s}(d(x)>0). \quad (5.17)$$

As a side note, we could also consider the distribution of the MAP estimate under either  $p(x|C=1)$  or  $p(x|C=-1)$ . Then, the distribution would be given by the two numbers

$$\begin{aligned} \Pr(\hat{C}=1|C=1) &= \Pr_{x|C=1}(d(x)>0) \\ \Pr(\hat{C}=1|C=-1) &= \Pr_{x|C=-1}(d(x)>0), \end{aligned} \quad (5.18)$$

where we use the notation  $\Pr_{x|C}$  to denote the probability of the statement following it under the probability distribution  $p(x|C)$ . These probabilities are illustrated in Fig. 5.11. They are analogous to Eq. (5.9) but would defeat the purpose of introducing multiple stimulus values, which was to get a richer description of behavior.



**Figure 5.11.** Class-conditioned distributions of the decision variable (the log posterior ratio).

Continuing from Eq. (5.17), we again assume the prior is flat. Since we derived in the previous subsection that  $d>0$  is equivalent to  $x>0$  when the prior is flat, we are simply interested in the probability that  $x>0$  for some true stimulus value  $s$ . This probability produces a point on the psychometric curve as in Fig. 5.1. It is obtained from the cumulative normal distribution:

$$\Pr(\hat{C}=1|s) = \text{Normcdf}(0; s, \sigma) = \int_0^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-s)^2}{2\sigma^2}} dx \quad (5.19)$$

Thus, the psychometric curve of a Bayesian observer, as in Fig. 5.1, is a cumulative normal distribution. The psychometric curve completely specifies the observer's distribution of MAP estimates. In an experiment, the psychometric curve is often fitted with a cumulative normal distribution to find  $\sigma$ , in accordance with Eq. (5.19).

#### 4.4 Detection

In our opening example, a radiologist determined whether a patient has a tumor based on an X-ray. This is an example of a *detection task*: is the tumor present or not? There are many other daily-life examples. As we are showering, we have to determine whether or not our phone rang. On the road, we have to determine whether there is a bump ahead or not. If we have a gas stove, detecting the smell of gas can keep us out of danger. In general, the task is to determine whether a signal is present in noise.

#### 4.4.1 Hit and false-alarm rates

Detection is closely related to discrimination. In the discrimination task discussed in Section 4.2, the observer had to discriminate between two stimulus values,  $s_+$  and  $s_-$ . In its simplest form, detection is the special case in which  $s_+$  is positive and  $s_- = 0$ , i.e., the observer is discriminating between a certain nonzero value and zero. In many cases, the variable  $s$  is somewhat abstract; for instance,  $s$  could be a composite of different image features that a radiologist uses to judge tumor presence. For simplicity, however, we still conceptualize  $s$  as a one-dimensional variable.

The Bayesian model we described for discrimination can thus also be used for detection.

For example, the decision variable is  $d(x) = \frac{s_+}{\sigma^2} \left( x - \frac{s_+}{2} \right)$  and probability correct is  $\frac{1}{2} + \frac{1}{2} \operatorname{erf} \frac{s_+}{2\sqrt{2}\sigma}$ . In a detection task, the probability of reporting “present” when the signal is

present is called the *hit rate*, *detection rate*, or *true positive rate*, whereas the probability of reporting “present” when the signal is absent is called the *false-alarm rate* or *false-positive rate*. These probabilities can all be recognized as areas under the curves in Fig. 5.11. This terminology stems from *signal detection theory*, the branch of Bayesian observer theory that deals mostly with binary variables. 1 minus the hit rate is the *miss rate* or *false-negative rate*, and 1 minus the false-alarm rate is called the *correct-rejection rate* or *true-negative rate* (see Table 5.1). These four terms can also be applied to a discrimination task, such as discriminating a  $-3^\circ$  from a  $3^\circ$  orientation, but in such tasks, it is arbitrary which stimulus is regarded as the “signal”.

In our example, hit and correct-rejection rates are equal, as are the false-alarm and miss rates. Consequently, hit and false-alarm rates sum to 1. In general, however, hit and false-alarm rate do not need to sum to 1. We will see an example in Section 4.4.4.

		Reported class		TOTAL
		present	absent	
True class	present	Hits (true positives)	Misses (false negatives)	1
	absent	False alarms (false positives)	Correct rejections (true negatives)	1

**Table 5.1.** Terminology for the four types of response frequencies in a detection task.

#### 4.4.2 Confidence

In a binary decision, the sign of the log posterior ratio determines the MAP decision. However, the log posterior ratio also has a magnitude or absolute value. A decision made with a log posterior ratio of 0.1 is made less confidently than one with a log posterior ratio of 1: after all, a lower absolute value means that the posterior probabilities of the two alternatives are closer to each other. Therefore, a natural measure of confidence in a binary decision is the magnitude of the log posterior ratio:

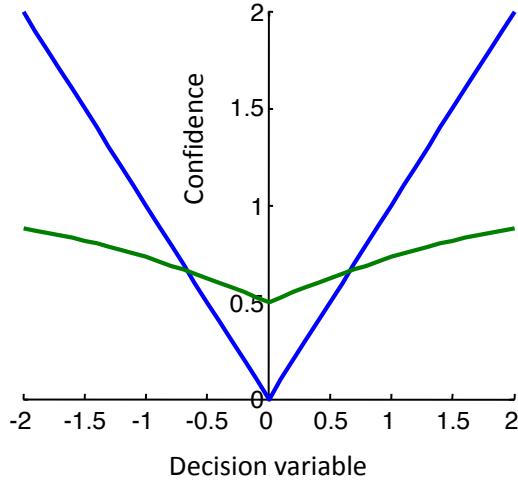
$$\text{confidence} = \left| \log \frac{p(C=1|x)}{1-p(C=1|x)} \right|. \quad (5.20)$$

Confidence can decrease due to a non-flat prior. For example, when the log likelihood ratio is 0.3, and the log prior ratio is  $-0.4$ , then confidence decreases from 0.3 to 0.1 due to the introduction of the non-flat prior.

How does this relate to the measure of confidence we discussed in Section 2.5.1, namely the posterior probability distribution evaluated at the observer's MAP estimate? In the current task, the posterior probability of the MAP estimate is  $p(C=1|x)$  if  $p(C=1|x) > 0.5$  and  $1-p(C=1|x)$  if  $p(C=1|x) < 0.5$ . It is related to the confidence in Eq. (5.20) through:

$$\text{posterior probability of MAP estimate} = \frac{1}{1+e^{-\text{confidence}}}. \quad (5.21)$$

Exercise 5.17: Verify this.

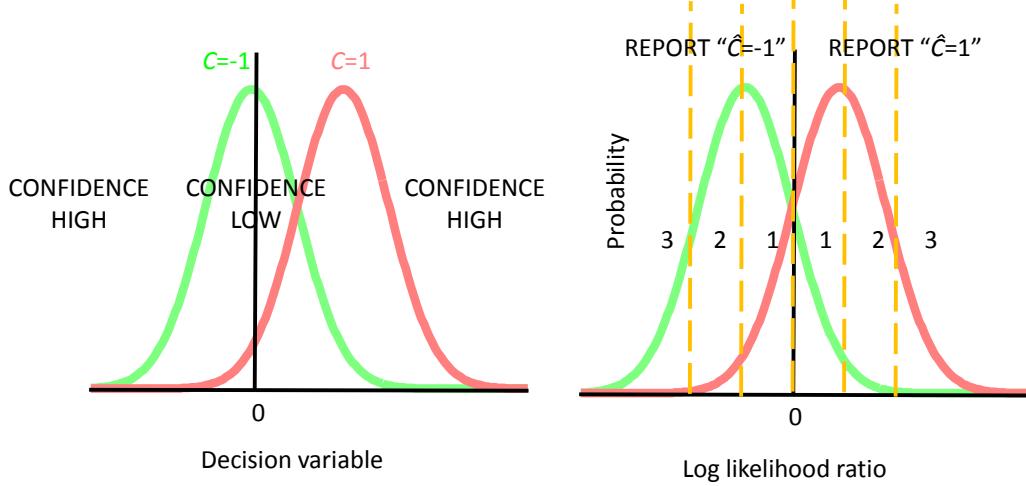


**Figure 5.12.** Two measures of confidence as a function of the log posterior ratio. Blue: the absolute value of the log posterior ratio. Green: the posterior probability of the MAP choice.

Since this mapping – also called the logistic function – is monotonically increasing, the two measures of confidence are in one-to-one correspondence, and either is a legitimate measure of confidence. In Figure 5.12, both measures are plotted as a function of the posterior probability of  $C=1$ . The measure in Eq. (5.21) might correspond most closely to the intuitive notion of *certainty*: one cannot be more than 100% certain, and in the absence of evidence, certainty is 50%. On the other hand, confidence does intuitively not have any limits (and in some people, it does not).

Having established that confidence (either measure) corresponds to distance from the origin on the decision variable axis, we can use the Bayesian model to predict not only the observer's responses on the discrimination (or detection) task, but also how often this decision is made with high or with low confidence. Of course, the dividing line between low and high confidence is unknown, but this is a number that could be fitted to a human subject's data. This is the idea behind a *confidence rating* experiment: the subject is asked not only for a binary judgment on the discrimination task, but afterwards also to rate confidence, let's say as low, medium, or high. Thus, there are now 6 possible responses: 2 class estimates times 3 confidence ratings. We saw before that the Bayesian observer makes a binary judgment by determining in which of two regions in the decision space (the positive and negative axes) the value decision variable falls. Similarly, the Bayesian observer now chooses one of the six possible responses by determining in which of six decision regions the value of the decision variable falls (Fig. 5.13). Three of these regions together form the negative axis, and three the positive axis. From left to right, these regions would correspond to estimating class as  $-1$  with high, medium, and low confidence, and estimating class as  $1$  with low, medium, and high confidence. A total of five

decision criteria separate these regions. When there are  $M$  confidence ratings, the number of points in the plot is  $2M-1$ .



**Figure 5.13.** a) The absolute value of the decision variable is a measure of confidence. b) Response categories in a hypothetical experiment with confidence ratings.

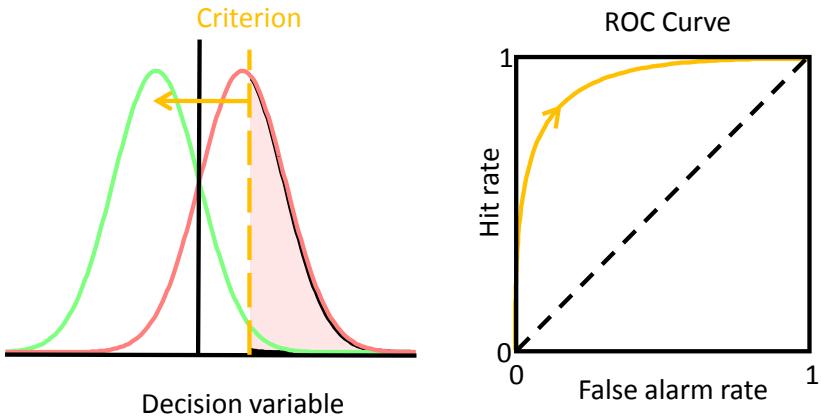
#### 4.4.3 Receiver operating characteristic

In Section 4.4.1, we defined hit and false-alarm rates with respect to one particular decision criterion. In a task with confidence ratings, we can associate a hit and a false-alarm rate with any criterion dividing two adjoining decision regions. In the example with three criteria, the highest criterion would separate class 1 estimates made with medium confidence from those made with high confidence. The generalized hit rate of the Bayesian model is equal to the area under the  $C=1$  distribution of the decision variable,  $p(d|C=1)$ , to the right of a particular criterion  $k$  (which was previously always zero). Similarly, the generalized false-alarm rate is equal to the area under the  $C=-1$  distribution of the decision variable,  $p(d|C=-1)$ , to the right of the same criterion  $k$ . In equations:

$$\begin{aligned} H(k) &= \Pr_{x|C=1}(d(x) > k) \\ F(k) &= \Pr_{x|C=-1}(d(x) > k) \end{aligned} \quad (5.22)$$

If there are three confidence ratings, this leads to five pairs of hit and false-alarm rates, one for each criterion. Plotting hit rate  $H(k)$  against false-alarm rate  $F(k)$  gives us five points in a plot with horizontal and vertical axes both equal to  $[0,1]$ . In the limit of having a very large number of confidence ratings, the plot would contain a smooth curve passing through the origin and through  $(1,1)$ . This would correspond to the decision criterion  $k$  moving continuously along the decision axis from right to left, at each value producing a hit and a false-alarm rate (Fig. 5.10). This curve is called the *receiver operating characteristic* (ROC). It characterizes the

distributions of the decision variable given either class in a more complete manner than the original hit and false-alarm rates can; the latter are essentially only one point on the ROC. The ROC is *parametrized* by the criterion. The ROC is one of the most important concepts in signal detection theory.



**Figure 5.14.** Distribution of the decision variable and the receiver-operating characteristic.

When deciding whether a stimulus is to the right or left of  $\mu$ , or whether a stimulus was equal to  $s_+$  or  $s_-$ , the hit rate is equal to the correct rejection rate and the false-alarm rate is equal to the miss rate. As a consequence, the ROC is symmetrical around the negative diagonal.

Exercise 5.18: Why is this a consequence?

However, this is not the case in general, and in the next chapter, we will see an example of an ROC that is asymmetric around the negative diagonal.

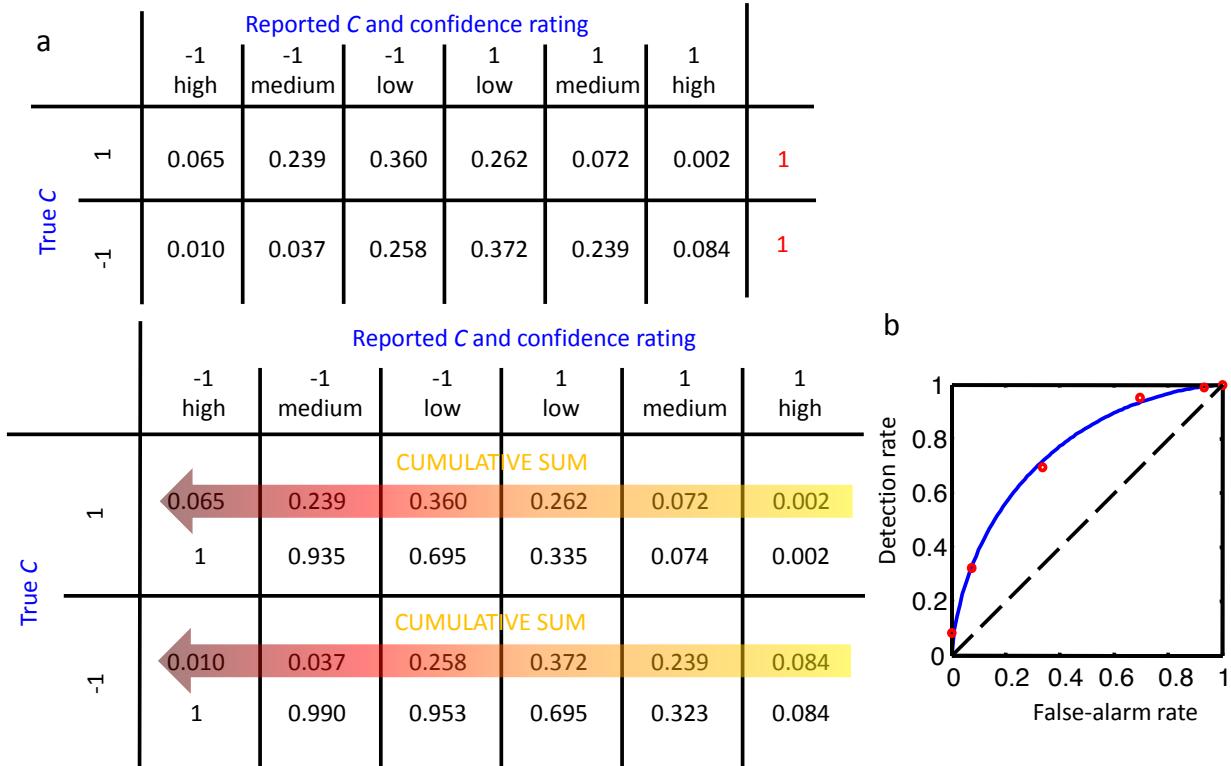


Figure 5.15. Obtaining an empirical receiver-operating characteristic from experimental data

In an actual experiment, an empirical ROC is obtained from the response frequencies in each of the  $2M$  response categories, for each of the two classes. The way to do this is by creating a table of 2 rows and  $2M$  columns (Fig. 5.15a). The top row corresponds to the true class being  $C=1$ , the bottom row to  $C=-1$ . Each column corresponds to a response category. The left  $M$  columns correspond to  $\hat{C}=-1$  responses, in order of decreasing confidence. The right  $M$  columns correspond to  $\hat{C}=1$  responses, in order of increasing confidence. Each cell in the table contains the frequency of responses in each category, divided by the total number of responses across all categories for that class. Thus, the sum of the numbers in each row equals 1. Next, create a new table in which each cell contains the sum of the number in the corresponding cell and all cells to the right of it in the same row in the original table. In other words, the new table is built by cumulatively summing the numbers in the original table from right to left, for each row separately. In the new table, each column corresponds to a (hit, false-alarm) rate pair. The leftmost pair should, by construction, always be equal to  $(1,1)$ . Finally, the hit rate is plotted against the false-alarm rate (Fig. 5.15b). When an observer is accurately described by a certain model, the ROC obtained from that model should go through the points of the empirical ROC.

#### 4.4.4 Bias

A prior different from 0.5 induces a bias in the optimal observer, in the sense that the optimal observer has a tendency to favor the class with the higher prior probability. This is reflected in Eq. (5.4), which states that the log likelihood ratio is compared not to zero, but to the negative log prior ratio. If we had used the log likelihood ratio as the decision variable instead of the log

posterior ratio, then the hit and false-alarm rates would have been different, since the probability that the log likelihood ratio exceeds 0 is different from the probability that it exceeds the negative of the log prior ratio. In spite of the hit and false-alarm rates being different, the receiver-operating characteristic would have remained the same: the ROC is obtained by considering every possible criterion, and the negative log prior ratio is just one possible criterion.

In other words, all decision rules of the form  $d(x) > k$ , where  $d(x)$  is a given function but  $k$  can take any value, produce the same ROC. Even if the observer had learned or assumed the wrong prior but was otherwise optimal, the ROC would be the same, since all that a wrong prior would do is to change  $k$ . Thus, the ROC is invariant under changes in bias. The introduction of a measure that is invariant under changes in bias was one of the main accomplishments of signal detection theory.

Following the example from Section 4.2 of discrimination between two stimuli, we consider the case of a general biased observer, who has the decision rule

$$\log \frac{p(x|C=1)}{p(x|C=-1)} > k.$$

We find for the hit and false-alarm rates:

$$\begin{aligned} H(k) &= \frac{1}{2} + \frac{1}{2} \operatorname{erf} \left( \frac{s_+ - s_-}{2\sqrt{2}\sigma} - \frac{k\sigma}{\sqrt{2}(s_+ - s_-)} \right) \\ F(k) &= \frac{1}{2} + \frac{1}{2} \operatorname{erf} \left( -\frac{s_+ - s_-}{2\sqrt{2}\sigma} - \frac{k\sigma}{\sqrt{2}(s_+ - s_-)} \right). \end{aligned} \quad (5.23)$$

This is an example where hit and false-alarm rates do not sum to 1. When the true prior is flat, the proportion of correct responses can be obtained from

$$\text{proportion correct} = \frac{H + 1 - F}{2}.$$

Exercise 5.19: Show this.

When the true prior is flat  $P(C=1) = P(C=0) = 0.5$  and

$$\text{proportion correct} = \frac{H}{2} + \frac{1-F}{2} = \frac{H + 1 - F}{2}$$

From Eq. (5.23), we see that the proportion of correct responses depends on the criterion,  $k$ . Thus, proportion correct is not a bias-invariant measure of performance. Instead, a common bias-invariant measure of performance is the area under the ROC. For details on this measure, we refer to signal detection theory books (see Further Reading). We discuss another such measure in the next section.

#### 4.4.5 Sensitivity

Sensitivity, also called discriminability and denoted  $d'$  (read “d prime”) is a way to quantify how well separated the class-conditioned distributions of the decision variable are regardless of the placement of the criterion. This measure is defined as

$$d' = \sqrt{2} \operatorname{erf}^{-1}(2H-1) - \sqrt{2} \operatorname{erf}^{-1}(2F-1), \quad (5.24)$$

where  $\operatorname{erf}^{-1}$  is the inverse function to the error function (do not confuse with 1 divided by the error function). The sense of this definition becomes clear when we apply it to discrimination between two stimuli under a Gaussian noise model. Substituting Eq. (5.23) in Eq. (5.24), we find

$$d' = \frac{s_+ - s_-}{\sigma}. \quad (5.25)$$

This remarkably simple expression does not depend on the criterion  $k$ ! No matter how much or how little the observer may be biased, sensitivity only reflects the class-conditioned distributions of the decision variable, i.e. the sensory evidence. The more these distributions overlap, the lower  $d'$  is. As the ratio between the difference between the two stimuli to be discriminated and the level of sensory noise,  $d'$  can be interpreted as a signal-to-noise ratio. From Eq. (5.23) it is clear that hit and false-alarm rates, and therefore proportion correct, can be expressed as a function of sensitivity and criterion: the former reflects the properties of the sensory evidence, the latter the observer’s bias.

#### Discriminability or accuracy?

There might seem to be a tension between discriminability,  $d'$ , and accuracy. In signal detection theory, it is common to regard discriminability as a better measure of performance than accuracy, because it is independent of the criterion. By contrast, accuracy is what is maximized by the Bayesian MAP observer, so it would make sense to use accuracy as a measure of performance. This apparent tension is resolved by noting that the Bayesian MAP observer does not use just any criterion, but uses the optimal one (the one that maximizes the posterior). Thus, accuracy is an entirely valid measure of performance. However, it might still be useful to split up accuracy into hit rate  $H$  and 1 minus false-alarm rate  $F$ . Discriminability is equivalent to accuracy (i.e. perfectly correlated with accuracy) when the class-conditioned distributions of the decision variable are Gaussian distributions with the same variance. In other cases, discriminability is of limited use.

In some texts, Eq. (5.24) is written as  $d' = z(H) - z(F)$  or  $d' = \Phi^{-1}(H) - \Phi^{-1}(F)$ . Here,  $z$  or  $\Phi^{-1}$  refers to the inverse of the cumulative standard normal distribution:  $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$ .

Exercise 5.20: Prove that Eq. (5.24) is equivalent to  $d' = \Phi^{-1}(H) - \Phi^{-1}(F)$ .

A definition of  $d'$  in terms of the means and variances of  $d$  on  $C=1$  and  $C=-1$  trials would be

$$d' = \frac{\langle d \rangle_{C=1} - \langle d \rangle_{C=-1}}{\sqrt{\frac{\text{var } d_{C=1} + \text{var } d_{C=-1}}{2}}}.$$

Exercise 5.21: Prove that this definition is equivalent to Eq. (5.25) for the discrimination task considered in this chapter.

Exercise 5.22: Argue why this definition might be preferable over the definition in Eq. (5.24).

#### 4.4.6 Applications

Signal detection theory has been widely applied in many domains, ranging from detecting objects on radar (for which the theory was originally developed), to determining the ability of a diagnostic test to detect a disease, to studying recall of words from memory (where the experimenter lets subjects study a list of words and then asks if they saw a word in that list). These applications have been described in detail elsewhere. All these situations are examples of the observer using noisy information (a radar image, physiological characteristics of the patient, or a sense of familiarity) to classify the world state into two classes (presence or absence of an object, presence or absence of the disease, having seen the word before or not).

Typically, signal detection theory studies are not described in the same texts as more recent studies of Bayesian perception, such as cue combination as described in the previous chapter. This might seem surprising, since both are based on computing the posterior distribution followed by MAP estimation. The reason for the separation is that historically, signal detection theory has mostly concerned itself with binary discrimination or detection tasks with a flat prior. As we have seen, the Bayesian MAP decision rule is then simply  $x > (s_+ + s_-)/2$ . Unlike the MAP estimate for cue combination, this decision rule does not require any knowledge of stimulus uncertainty, i.e. the width of the likelihood function,  $\sigma$ . (If the prior is not flat, then the decision rule does require such knowledge.) By contrast, more recent studies of Bayesian perception, like those discussed in Chapters 2-4, have focused on tasks that do require knowledge of stimulus

uncertainty. There, we emphasized that the MAP estimate was given by an expression in which the measurement  $x$  is weighted by its reliability or precision,  $1/\sigma^2$ . In the current chapter, we have not emphasized this, but the same feature is present in Eq. (5.6), the log posterior ratio in the discrimination task. In an orientation discrimination task (Fig. 5.4a), the precision could be manipulated for example through presentation time or stimulus contrast. We saw that when the prior is flat, the decision rule simplifies to an equation from which  $\sigma^2$  disappears,  $x > (s_+ + s_-)/2$ . However, in the presence of a non-trivial prior, the weighting by reliability survives. Studies in which observers are optimal even when optimality requires knowledge of sensory uncertainty provide clues that neurons encode entire likelihood functions over stimuli, rather than only maximum-likelihood estimates. The dichotomy between the two branches of Bayesian modeling is therefore mostly a matter of whether the expression for the MAP estimate contains the stimulus uncertainty.

Nevertheless, many concepts from signal detection theory, such as the ROC curve, are generally useful and can also be applied to binary decision tasks in which the expression for the MAP estimate contains the stimulus uncertainty. In fact, we will do so in the next chapter, when discussing more complex generative models.

#### 4.5 Where does the optimal criterion lie?

In a binary decision, when the prior is flat, the optimal decision rule is to compare the log likelihood ratio to zero. We saw before (Fig. 5.5b) that the decision criterion zero would lie exactly halfway between the two class-conditioned distributions of the decision variable, in other words, at the intersection point of the two curves. While this is the most logical position in this case, the placement at the crossing point is no coincidence. To see this, consider the fact that the log likelihood ratio,  $d$ , is computed in a deterministic way from the measurement,  $x$ . Thus, we could replace the generative model in Fig. 5.2a which contained a noise distribution  $p(x|C)$ , by an equivalent generative model in which the log likelihood ratio itself is regarded as the measurement, i.e.  $C \rightarrow d$ , with a noise distribution  $p(d|C)$ . If  $d$  is the measurement, then the Bayesian observer would choose  $\hat{C}=1$  when  $p(d|C=1) > p(d|C=1)$ . This means that the critical value of  $d$  is the one for which both sides are equal,  $p(d|C=1) = p(d|C=1)$ . This is at the intersection point of the two curves. Thus, when the prior is flat, the optimal criterion lies at the intersection point of the noise distributions, regardless of the functional forms of these noise distributions. When there are multiple intersection points, the region in the measurement space producing a  $\hat{C}=1$  judgment will consist of multiple disjointed pieces.

#### 4.6 Concluding remarks

We have introduced the Bayesian framework for binary decision-making, also known as signal detection theory. We have seen that proportion correct is an impoverished way of representing performance, since it does not distinguish between the two types of correct responses, hits and correct rejections. Hit rate and correct rejection rate themselves depend on sensitivity and criterion or bias. The receiver-operating characteristic is a curve obtained by varying the

criterion. Both sensitivity and the area under the ROC are bias-independent measures of performance.

It is important to recognize which of the concepts discussed in this chapter extend beyond the examples of discrimination and detection, and which ones do not. Hit and false-alarm rates, as well as the ROC, extend to any binary decision. Sensitivity, however, is specific to the Gaussian assumption made. When the distributions of the decision variable are not Gaussian with equal variance as they were here, Eq. (5.25) no longer follows from the definition, Eq. (5.24). Whatever its form,  $d'$  loses its significance if it depends on the criterion. The lack of generality of Eq. (5.24) can be recognized in the appearance of the inverse error function, which is associated with the Gaussian distribution. Ultimately, the predictions of a model are not given by any one summary statistic, but by the probabilities of each possible response given the experimental condition on each trial, i.e. Eq. (5.22). The criterion  $k$  might be unknown, either because the observer has an unknown bias or assumed prior, or because the criterion is associated with the boundary between two confidence regions. When this is the case, the criterion can be fitted to the data.

Binary and continuous variables are two ends on a spectrum. A world state variable that is discrete but has a large number of possible values comes close to being continuous. An example would be choosing in which of 8 directions a cloud of dots is moving. All probability distributions in the Bayesian model would be probability mass functions rather than probability density functions. In that sense, all Bayesian inference on a discrete world state variable is very similar to binary decisions. However, many of the concepts introduced in this chapter, such as log posterior ratios, decision rules, and ROCs, are not natural concepts when there are multiple alternatives.

The type of binary decisions considered in this chapter have been rather limited, namely only those where the class  $C$  uniquely specifies the stimulus (whose values we denoted  $s_+$  and  $s_-$ ). Much more general is the case where each class  $C$  determines a *distribution* over the stimulus. For example, in a typical orientation discrimination task, the subject is not asked to discriminate between  $2^\circ$  to the right and  $2^\circ$  to the left of the vertical, but between any leftward tilted and any rightward tilted stimulus. To treat this case properly, we need to introduce the concept of marginalization, which we will do in the next chapter. Marginalization is key in any reasonably complex inference task.

#### Box: Good experimental design for Bayesian modeling

To allow for successfully building a Bayesian model of your psychophysical data, the first requirement is to design an experiment well. There are well-known general guidelines for this, and in addition some that are specific to Bayesian models. In general, one would like to control as many of the parameters that are not of interest to the scientific question. For example, to study how humans perform discrimination, we want to present stimuli for a short

time (a few tens of milliseconds), to avoid complications associated with eye movements, the time course of attention, and the integration of information over time, all of which can effect the quality of encoding (i.e. the standard deviation of the noise distribution) in a potentially complex way. Reaction time experiments are typically more complex to model than accuracy experiments with short presentation times. Therefore, if your scientific question allows to do an accuracy version of the same experiment, it will likely save you work and computation time during modeling.

Similarly, we want to keep attributes of the stimuli that are not of interest as much the same between stimuli. Specifically, make sure to carefully control the reliability/precision/noise level of the stimuli. For example, if you arrange multiple items in a display, arranging them on a circle around the fixation point instead of in a rectangular grid ensures that the eccentricity (distance from the fixation point) is the same and therefore encoding precision is at least approximately the same, allowing to model reliability with a single parameter.

Furthermore, one needs to be well aware of domain-specific effects that can influence performance in the task. For example, when two stimuli are brought close together, an effect known as crowding can occur, in which the internal representations of both stimuli influence each other. If crowding is not of primary interest, it is best to minimize it by placing the stimuli sufficiently far apart from each other. Specific to Bayesian modeling, it is often useful to use stimuli whose feature of interest is one-dimensional or at most two-dimensional. For example, when studying cue combination, it is easier to model a flash and a beep presented on a horizontal line, than to model the integration of the auditory and visual information in speech perception. In the perceptual experiments discussed in this book, we use stimuli that are as simple as possible: they have only a single relevant dimension, for example orientation. Stimuli like letters, line drawings, photographs of objects, or natural scenes are much more difficult to cast into a model because they have many features and in some cases it is even clear what the relevant features (perceptual building blocks) are. Moreover, large number of features translates into a large number of dimensions, and noise models in high numbers of dimensions have even higher numbers of free parameters. This is not to say that studying complex stimuli not interesting, on the contrary. However, more assumptions have to be made in defining the stimulus spaces.

Finally, we recommend that anyone interested in building a Bayesian model of their task write out the model and simulate it before even starting to collect data. For Bayesian models, since they are based on principles of optimality, this is always possible. This process will usually highlight potential problems in the experimental design.

## 4.7 Further reading

Introductory books on signal detection theory that we like:

- Green, D. M. and J. A. Swets (1966). *Signal detection theory and psychophysics*. Los Altos, CA, John Wiley & Sons.

- Macmillan, N. A. and C. D. Creelman (2005). *Detection Theory: A User's Guide*. Mahwah, N.J., Lawrence Erlbaum Associates.
- Wickens, T. D. (2002). *Elementary signal detection theory*. New York, Oxford University Press.

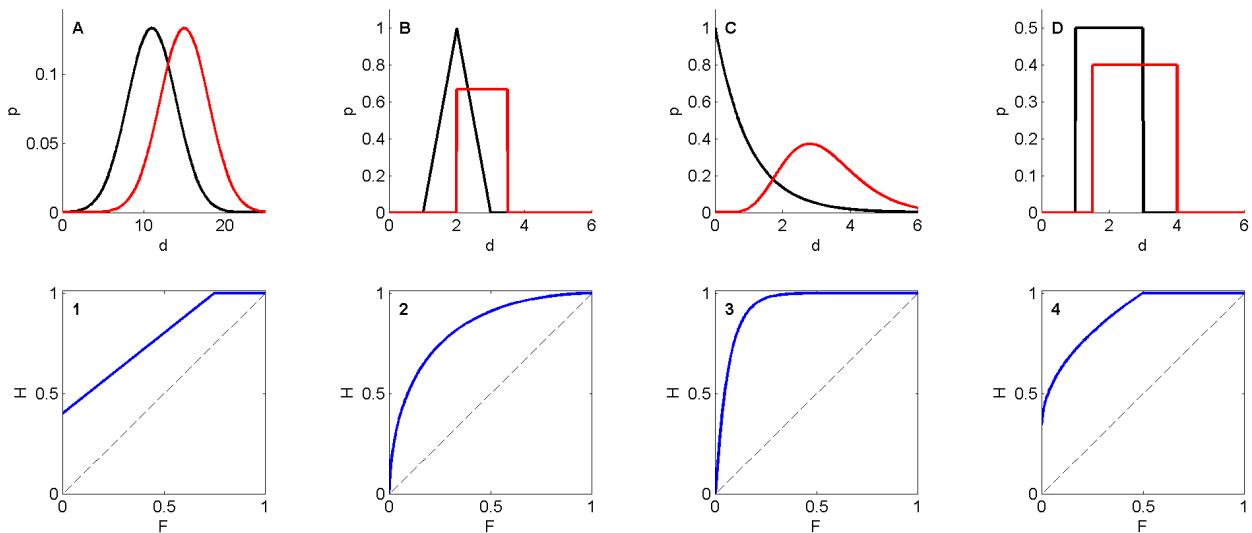
## 4.8 Problems

**Problem 5.1.** Discrete decisions can be viewed as a special case of continuous decisions. The two alternatives in a discrimination task are sometimes two points on a continuum. For example, the two values  $s_+$  and  $s_-$  in Section 4.2 are two orientations on the continuum of all possible orientations (from  $-90^\circ$  to  $90^\circ$ ). Show that inferring the stimulus in such a task can be regarded as a special case of inference on the underlying continuous variable with a suitably chosen prior over that variable.

**Problem 5.2.** In a binary choice, if the prior is given by Fig. 5.2b and the posterior by Fig. 5.2c, calculate the log likelihood ratio. Do the prior and the likelihood favor the same alternative?

**Problem 5.3.** If the log posterior ratio is 0.1, what are the posterior probabilities of the two choice alternatives? What if the log posterior ratio is 1?

**Problem 5.4.** In the first row of the below figure, each plot shows the distributions of the decision variable under each of the two alternatives in a binary decision task. The second row displays receiver-operating characteristic (ROC) curves. Indicate for each ROC to which plot in the top row it belongs.



**Problem 5.5.** An observer is presented with  $N \geq 2$  conditionally independent measurements of a stimulus that can take on two values,  $s_+$  and  $s_-$ . The task of the observer is to respond which of

the two stimuli was presented. The noise distribution is Gaussian with variance  $\sigma^2$  and  $p(s_+)=p(s_-)=0.5$ . Show that the decision rule depends on the sum of the measurements.

**Problem 5.6.** A police investigator is trying to determine whether a suspect is guilty. His prior is 0.5. The investigator has three conditionally independent pieces of evidence. Based on each piece individually, the probability that the suspect is guilty is 60%. What is the probability that the suspect is guilty based on all three pieces of evidence taken together?

**Problem 5.7.** In Sections 4.1.1 and 4.1.2, we considered a Gaussian stimulus distribution. Under what stimulus distribution does the optimal decision rule remain the same?

**Problem 5.8** (Distribution of the decision variable) In the context of Section 4.2, assume a flat prior, so that the Bayesian decision variable becomes the log likelihood ratio. This variable is a random variable that “inherits” its distribution from the distribution of  $x$ . Show that the class-conditioned distributions of the decision variable are

$$p(d | C=1) = \text{Normal}\left(d; \frac{(s_+ - s_-)^2}{2\sigma^2}, \frac{(s_+ - s_-)^2}{\sigma^2}\right)$$

$$p(d | C=-1) = \text{Normal}\left(d; -\frac{(s_+ - s_-)^2}{2\sigma^2}, \frac{(s_+ - s_-)^2}{\sigma^2}\right)$$

**Problem 5.9** (The error function)

The *error function* (see Appendix Section 7.4) is a sort of standardized cumulative normal distribution. Show that Eq. (5.9) can be rewritten as

$$\Pr(\hat{s} = s_+ | s = s_+) = \frac{1}{2} + \frac{1}{2} \operatorname{erf} \frac{s_+ - s_-}{2\sqrt{2}\sigma} \quad (5.26)$$

$$\Pr(\hat{s} = s_+ | s = s_-) = \frac{1}{2} - \frac{1}{2} \operatorname{erf} \frac{s_+ - s_-}{2\sqrt{2}\sigma}$$

**Problem 5.10** (Unequal variances) In the chapter, we discussed a discrimination task with two possible stimulus values,  $s_+$  and  $s_-$ . We assumed that the distributions of the measurement,  $p(x|s_+)$  and  $p(x|s_-)$ , were both Gaussian with equal variances. Now assume instead that their variances are different and have values  $\sigma_+^2$  and  $\sigma_-^2$ , respectively. You may assume that the priors are equal,  $p(s_+) = p(s_-) = 0.5$ .

- a) Derive the log posterior ratio and write down the Bayesian decision rule. Simplify this rule to a set of inequalities for  $x$ . Hints: ....

- b) Derive an expression for the probability that the observer reports  $s_+$  when the stimulus was  $s_+$  and similarly when the stimulus was  $s_-$ , in terms of the error function (see Appendix section 7.4).

**Problem 5.11** (Unmodeled mistakes) Suppose that we perform an experiment with binary responses ( $r=0, r=1$ ) and that  $p(r|s)$  expresses the predicted probability of the observer's response – under an arbitrary model – when the stimulus is  $s$ .

- a) Suppose that the observer accidentally presses the wrong key on a proportion  $\lambda$  of all trials. How does this change the predicted probability of the observer's response?
- b) Suppose that the observer makes a random guess on a proportion  $g$  of all trials (e.g. because he sometimes did not pay attention and didn't see the stimulus). How does this change the predicted probability of the observer's response?

**Problem 5.12** (Cue combination for a binary world state) In Chapter 3, we discussed cue combination as a continuous task with a flat prior in which the Bayesian MAP observer weights each measurement by its precision. Cue combination can also be done for a binary variable. Consider the case of two conditionally independent measurements of the class  $C$ ,  $x_1$  and  $x_2$ , drawn from normal distributions with the same mean ( $s_+$  or  $s_-$ , depending on class). We take the prior over class to be flat. What is the log posterior ratio?

**Problem 5.13.** We worked out the proper Bayesian model for the yes/no discrimination task in Section 4.2. Do the same for the 2AFC discrimination task.

**Problem 5.14** (Binary classification – other distributions)

Consider the classification problem in Section 4.3 with a flat prior.

- a) Derive the decision rule of the Bayesian MAP observer when  $p(s|C)$  is a uniform distribution on the interval  $[-a,0]$  when  $C=-1$ , and a uniform distribution on  $[0,a]$  when  $C=1$ .
- b) Assume that the class distribution is symmetric between the two classes, i.e.,  $p(s|C=-1) = p(-s|C=1)$ , and that  $p(s|C=1)=0$  for  $s<0$ . Show that the Bayesian MAP observer has the decision rule  $x>0$ .

**Problem 5.15.** Approximating the optimal decision rule in classification in the presence of a prior. Start from Eq. (5.15), continuous case, but now don't assume that the prior over  $C$  is flat. Assume that the class-conditioned stimulus distribution is uniform on  $[-s_{\max},0]$  when  $C=-1$  and on  $[0,s_{\max}]$  when  $C=1$ .

- a) Express the optimal decision rule using the cumulative normal  $\Phi$  function. A problem of this rule is that it is complicated and cannot be simplified.

- b) Explain at an intuitive level why  $s_{\max}$  appears in the decision rule.
- c) Now approximate the rule in the limit that  $s_{\max}$  is very large ( $s_{\max} \rightarrow \infty$ ). Two terms in your decision rule should now disappear.
- d) Simulate in Matlab the original decision rule (from part a) and the approximate rule (from part c). Take  $\sigma=1$ . Examine how similar the psychometric functions are for different values of  $s_{\max}$ .
- e) What is the effect of changing the prior over  $C$  using either the original or the approximate rule?

**Problem 5.16.** On each trial, you are shown two stimuli. One of them is tilted to the right and one to the left of vertical. Your task is to report which of the two is tilted to the right. The measurements follow Gaussian distributions with the same variance.

- a) What is the Bayesian decision rule in this task?
- b) Calculate the psychometric curve.

## LAB PROBLEMS

### Problem 5.17: Simulating the ROC in a detection problem

An observer is trying to detect a signal of strength  $s=3$  in noise. The noise has a normal distribution with standard deviation  $\sigma=2$ . On each trial, an experimenter presents noise, or noise plus signal, each with 50% probability. The task of the observer is to respond whether the signal is present or absent.

- a) We will simulate such an observer in Matlab. Start by simulating the measurement on each of 10,000 trials (use the “randn” function). Plot two histograms of the measurements: one for the trials when the signal was present and one for the trials when the signal was absent. Plot them in the same plot (use the “hist” function).
- b) Calculate the decision variable (log posterior ratio) on each trial. Plot the histograms of the decision variables in the same way as you did in (a). How do these histograms compare to the ones in (a), and why?
- c) Assume now that on each trial, the observer also provides a confidence rating by reporting “high confidence” when the absolute value of the log likelihood ratio exceeds 2, “medium confidence” when it lies between 1 and 2, and “low confidence” when it lies between 0 and 1. Create a 2-by-6 table of the two possible stimuli (signal present or absent) and the six possible responses. In each cell, put the numerical frequency of the response conditioned on the stimulus.
- d) Calculate the empirical ROC by cumulatively summing the response frequencies. Plot the resulting points on top of the theoretical ROC based on Eq. (5.23).
- e) What happens when you reduce the signal strength to  $s=2$ ? Interpret.

## Contents

6 Chapter 6: Marginalization .....	6-2
6.1 What is marginalization? .....	6-3
6.2 Nuisance variables and marginalization in single-object perception.....	6-6
6.2.1 Viewpoint invariance .....	6-6
6.2.2 Contrast in orientation estimation.....	6-8
6.2.3 Light and color: light-from-above prior.....	6-10
6.2.4 Inferring reflectance.....	6-10
6.2.5 Random-dot kinematograms .....	6-10
6.2.6 Classification under ambiguity .....	6-12
6.2.7 Size-depth ambiguity .....	6-13
6.3 Multi-object classification: visual search.....	6-19
6.3.1 Camouflage .....	6-20
6.3.2 A visual search experiment.....	6-21
6.3.3 Generalizations .....	6-25
6.3.4 Approximating the Bayesian decision rule .....	6-26
6.4 Posteriors for general generative models: “following the arrows” .....	6-27
6.5 Applications .....	6-28
6.6 Bayesian view of structure perception .....	6-29
6.7 Structure perception: causal inference .....	6-31
6.7.1 Generative model .....	6-32
6.7.2 Inference .....	6-33
6.7.3 Estimate distribution and Monte Carlo simulation .....	6-35
6.7.4 Posterior distribution over the stimulus .....	6-35
6.8 Structure perception: sameness judgment.....	6-37
6.9 Structure perception: contour integration .....	6-39
6.9.1 Simple contours .....	6-39
6.9.2 Natural contours.....	6-41
6.10 Structure perception: Gestalt principles.....	6-42
6.11 Concluding remarks .....	6-46
6.12 Further reading.....	6-48

## 6 Chapter 6: Marginalization

*How can we deal with aspects of the world that are not directly related to the question we want to ask?*

We have seen that Bayesian inference can model a multitude of perceptual tasks, ranging from binary estimation to the estimation of continuous world states, and from tasks involving a single observation to ones involving cue combination. Bayesian inference can model such a wide variety of tasks because it is flexible: the terms entered into Bayes' formula represent the statistical structure of the task at hand, as expressed in the generative model.

In previous chapters, we mainly considered generative models in which the world state of interest gave rise in a rather straightforward manner to the measurement(s). In this chapter, we consider tasks characterized by more complex generative models, in which the path from world state of interest to measurement(s) is less straightforward. In these generative models, additional variables must be taken into account, even though they are not of primary interest to us.

Mathematically, the way to deal with variables in the generative model other than the world state of interest and the measurements is to consider each possible value for those variables according to their probability and average over them. This procedure is called *marginalization*. Marginalization is common in Bayesian models and inevitable in all but the simplest problems. We have already come across marginalization briefly in previous chapters; here, we consider it in more detail. Bayes rule and Marginalization together define the two equations that underlie virtually all of Bayesian statistics.

Plan of the chapter: After describing the marginalization procedure, we discuss several generative models requiring marginalization. Using a variety of visual perception tasks as examples, we begin with relatively simple generative models and progress to more complex ones. We discuss in detail the three steps of a Bayesian model for visual search, whose generative model involves a large number of variables.



**Figure 1:** how do we know this is a bike?

### 6.1 What is marginalization?

In probability theory, marginalization is the operation of turning a probability distribution over multiple variables into a distribution over one of them. For example, if A and B are random variables, and  $p(A,B)$  is their joint distribution, then summing over B produces the distribution over A:

$$p(a) = \sum_b p(a,b) = \sum_b p(a|b)p(b). \quad (6.1)$$

As an example, suppose we roll two regular dice, one at a time. The game we are playing rewards us if the total score from the two rolls is 10. What is the probability that this will occur? To find out, we can consider the probability of every value resulting from the roll of the first die (the ancillary variable), and the probability of a total of 10 given that first value:

$$p(\text{total} = 10 \mid \text{two rolls}) = \sum_{i=1}^6 p(\text{total} = 10 \mid \text{first roll} = i)p(\text{first roll} = i)$$

We are *marginalizing* over the irrelevant value of the first roll. To express the marginalization formula in words, we replace each product with “and” and each addition with “or”. We are stating that the probability of a total of 10 is the probability that the first die lands 1 AND the total will be 10 given that the first lands 1, OR that the first lands 2 AND the total will be 10 given that the first lands 2, and so on. To compute the marginalization sum, we note that if the first die lands 1, 2, or 3, it is impossible for the total of the two dice to reach 10; if the first die lands 4, 5, or 6, then the second die would need to land 6, 5, or 4, respectively, and each of these occurs with probability 1/6. Thus, we have:

$$p(\text{total} = 10 \mid \text{two rolls}) = 0 \cdot \frac{1}{6} + 0 \cdot \frac{1}{6} + 0 \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} = \frac{3}{36} = \frac{1}{12}.$$

For a second example, suppose we want to know the probability that a randomly selected Canadian is a farmer. Canada has 13 geographic regions (provinces plus territories). Suppose we find an almanac that reports the proportion of all Canadians living within each region, and also the proportion of farmers within each region. To obtain the answer, we multiply those two proportions for every region, and then sum over all regions. Here, the region of residence is the ancillary variable:

$$p(\text{farmer} \mid \text{Canada}) = \sum_{i=1}^{13} p(\text{farmer} \mid \text{region}_i, \text{Canada})p(\text{region}_i \mid \text{Canada})$$

We are stating that the probability of randomly selecting a farmer is the probability that we will randomly select a person from region 1 AND that a randomly selected person from region 1 is a farmer, OR that we will randomly select a person from region 2 AND that a randomly selected person from region 2 is a farmer, and so on (Equivalently, we can conceptualize our procedure as

first randomly selecting a region, with a probability proportional to the population of the region, and then randomly selecting a person from within that region.)

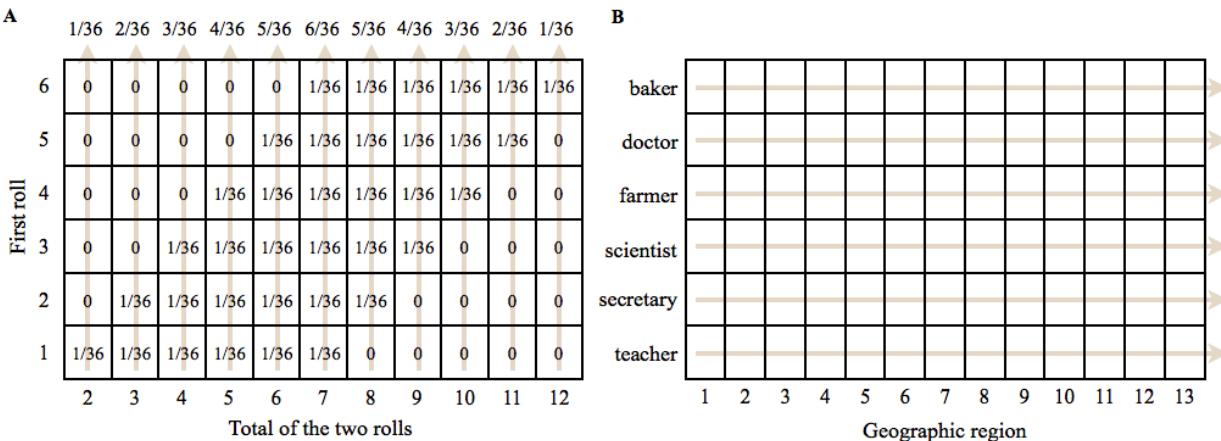
Where does the name “marginalization” come from? According to the product rule,  $p(B|A)p(A)=p(A,B)$ . Our example marginalization equations can therefore be rewritten:

$$p(\text{total} = 10 \mid \text{two rolls}) = \sum_{i=1}^6 p(\text{total} = 10, \text{first roll} = i)$$

and

$$p(\text{farmer} \mid \text{Canada}) = \sum_{i=1}^{13} p(\text{farmer}, \text{region}_i \mid \text{Canada})$$

Thus, we can think of marginalization as the sum over one dimension – the dimension we are not interested in estimating – of a joint probability distribution. When the sum is repeated for each value of the relevant dimension (e.g., not just for dice totals of 10, but for all totals; not just for farmers, but for all occupations), then marginalization reduces the joint distribution to a distribution over just the dimension of interest. Represented graphically, the summation occurs towards the “margin” of the joint distribution, giving marginalization its name (Fig. 6.1).



**Figure 6.1.** Marginalization. Each panel shows a joint probability distribution over two variables. The brown lines represent marginalization over the irrelevant variable, a procedure that reduces the two-dimensional distribution to a one-dimensional distribution over the variable of interest. **A.** Dice example. The value in each square is the probability of that particular (first roll value, total value) pair. Marginalization over the first roll value results in a probability distribution over the total value (top numbers). **B.** Canadian farmer example. Values within each square (not shown) would represent the proportion of Canadians characterized by the corresponding (occupation, geographic region) pair. Marginalization over region results in a probability distribution over occupation. (Only a small subset of occupations is shown).

The marginalization procedure readily generalizes to joint distributions over any number of variables. For example, suppose the almanac had listed, not the proportion of farmers in each region, but the proportion of farmers in each county within each region, as well as the proportion of the region's population living in each county. We could then marginalize over both county and region. Abbreviating farmer  $f$  and Canada  $C$ , and summing over all  $i = 1$  to 13 regions and  $j = 1$  to  $N_i$  counties within each region, we would have:

$$\begin{aligned} p(f | C) &= \sum_{i=1}^{13} \sum_{j=1}^{N_i} p(f | \text{county}_j, \text{region}_i, C) p(\text{county}_j | \text{region}_i, C) p(\text{region}_i | C) \\ &= \sum_{i=1}^{13} \sum_{j=1}^{N_i} p(f, \text{county}_j, \text{region}_i | C) \end{aligned}$$

### Box: Conditioned marginalization

The general form of marginalization where we integrate over all possible values of an unmeasured variable naturally comes from the product rule. However, the whole calculation is unchanged if probabilities are already conditioned on other variables, e.g.  $c$ :

$$p(a | c) = \sum_b p(a, b | c) = \sum_b p(a | b, c) p(b | c).$$

In many cases we have to deal with latent variables that we need to marginalize that are continuous. If  $b$  is a continuous variable, marginalization consists of an integral:

$$p(a) = \int p(a, b) db = \int p(a | b) p(b) db. \quad (6.2)$$

In such cases, everything about marginalization stays the same, the only difference is that sums are replaced with integrals.

One example of such a continuous marginalization is if we want to calculate the probability distribution of the sum of two variables  $a=b+c$  where both  $b$  and  $c$  have Gaussian distributions.

In this case we have  $p(a) = \int p(a|b)p(b)db$  where  $p(b)$  has a Gaussian distribution  $N(\mu_b, \sigma_b)$  and  $p(a|b)$  has a gaussian distribution  $N(\mu_c, \sigma_c)$ . A little bit of algebra, keeping in mind that this is a convolution, shows that  $p(a) \sim N(\mu_b + \mu_c, \sqrt{\sigma_b^2 + \sigma_c^2})$ . This derivation, apart from now being continuous, is exactly equivalent to the sum of dice example we treated earlier.

Exercise 6.0: Prove this mathematically.

## 6.2 Nuisance variables and marginalization in single-object perception

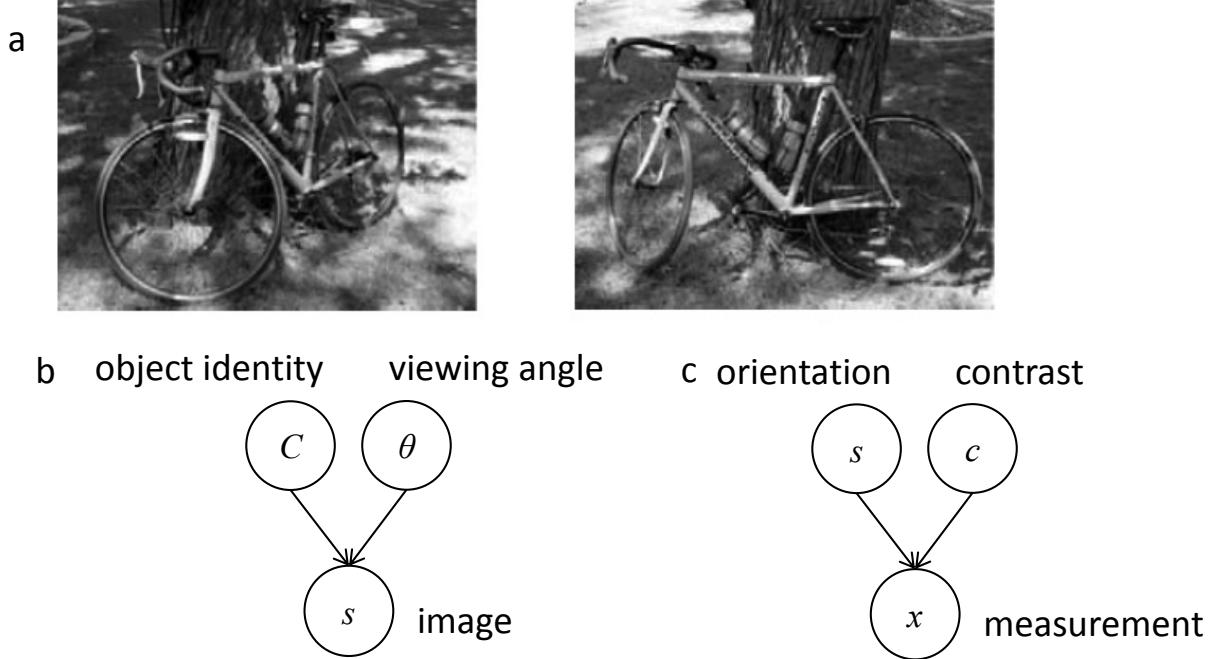
Many generative models relevant to perception involve variables that we are not interested in estimating, but that nonetheless are necessary to deal with, because they constitute an essential part of the formula that links the world state variable of interest with the measurement(s). Marginalization always involves a sum or integral over the possible values of the ancillary variable(s), weighted by the probability of each value.

In many perceptual tasks, ancillary variables play a crucial role in the generative model. Although these variables are not of primary interest to the observer, the observer must take their influence into account, via marginalization, in order to generate an accurate perception. Perhaps because this requires extra computational effort, these ancillary variables have traditionally been called *nuisance parameters*. Conforming we tradition, we adopt that somewhat pessimistic term here, although we think a more adequate description might be *perceptual helper variable*, since without these variables perception in many cases would fail. Here we consider a variety of visual perceptual examples that involve nuisance parameters, and we show how the observer can marginalize in order to compute the likelihood function over the world state variable of interest.

### 6.2.1 Viewpoint invariance

Suppose you want to identify the object in a photograph (Fig. 6.2a). You care only about the object's identity, not the angle from which it was photographed. Yet, the camera angle does help determine the image and therefore the visual information received by your retina. In making the identification, your brain has to somehow *discount* camera angle, and infer only the value of the state-of-the-world variable of interest to you, object identity. In other words, your brain needs to realize that you could be viewing the object from any angle, and take into account how each object (e.g., a bicycle, a car, etc) would look from each angle. If you are able to identify the bicycle from any angle, your identification ability is *viewpoint-invariant*.

The generative model of this task is given in Fig. 6.2b. Besides a node for object identity, it has a node for all irrelevant variables, such as viewing angle. The observation is in this case the image. We are assuming zero sensory noise; if sensory noise were present, there would be an additional node in the generative model, representing the noisy internal representation of the image.



**Figure 6.2.** Object recognition and nuisance parameters. (a) The same object looks different when viewed from a different angle. Viewing angle is a nuisance parameter. Picture and example from Kersten and Yuille (2003). (b) Generative model of the object recognition task.

We denote class by  $C$ , viewing angle by  $\theta$ , and the image by  $s$ . We assume class and viewing angle are independent; this means that it is not the case that certain objects are photographed more often from particular angles than other objects are. The probability distributions in the generative model are the distribution over class,  $p(C)$ , the distribution over viewing angle,  $p(\theta)$ , and the distribution of the stimulus conditioned on both class and viewing angle,  $p(s|C, \theta)$ . Optimal inference in more complex generative models frequently requires knowledge of multiple distributions over world state variables, and properly incorporating those as priors. The posterior distribution over  $C$  is  $p(C|s)$ . This is obtained by first applying Bayes' rule:

$$p(C|s) = \frac{p(s|C)p(C)}{p(s)}$$

and then writing out the class likelihood as a marginalization over  $\theta$ :

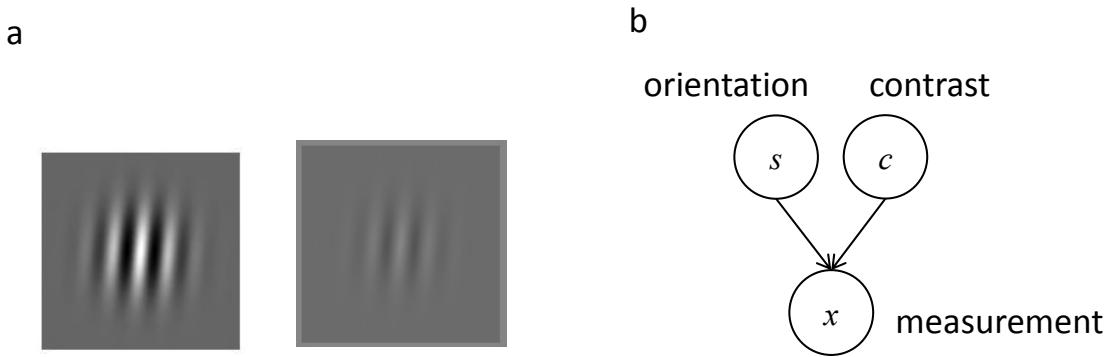
$$p(s|C) = \int p(s|C, \theta)p(\theta|C)d\theta = \int p(s|C, \theta)p(\theta)d\theta \quad (6.2)$$

In the last equality, we have used the information that  $C$  and  $\theta$  are independent random variables, so that  $p(\theta|C) = p(\theta)$ . The formula has the same interpretation as the marginalization

sums discussed in the examples of section 6.1. To appreciate the meaning of Eqn. (6.2), let's consider a particular class, C: bicycle. The marginalization Eqn. (6.2) states that the probability of the visual image, given the object is a bicycle, is the probability the photographer chose to shoot at an angle  $\theta$  relative to the object, AND that a bicycle shot at that angle would produce the visual image we are seeing, OR that the photographer chose to shoot at an angle  $\theta + d\theta$ , AND that a bicycle shot at that angle would produce the visual image we are seeing, and so on, for all  $\theta$ . By computing equation Eqn. (6.2) for many different classes of object, C (bicycle, car, person, etc), the observer can in principle generate a class likelihood function and thereby (via Bayes' rule) a posterior probability distribution whose mode is the most probable object identity.

Note that in Chapter 4 (Binary decisions), we also formulated judging whether an orientation was tilted to the left or to the right as a classification task. Though formally “all orientations to the left” do form a class, the inference was still primarily aimed at identifying a physical stimulus (orientation) there. The left/right distinction was just a convenient method to ask for a subject's response (a more cumbersome method would have been to adjust a probe to match the seen orientation; this would take more time and involve short-term memory). By contrast, in the example above, class captures the meaning of a stimulus. In fact, it is a “natural” category. [THIS DISTINCTION IS NOT HARD; THIS PARAGRAPH NEEDS WORK, SEEMS OK TO KPK]

### 6.2.2 Contrast in orientation estimation



**Fig. 6.3 (a)** Orientation estimation under varying contrast. The observer's task is to reproduce the seen orientation. Shown are a high-contrast and a low-contrast trial. (b) Generative model of orientation estimation when contrast is a nuisance parameter.

We consider an orientation estimation task similar to spatial localization in Chapter 2 (Fig. 6.3). In Chapter 2, we made the implicit assumption that the standard deviation of the noise,  $\sigma$ , was fixed and known throughout the experiment. One prominent nuisance parameter that affects this standard deviation is the visual contrast of the stimulus. Contrast is commonly defined as the luminance of the stimulus relative to the background, divided by the maximum possible luminance of the stimulus relative to the background. A stimulus that is equal in luminance to the

background has contrast 0; one that is maximally bright has contrast 1. Contrast affects the precision of encoding of a stimulus: higher contrast means lower  $\sigma$  and more precise encoding. Imagine now that the experimenter varies contrast randomly from trial to trial. Then,  $\sigma$  is a function of contrast, denoted  $c$ :

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

(Sigma might also depend on  $s$ , but we ignore that here.) This means that the generative model is as in Fig. 6.2c: the distribution of the measurement is determined by two random variables,  $s$  and  $c$ , rather than only by  $c$ . Associated with  $c$  is a world state distribution  $p(c)$ .

In the inference,  $c$  is not of interest to the observer, since the task requires us to estimate orientation. The observer wants to compute solely the likelihood function over orientation,  $s$ . Therefore, she has to integrate out contrast:

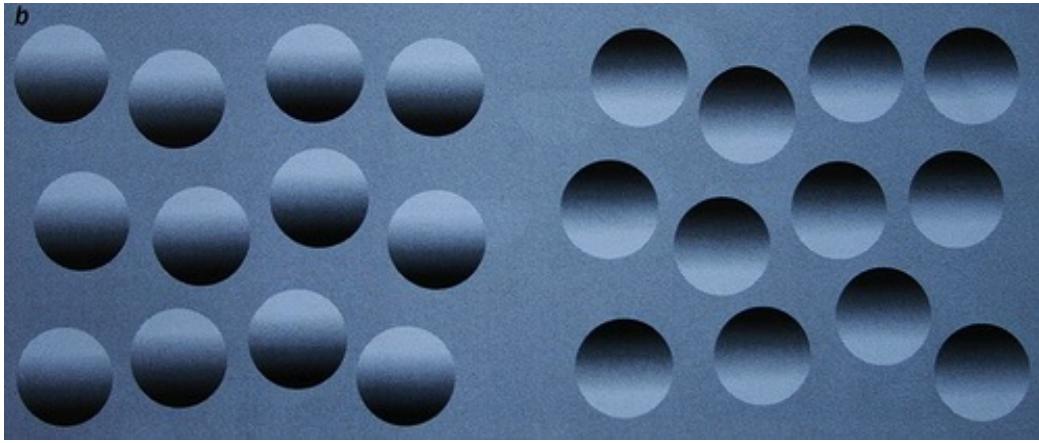
$$\begin{aligned} L(s) &= p(x|s) \\ &= \int p(x|s,c) p(c) dc \\ &= \int \frac{1}{\sqrt{2\pi\sigma(c)^2}} e^{-\frac{(x-s)^2}{2\sigma(c)^2}} p(c) dc \end{aligned}$$

The likelihood function depends on the distribution over contrast,  $p(c)$ . The good news is that the maximum-likelihood estimate of  $s$  is still  $x$ , regardless of  $p(c)$ .

Exercise 6.1: Prove this mathematically.

The bad news is that whichever other operation the observer performs on the likelihood function, such as combining it with a prior over  $s$ , or cue combination, does require knowledge of  $p(c)$ . One way to avoid this is to consider a richer internal representation than just a scalar  $x$ . For example, we will see in Chapter 12 that a population of neurons can simultaneously encode  $x$  and  $\sigma$ , allowing for a code in which marginalization over  $c$  is not needed for computation.

### 6.2.3 Light and color: light-from-above prior



**Figure 3.3.** Illustration of the light-from-above prior. Which of the two sides is seen as hollow, and why?

We assume for simplicity that light can only come from above or below, and that convexity/concavity and light direction completely determine the image (in other words, we restrict ourselves to images like the ones above). The observer needs to marginalize over light direction:

$$\begin{aligned}
 \frac{p(\text{convex} | I)}{p(\text{concave} | I)} &= \frac{p(\text{convex})}{p(\text{concave})} \frac{p(I | \text{convex})}{p(I | \text{concave})} \\
 &= \frac{p(\text{convex})}{p(\text{concave})} \frac{p(I | \text{convex, light } \downarrow) p(\text{light } \downarrow) + p(I | \text{convex, light } \uparrow) p(\text{light } \uparrow)}{p(I | \text{concave, light } \downarrow) p(\text{light } \downarrow) + p(I | \text{concave, light } \uparrow) p(\text{light } \uparrow)} \\
 &= \frac{p(\text{convex})}{p(\text{concave})} \frac{0 \cdot p(\text{light } \downarrow) + 1 \cdot p(\text{light } \uparrow)}{1 \cdot p(\text{light } \downarrow) + 0 \cdot p(\text{light } \uparrow)} \\
 &= \frac{p(\text{convex})}{p(\text{concave})} \frac{p(\text{light } \uparrow)}{p(\text{light } \downarrow)}
 \end{aligned}$$

### 6.2.4 Inferring reflectance

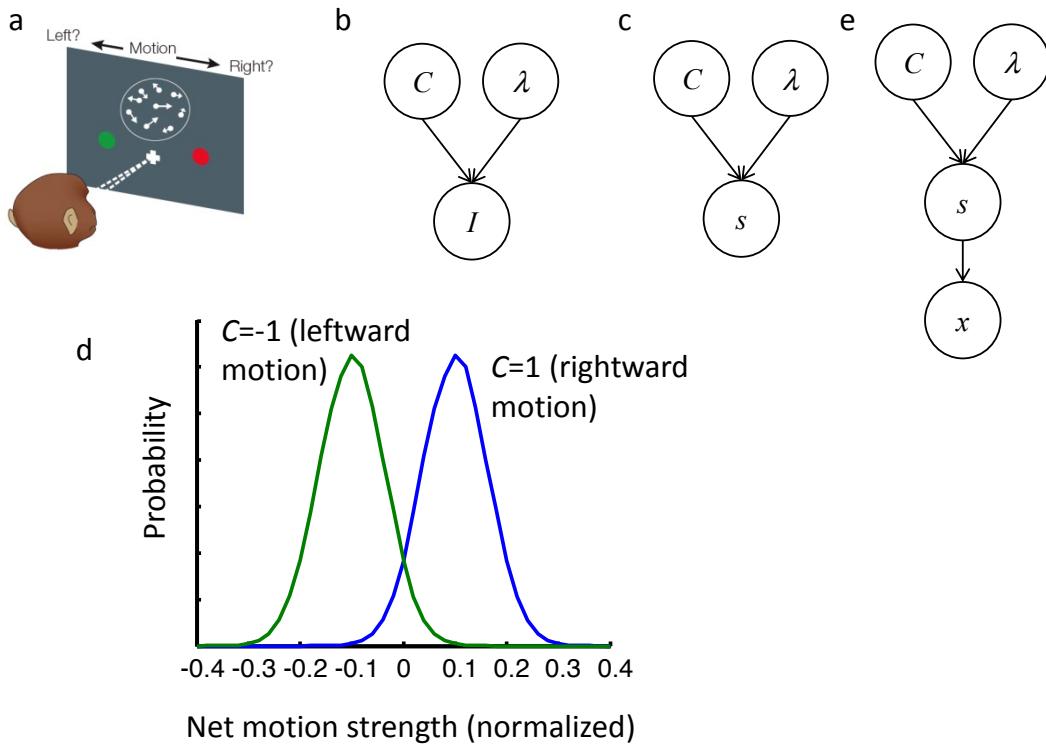
<IS THERE SUPPOSED TO BE STH HERE?>

### 6.2.5 Random-dot kinematograms

Fig. 6.3a shows a random-dot kinematogram, used often in studying decision-making in monkeys. The world state variable of interest is direction of motion, which can be to the left or the right. The nuisance parameter is coherence, defined as the proportion of dots moving in the “correct” direction, while all other dots move in random directions. When coherence is 1, all dots

move in the same direction (left or right). When coherence is 0, there is no net motion signal on average. Even in the absence of noise, perfect performance on this task is impossible, because a true motion direction to the right can give rise to a net dot motion direction to the left, simply because of the randomness in the non-coherently moving dots. Thus, this task contains ambiguity. Coherence is typically varied randomly from trial to trial. The generative model of this task is shown in Fig. 6.3b. By  $I$ , we denote the visual information on the screen – this does not even take into account any noise added between the screen and the neural representation. Similarly to Section 6.2.2, the likelihood over  $C$  based on  $I$  is obtained by marginalizing over coherence  $\lambda$ :

$$\begin{aligned} L(C) &= p(I|C) \\ &= \int p(I|C, \lambda) p(\lambda) d\lambda \end{aligned}$$



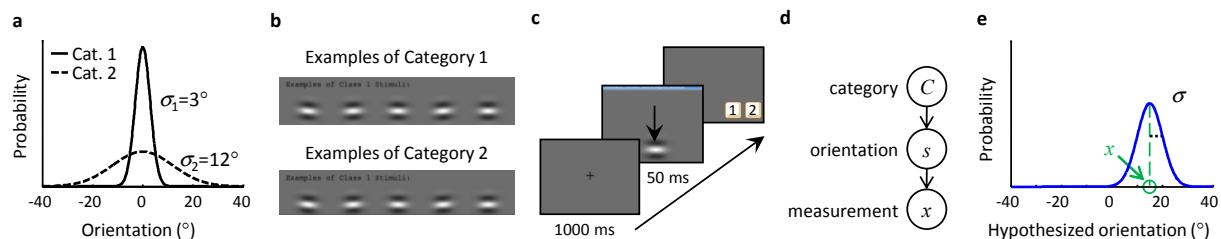
**Figure 6.3. A motion discrimination task.** (a) Observers view a display of moving dots. Some dots move coherently to the right or to the left, others in random directions. The observer has to indicate the direction of motion of the coherently moving dots. The percentage of coherently moving dots is called coherence and denoted  $\lambda$ . (b) Generative model of this task.  $C$  indicates class (coherent motion to the left or to the right), and  $I$  is the sequence of visual images produced. (c) Simplified generative model, where the display is summarized by the relevant variable, namely the mean direction of motion of all dots. (d) Distribution of the net motion strength  $s$  when coherent motion is leftward ( $C=-1$ ) or rightward ( $C=1$ ). We simulated 100 dots, a coherence of 10%, and the motion vector of each dot had length 1. Although these distributions overlap, the optimal decision rule is still to report “rightward” when  $s>0$ . (e) There might be noise in the measurement as well, giving rise to an extra measurement node,  $x$ .

To be a bit more concrete, we can summarize the image  $I$  using the net motion direction of the dots, denoted  $s$  (Fig. 6.3c) assuming that there are  $N$  dots. Then we can calculate the distribution of  $s$  conditioned on  $C$  and  $\lambda$ . For example, when  $C=1$  (rightward motion) and  $\lambda=0.1$ , then each dot has a 10% probability of moving rightward and a 80% probability of moving in any direction. The net motion direction consists of two terms: a constant term of magnitude 0.1, and a variable term reflecting the mean of  $N$  identically distributed random variables, each of which has the form  $\cos(\alpha)$ , where  $\alpha$  has a uniform distribution. This can only be calculated numerically (Fig. 6.3d) to calculate the distribution of net motion strengths.

The optimal decision rule in this problem is always  $s>0$ . One could model sensory noise in this problem by adding a measurement variable  $x$  to the generative model, as in Fig. 6.3e. Then, the optimal decision rule would always be  $x>0$ .

### 6.2.6 Classification under ambiguity

Suppose that an observer's task is to classify an orientation as class 1 or class 2. The prior over class is flat. Class 1 orientations are drawn from a Gaussian distribution with mean 0 and standard deviation  $\sigma_1$ . Class 2 orientations are drawn from a Gaussian distribution with mean 0 and standard deviation  $\sigma_2$ . The distributions are illustrated in Fig. 6.4a, with example stimuli in Fig. 6.4b.



**Figure 6.4. Task and generative model.** (a) Probability distributions over orientation that define Categories 1 and 2. The distributions have the same means but different standard deviations (for Monkey L,  $\sigma_2$  was  $15^\circ$ ). (b) Examples of stimuli drawn from each category. (c) Trial procedure for humans. A grating drifts in a direction perpendicular to its orientation. The subject reports category through a key press. The trial procedure for the monkeys was similar (see Methods). (d) Generative model of the task. (e) Likelihood function over orientation on a trial when the measurement is  $x$ . The width of this function,  $\sigma$ , measures sensory uncertainty on this trial.

We assume class 1 corresponds to the narrower distribution, so  $\sigma_1 < \sigma_2$ . Because these distributions overlap, any given orientation could come from either class, though usually not with the same probability. The Bayesian observer picks the class with the highest probability. For now, we assume there is no sensory noise. Thus, the generative model is simply  $C \rightarrow s$ . The log posterior ratio is equal to

$$\log \frac{p(C=1|s)}{p(C=2|s)} = \log \frac{\sigma_2}{\sigma_1} - \frac{s^2}{2} \left( \frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} \right) \quad (6.2)$$

Exercise 6.2: Show this.

It follows from Eq. (6.2) that the MAP observer estimates  $\hat{C}=1$  when

$$s^2 < \frac{2 \log \frac{\sigma_2}{\sigma_1}}{\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}}. \quad (6.2)$$

This is the first time we have encountered a decision variable that is quadratic in the observation (remember that there is no sensory noise, so the observation is equal to  $s$ ). This makes perfect sense, however. It is clear from Fig. 6.4a that the probability of  $s$  is higher under class 1 than under class 2 when  $s$  falls in a narrow region around 0. There are two decision criteria on  $s$ , one on each side of 0. The absolute values of these criteria are equal and given by the square root of the right-hand side of Eq. (6.2).

As we have framed the example so far, the estimate distribution of the MAP observer is trivial because we assumed zero sensory noise. Because of this assumption, the observation is  $s$  whenever the stimulus is  $s$ . Therefore, the probability of estimating  $\hat{C}=1$  is simply 0 or 1, depending on the value of  $s$ .

However, if sensory noise is present in the task, the generative model is as in Fig. 6.4d, and the likelihood of class  $C$  is then given by a marginalization integral:

$$p(x|C) = \int p(x|C, s) p(s|C) ds$$

The Bayesian model for this case is worked out in a Problem.

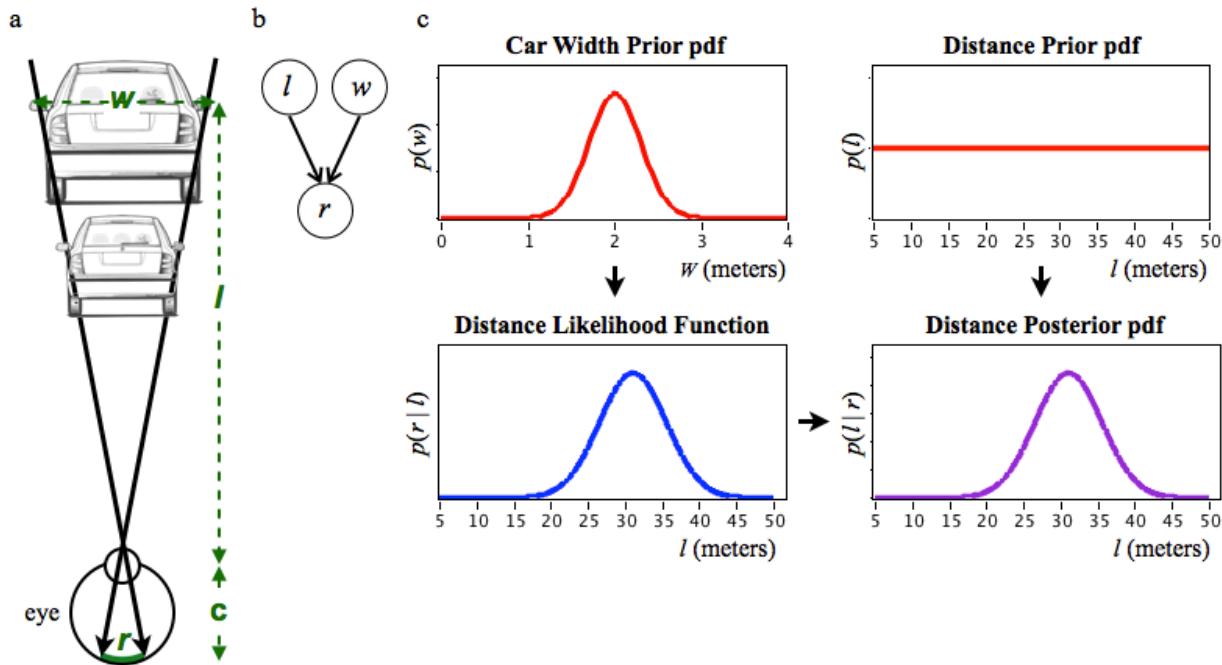
### 6.2.7 Size-depth ambiguity

Consider a defensive driver who wants to maintain a safe distance between her car and the one ahead. To do so, she must accurately perceive the distance to the next car. Under good visual conditions, a driver has many cues to aid depth perception. However, when visual conditions are poor, as for instance in darkness or fog, the number of distance cues is diminished. One cue a driver can use provided simply that she can see the taillights of the car ahead, even if with just a single eye, is the size of the image of the car on her retina. The observer's task, then, is to estimate the distance (i.e. depth) to the next car,  $l$ , from the size of the retinal image of that car,  $r$ . Let's assume that she has a flat (uniform) prior over distance (If she were familiar with traffic

conditions on the road in question, or similar roads in the same city, she could express a non-uniform prior, which would increase the accuracy of her inference). Then Bayes' rule tells us

$$p(l|r) \propto p(r|l)p(l) \propto p(r|l)$$

But how do we calculate the likelihood function,  $p(r|l)$ ? More precisely, how does  $l$  influence  $r$ ? The farther away an object is, the smaller is the image that it projects onto our retina. Nevertheless, the size of a *completely unfamiliar* object's retinal image provides no information regarding its distance away from us. This is because a larger object that is farther away will produce the same retinal image as a smaller, closer object (Fig. 6.5a). It is not possible, then, to accurately estimate the distance to an entirely unfamiliar object, based only on the size of its retinal image. Luckily, however, the vast majority of objects that we encounter in life – cars included – are *not* entirely novel to us. In fact, most of us have a great deal of familiarity with cars, and so we have considerable knowledge, gained from years of experience, regarding the distribution of the sizes and shapes of cars on the road.



**Figure 6.5.** Depth perception from image size. **(a)** At distance  $l$  from the observer, a car of width  $w$  produces a retinal image of width  $r$ . From trigonometry,  $r/c = w/l$ , where  $c$  is the distance from the observer's pupil to her retina; a smaller car, closer to the observer, would subtend the same visual angle and produce the same retinal image. The observer can therefore infer the distance to the car only if she has a prior opinion about the car's size. **(b)** The generative model. The nuisance parameter,  $w$ , and the world state variable of interest,  $l$ , are both needed to generate the observation,  $r$ . We assume no noise in the observation. **(c)** Inference. The observer's prior over car width (here a Gaussian with mean 2 m and standard deviation 0.3 m) translates

into a likelihood function over distance. The observer's likelihood function combines with her prior over distance (here uniform) through Bayes' formula, to produce a posterior over distance. The car width used in this example was 1.65m, and its distance was 25m. The observer's MAP estimate for distance is 31m: the car appeared to be farther away than it really was, because it was somewhat smaller than the observer expected.

Importantly, we can use our knowledge of car widths to generate the likelihood function that relates the car's depth,  $l$ , to the size of the retinal image,  $r$ . Thus, in the generative model (Fig. 6.5b), car width,  $w$ , is a nuisance parameter that is needed to bridge the gap between the world state variable of interest (depth) and the observation (retinal image). In order to calculate  $p(r|l)$ , we marginalize over the nuisance parameter, as in Eq. (6.2):

$$p(r|l) = \int_w p(r|w,l) p(w|l) dw = \int_w p(r|w,l) p(w) dw$$

In the second equality we have replaced  $p(w|l)$  with  $p(w)$ , because  $w$ , the width of the car, is independent of the car's distance from us,  $l$ .

Before we evaluate this integral, let's take a moment to appreciate its meaning. Recall that  $p(r|l)$ ,  $p(r|w,l)$ , and  $p(w)$  are all probability densities. Thus,  $p(w)dw$  is the prior probability that  $w$  occupies the range  $w \pm dw/2$ ;  $p(r|w,l)dr$  is the probability, given  $w$  and  $l$ ,  $r$  occupies the range  $r \pm dr/2$ ; and  $p(r|l)dr$  is the probability, given  $l$ , that  $r$  occupies the range  $r \pm dr/2$ . In the equation, only one probability occurs:  $p(w)dw$ . The other terms are probability densities over  $r$ . With this in mind, let's multiply both sides of the equation by  $dr$ , in order to convert those densities into probabilities as well:

$$p(r|l)dr = \int_w p(r|w,l)dr p(w)dw$$

We can now interpret this equation simply, using the sum (AND) and product (OR) rules of probability: The probability, given  $l$ , that  $r$  occupies the range  $r \pm dr/2$ , is the probability that  $w$  occupies the range  $w_0 \pm dw/2$  AND  $r$  occupies the range  $r \pm dr/2$ , OR that  $w$  occupies the range  $w_1 \pm dw/2$  AND  $r$  occupies the range  $r \pm dr/2$ , and so on for all possible values of  $w$ , where  $w_0$  is the lowest value,  $w_1 = w_0 + dw$ ,  $w_2 = w_1 + dw$  and so on.

Now, what is the value of  $p(r|w,l)dr$  as a function of  $w$ ? Since  $w$  and  $l$  together specify  $r$  (from trigonometry,  $r = cw/l$ ; see Fig. 6.5a), it follows that, for any particular  $r$  and  $l$ ,  $p(r|w,l)dr = 0$  for all values of  $w \neq lr/c$ , and  $p(r|w,l)dr = 1$  when  $w = lr/c$ . Therefore, the integral becomes simply

$$p(r|l)dr = p(w = lr/c)dw$$

It follows that:

$$p(r|l) = p(w = lr/c) \frac{dw}{dr} = \frac{l}{c} p(w = lr/c) \quad (6.2)$$

### Transformation of variables

At a given  $l$ , there is a unique mapping from  $w$  to  $r$ :  $r = cw/l$  (Fig. 6.5a). Accordingly, we might have thought, mistakenly, that we need simply to read off the height of the prior pdf over  $w$ , at the value  $w = lr/c$ , in order to find the likelihood,  $p(r|l)$ . This would have resulted in the incorrect conclusion that  $p(r|l) = p(w = lr/c)$ . This mistaken reasoning overlooks a crucial point: the distributions  $p(r|l)$  and  $p(w)$  are not probabilities but probability densities. The area  $p(r|l)dr$  is the probability, given  $l$ , that  $r$  falls within the infinitesimal range  $r \pm dr/2$  and the area  $p(w)dw$  is the prior probability that  $w$  falls in the infinitesimal range  $w \pm dw/2$ . Since, for a given  $l$ ,  $w$  maps to  $r$ , it is these areas – not the densities – that are equal. Thus,  $p(r|l)dr = p(w = lr/c)dw$ , as derived above. Note that  $p(r|w, l)$  is a *delta function*. For an explanation of the delta function and a more detailed and formal exposition of the transformation of variables procedures, see the Appendix, Sections 8 and 11.1.

Having calculated the likelihood function  $p(r|l)$ , we can now plot this as well as the driver's posterior pdf,  $p(l|r)$  (which as explained above is proportional to  $p(r|l)$ , since the prior over distance is uniform) (Fig. 6.5c). Entering into Eqn. (6.2) the observer's Gaussian prior over car width (described by mean  $\mu_w$  and standard deviation  $\sigma_w$ ) evaluated at  $w = lr/c$ , we can write her posterior pdf over distance as:

$$p(l|r) \propto \frac{l}{c} e^{-\frac{\left(\frac{lr}{c} - \mu_w\right)^2}{2\sigma_w^2}}$$

We can now derive the observer's MAP distance estimate. To do so, we first take the logarithm of her posterior pdf:

$$\log p(l|r) = \log l - \frac{\left(\frac{lr}{c} - \mu_w\right)^2}{2\sigma_w^2} + \text{const.}$$

Here we have subsumed all terms that do not depend on  $l$  into the final constant. Next, we take the first derivative with respect to  $l$ :

$$\frac{d}{dl} \log p(l|r) = \frac{1}{l} - \frac{\left(\frac{lr}{c} - \mu_w\right)r}{\sigma_w^2}$$

Now recall that, although the observer does not know the car's true width or distance, she does know that the ratio of these values is  $r/c$  (see Fig. 6.6a):

$$\frac{r}{c} = \frac{w_{true}}{l_{true}}$$

Substituting this expression into the derivative of the log posterior, and setting that derivative to zero in order to find its mode  $l_{MAP}$ , we have:

$$\frac{1}{l_{MAP}} - \frac{\left( l_{MAP} \frac{w_{true}}{l_{true}} - \mu_w \right) \frac{w_{true}}{l_{true}}}{\sigma_w^2} = 0$$

This solution to this equation is:

$$l_{MAP} = l_{true} \frac{\mu_w}{w_{true}} \cdot \frac{1}{2} \left( 1 + \sqrt{1 + 4 \left( \frac{\sigma_w}{\mu_w} \right)^2} \right) \quad (6.2)$$

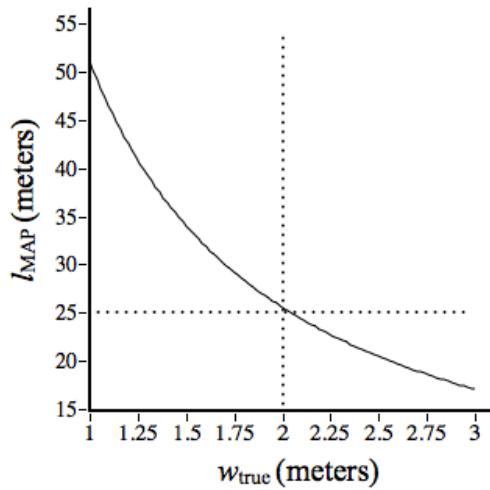
Exercise 6.3: Show this.

Notice that, to the extent that the car's width,  $w_{true}$ , exceeds the mean of the observer's prior,  $\mu_w$ , she will underestimate the car's distance (Fig 6.6): she will perceive a wider-than-average car as being closer than it really is. Conversely, to the extent that the car's width falls short of the mean of her prior, she will overestimate its distance: she will perceive a smaller-than-average car as farther away than it really is. Nevertheless, the observer has gained valuable information regarding the car's probable distance, based on a single visual cue (retinal image size).

Exercise 6.4: In the special case that  $\sigma_w = 0$ , Equation (6.2) reduces to  $l_{MAP} = l_{true} \left( \frac{\mu_w}{w_{true}} \right)$ .

Explain why this makes sense.

Exercise 6.5: When  $\sigma_w > 0$  and  $w_{true} = \mu_w$ , the MAP estimate  $l_{MAP}$  is still greater than  $l_{true}$ . Explain why this makes sense.



**Figure 6.6.** The observer's MAP estimate when viewing cars of different sizes. The observer's prior over car width was a Gaussian centered at  $\mu_w = 2m$  (vertical dotted line) and with standard deviation  $\sigma_w = 0.3m$ . The actual distance to the car was  $l_{true} = 25m$  (horizontal dotted line). The plot shows that when the observer views a car of normal size (i.e., a car of the expected width, 2m), she perceives its distance with good accuracy. In contrast, cars that are smaller than expected ( $w < 2m$ ) are perceived to be farther away than they really are, while cars that are larger than expected ( $w > 2m$ ) are perceived to be closer than they really are.

The previous example illustrates how our understanding of an object's size affects our perception of its distance. Conversely, our understanding of an object's distance also affects our perception of its size. Figure 6.7 illustrates three examples of the *Ponzo illusion*. Within each panel, the figures have the same size on the page (and therefore on the retina), but the topmost figure appears larger. This occurs because the brain interprets the two-dimensional picture as a three-dimensional scene, where the context suggests that the topmost figure is farther away. The only way one object can have the same retinal image size as another and yet be farther away is if the farther object is larger. Your thumb, viewed at arm's length, may occupy the same retinal area as a distant mountain (for this reason, you can use your thumb to block your view of the mountain). Size-depth ambiguity occurs because objects with an infinite set of possible physical sizes can produce the same retinal image size, depending on their distance from the observer. The more confident we are of the distance to an object, the more confident we can be of its size.



**Figure 6.7.** Three variants of the Ponzo illusion. In each case, the topmost figure (line segment) looks taller (wider) even though it is physically equally tall/wide in the drawing.

### 6.3 Multi-object classification: visual search

We now work out in detail the Bayesian model for a task that is more complex than the examples we have seen in the previous sections, namely visual search. This is a classification task, but it involves multiple objects as well as marginalization over spatial location.



**Figure 6.8.** Visual search (a) In the animal world. (b) In the human world. (c) In the laboratory.

To an animal in the wild, performing efficient visual search can be a matter of life or death. Frequently, animals must detect whether a predator is present in a visual scene. The predator might be hidden or camouflaged, making it very difficult to distinguish from the surrounding visual elements (Fig. 6.8a). As a modern-day human, we might want to find a particular piece of paper on a cluttered desk (Fig. 6.8b), or determine whether our keys are present among a large set of objects some of which (coins, other keys) bear similarity to our keys.

Doing psychophysics with natural scenes is difficult for several reasons. First, they are so rich in content that describing them mathematically would require a very high-dimensional space. Second, it is not clear what exactly an object is: is the tree the object, or an individual branch, or perhaps even a leaf? Third, the noise in the different dimensions of a natural scene typically has a complex and largely unknown correlation structure that goes far beyond the Gaussian noise distribution that we have considered so far. Therefore, in laboratory visual search tasks, as well as in other laboratory psychophysics, it is common to use extremely simplified search scenes that contain a relatively small number of highly distinct objects that differ only along a single stimulus dimension (Fig. 6.8c). Obviously, a big gap exists between such simple, artificial scenes and natural tasks, but it is hoped that the majority of the computational mechanisms we can unveil using laboratory tasks are also relevant in natural tasks – in essence, that the laboratory task allows us to study the minimal building blocks of the computation that the brain performs in the real world.

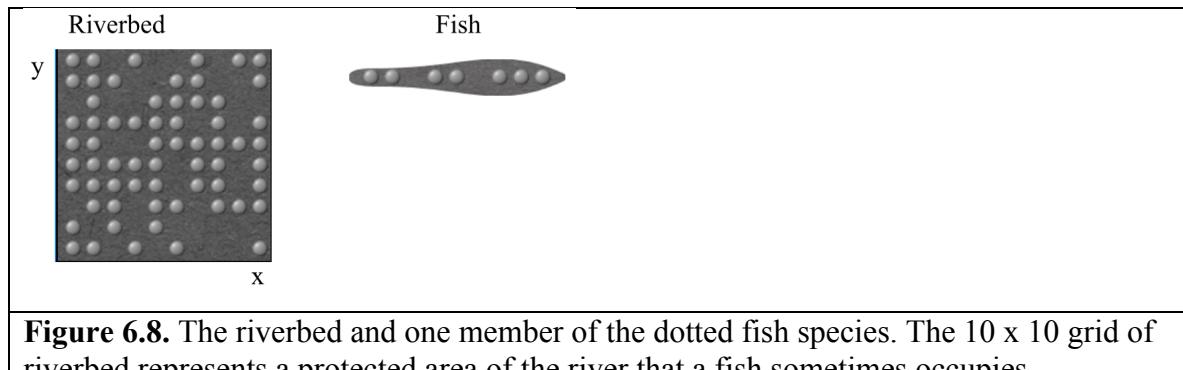
There exists a large cognitive psychology literature on visual search, with many ad hoc models. A line of work dating back to the 1950s has built probabilistic models of visual search, from the earlier pioneers (Peterson, Birdsall, Jaarsma) to later researchers (Palmer, Eckstein, Verghese, Shimozaki, Baldassi, and others). It has been shown recently (Ma, 2011) that visual

search, at least for simple stimuli, can under rather general conditions be well described using a Bayesian model.

We consider visual search tasks in which the observer has to report whether a target stimulus is present among a set of stimuli. The location at which the target will appear, if at all, is a priori unknown to the observer. For this reason, the observer has to take into account every possible location of the target. The Bayesian observer solves this by averaging over all possible target locations – in other words, by marginalizing over location. We first consider a simple animal camouflage example that involves no sensory noise. Next, we consider a laboratory visual search experiment.

### 6.3.1 Camouflage

In Chapter 1, we discussed camouflage as an evolved strategy for producing broad likelihood functions in the observer. Here we consider a toy problem inspired by the image of the flounder (Fig. 1.10 C). Suppose that a dotted fish species inhabits a river, and that a fish sometime hovers just above the pebble-covered riverbed. River fish often point upstream, as this orientation provides easier stability against the current, and faces the fish in the direction of potential food that is being swept downstream. We assume that the dotted fish takes this orientation (to the right, in Fig 6.8). We divide the area under consideration into a 10x10 grid and consider a species of dotted fish that has size 1 x 10. We assume that the fish skin color is identical to that of the riverbed mud, and that each dot on the skin closely resembles a pebble (Fig 6.8).



**Figure 6.8.** The riverbed and one member of the dotted fish species. The 10 x 10 grid of riverbed represents a protected area of the river that a fish sometimes occupies.

Generative model: suppose that each grid square on the riverbed independently has probability  $a$  of containing a pebble and  $1-a$  of not containing a pebble. For the fish species, the probability of a dot within any grid square on the skin is  $b$ , and of no dot is  $1-b$ .

We wish to determine, given a visual observation such as shown in Fig. Y, whether the fish is present ( $C=1$ ) or not ( $C=0$ ). The probability of the observation, given  $C=0$ , is:

$$p(obs|C=0) = a^n(1-a)^{n-1}, \text{ where } n \text{ is the total number of dots.}$$

The probability of the observation, given  $C=1$  is:

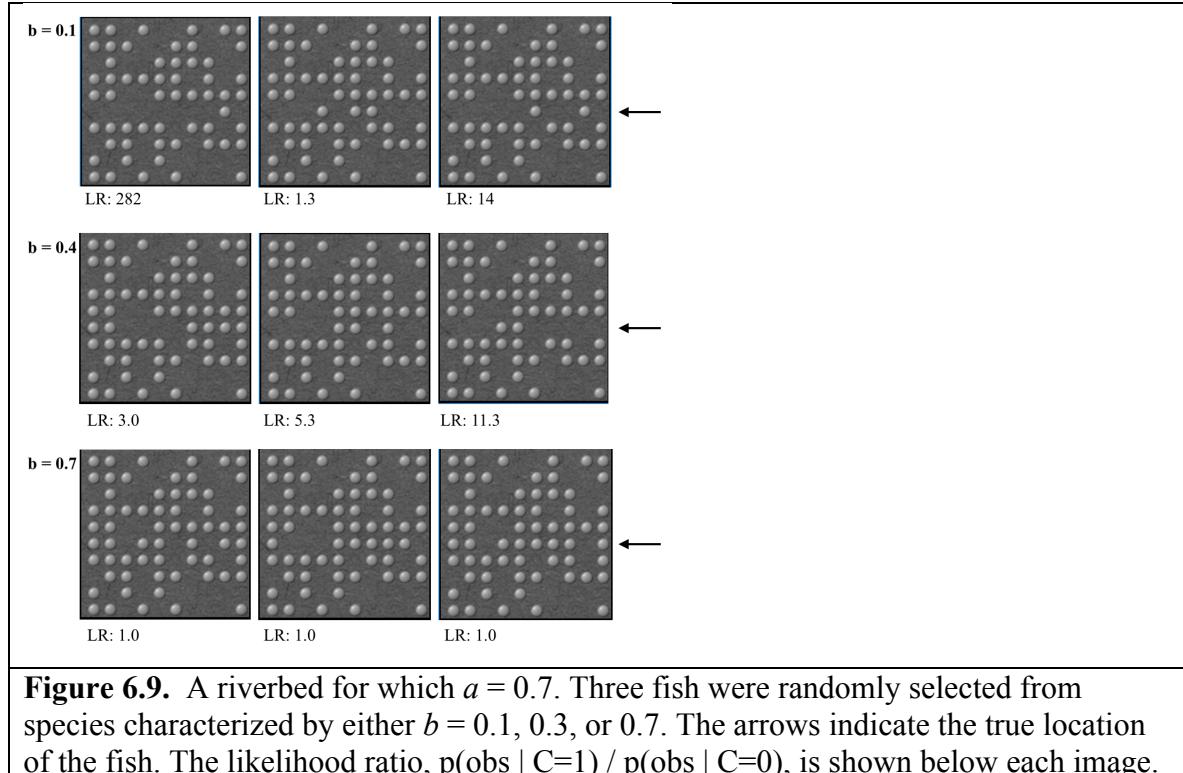
$$p(obs | C=1) = \sum_{y=1}^{10} p(obs | y, C=1) p(y | C=1)$$

will assume that, if the fish is present, it can occupy any of the 10 y-locations with equal probability. We then have:

$$p(obs | C=1) = (0.1) \sum_{y=1}^{10} b^{n_y} (1-b)^{10-n_y} a^{n_{-y}} (1-a)^{90-n_{-y}}$$

where  $n_y$  is the number of dots observed within row  $y$  of the riverbed, and  $n_{-y} = n - n_y$  is the number of dots observed within all rows other than  $y$ .

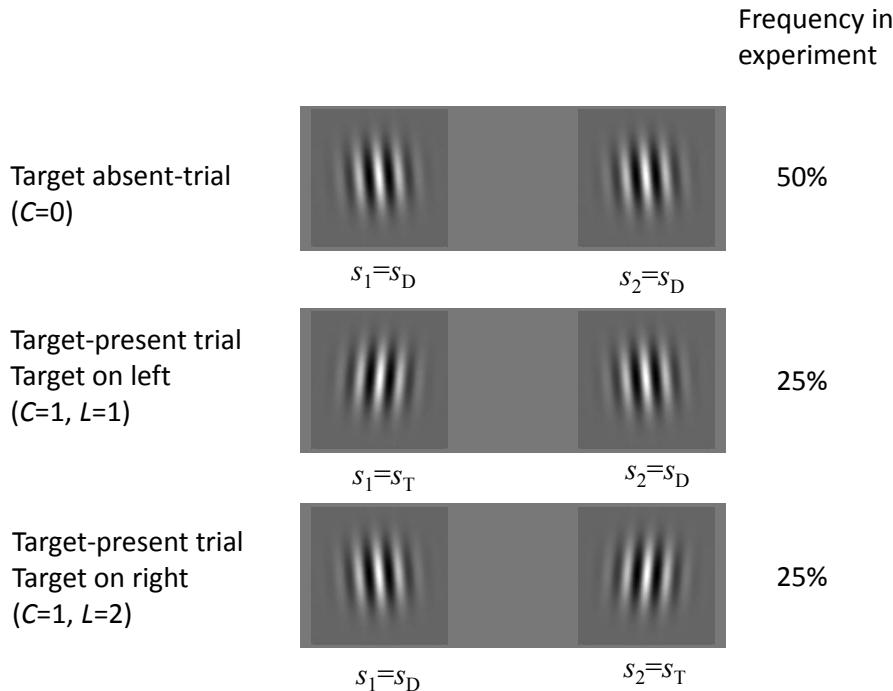
Fig. 6.9 shows the results for 3 randomly selected fish, where  $b=0.1, 0.4$ , or  $0.7$ , hovering above a riverbed characterized by  $a=0.7$ .



Within the context of this problem, we can define a perfectly camouflaged species as one for which  $b=a$ . In that case, the Eqns. show that  $p(obs | C=0) = p(obs | C=1)$ , indicating that the visual observation contributes no information regarding the presence or absence of a fish: the observer's likelihood function is flat, and the posterior,  $P(C=1 | obs)$ , will equal the prior,  $P(C=1)$  – just as if the fish were transparent. The presence of the fish becomes more apparent as the camouflage is made progressively less optimal (i.e., as  $|a-b|$  increases). It is possible to show that the visibility of a smaller fish is less affected by suboptimal camouflage (see Problems).

### 6.3.2 A visual search experiment

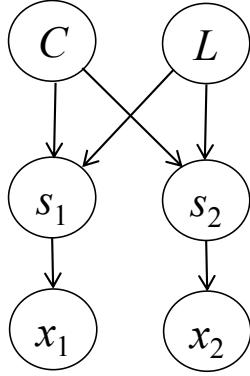
Consider a very simple search display with only two stimuli, as in Fig. 6.9. The target orientation is  $5^\circ$  tilted to the right of the vertical; we denote this as  $s_T=5^\circ$ . Any non-target stimulus – a distractor – is tilted  $5^\circ$  to the left of vertical; we denote this as  $s_D=-5^\circ$ . On each trial, the target is present with 50% probability. If it is absent, both stimuli have orientation  $s_D$ . If it is present, it can be in either location with equal probability. The figure shows the three possible combinations that can occur in this task.



**Figure 6.9.** A simple laboratory visual search task. The observer reports whether the  $5^\circ$  rightward tilted grating is present or not. Shown are the possible displays with their respective frequencies.

The observer is presented with a display like this for a brief period of time – for example, 100 ms – and asked to judge whether the target is present or not.

The generative model (Fig. 6.10) as usual assumes that the observation of each stimulus is corrupted by Gaussian noise. Since there are two stimuli, we assume the noise is independent between locations; this is a reasonable assumption when the stimuli are not close to each other on the screen. Note that, like classification, this is a hierarchical structure. As in any detection task, the state-of-the-world variable of interest is  $C$ , which can take values 0 (target absent) and 1 (target present). We denote by  $L$  the location of the target, by  $s_1$  and  $s_2$  the orientations of the stimuli on the left and right sides of the display, and by  $x_1$  and  $x_2$  their corresponding noisy observations, drawn from normal distributions with standard deviations  $\sigma_1$  and  $\sigma_2$ , respectively.



**Figure 6.10.** Generative model of the simple visual search task in Fig. 6.9. Note that when  $C=0$  (target absent),  $L$  (target location) is a meaningless variable and does not influence the distribution of  $s_1$  or  $s_2$ .

Exercise 6.5: Formulate the probability distribution associated with each node in the generative model.

On a given trial, the observer receives noisy observations,  $x_1$  and  $x_2$ . The Bayesian observer decides whether the target is present based on the log posterior ratio of target presence, which must be computed from  $x_1$  and  $x_2$ . Given that the prior over  $C$  is flat (the target is present or absent with 50% probability), the log posterior ratio is equal to the log likelihood ratio:

$$d = \log \frac{p(x_1, x_2 | C=1)}{p(x_1, x_2 | C=0)}.$$

We first examine the denominator, which is the likelihood of target absence. When the target is absent, there is only one possible stimulus combination: both orientations are equal to  $s_D$ . Since noise is assumed to be independent at both locations, the denominator is the product of two probabilities:

$$p(x_1, x_2 | C=0) = p(x_1 | s_D) p(x_2 | s_D). \quad (6.2)$$

Now we examine the numerator – the likelihood that the target is present. As we saw in Figure 6.9, there are two possibilities when the target is present: it appears either in the left location ( $L=1$ ) or in the right location ( $L=2$ ), with equal probability. The Bayesian observer averages – marginalizes – over both possibilities. In one case ( $L=1$ ), the probability of the observations given the stimuli is the product of the probability that  $x_1$  is generated by  $s_T$  and  $x_2$  by  $s_D$ , and in the other case ( $L=2$ ) it is the other way round. Averaging over both possibilities gives us the likelihood of target presence:

$$\begin{aligned}
p(x_1, x_2 | C=1) &= p(x_1, x_2 | L=1, C=1) p(L=1 | C=1) + p(x_1, x_2 | L=2, C=1) p(L=2 | C=1) \\
&= \frac{1}{2} p(x_1 | s_T) p(x_2 | s_D) + \frac{1}{2} p(x_1 | s_D) p(x_2 | s_T)
\end{aligned} \tag{6.2}$$

Now we are ready to compute the log likelihood ratio,  $d$ . We do this by substituting the two individual likelihoods and simplifying a bit:

$$d = \log \left( \frac{1}{2} \frac{p(x_1 | s_T)}{p(x_1 | s_D)} + \frac{1}{2} \frac{p(x_2 | s_T)}{p(x_2 | s_D)} \right). \tag{6.2}$$

Exercise 6.6: Show this.

The expression inside the parentheses is the average of two likelihood ratios, one associated with each location. We could call these the “local” likelihood ratios. These likelihoods might look familiar, as they are exactly the ones we encountered in Section 5.4, when we discussed discrimination between two stimulus values. Here, those stimulus values are  $s_T$  and  $s_D$ . Thus, the intuition of Eq. (6.2) is that in this particular search task, the evidence that the target is present anywhere in the display is computed from the evidence that the target is present at each individual location. This is not true for any search task; it follows from the statistical structure of the particular task (the generative model).

Having recognized the likelihood ratios from the discrimination task, we can simplify them by substituting the expressions we found in Section 5.4 (first exponentiate both sides). This gives our final expression for the decision variable:

$$d = \log \left( \frac{1}{2} e^{\frac{s_T - s_D}{\sigma_1^2} \left( x_1 - \frac{s_T + s_D}{2} \right)} + \frac{1}{2} e^{\frac{s_T - s_D}{\sigma_2^2} \left( x_2 - \frac{s_T + s_D}{2} \right)} \right). \tag{6.2}$$

This expression cannot be simplified any further in general, but we can substitute the specific values  $s_T=5^\circ$  and  $s_D=-5^\circ$ :

$$d = \log \left( \frac{1}{2} e^{\frac{10x_1}{\sigma_1^2}} + \frac{1}{2} e^{\frac{10x_2}{\sigma_2^2}} \right).$$

Unlike the simple discrimination or detection discussed in Chapter 5, this log posterior ratio is a very nonlinear function of the two observations,  $x_1$  and  $x_2$ . We could not have guessed

beforehand that the optimal strategy in this task involves computing this particular combination of the observations. Maximum-a-posteriori estimation amounts to reporting “target present” when  $d$  is positive and “target absent” when  $d$  is negative.

Looking back at the inference computation, we notice that the essential ingredient that we had not encountered in tasks discussed earlier in this chapter is the marginalization over location, Eq. (6.2). Nevertheless, the essence of the marginalization procedure is the same: the Bayesian observer always averages over the possible values of unknown variables.

In discussing cue combination and combining a cue with a prior, we emphasized the importance of weighting observations by their reliabilities. Here, we see the same weighting in a different form: the inverse of the variance of the noise appears in the local likelihood ratio, through  $\frac{s_T - s_D}{\sigma_i^2} \left( x_i - \frac{s_T + s_D}{2} \right)$ . The interpretation of this weighting is that if the distance of  $x_i$  to

the midpoint between target and distractor orientations,  $(s_T + s_D)/2$ , is the same at two locations, the location that comes with the highest reliability (lowest  $\sigma_i$ ) will contribute most towards the overall decision about target presence. In the extreme case that  $\sigma_i$  is infinitely large at one location, the observation at that location does not contribute at all to the decision.

Exercise 6.7: Verify using the final expression for  $d$  above, that when  $\sigma_1$  is infinite, the MAP decision rule  $d>0$  reduces to a discrimination task at the second location.

Thus, we see another illustration of the fact that Bayesian observers weight observations by their reliabilities, now in a more complex task.

As the last step in Bayesian modeling, we characterize the behavior of a Bayesian observer across many trials. The distribution of the MAP estimate consists of the probability that the observer will report “target present” when the target is really present, and when it is absent – in other words, the hit and false-alarm rates. These rates are the probabilities that  $d$  is positive when  $x$  is generated by  $C=1$  and  $C=0$ , respectively. More generally, as in Chapter 5, we can compute a receiver-operating characteristic curve by calculating the probabilities that  $d>k$  when  $x$  is generated by  $C=0$  and by  $C=1$ , for a large range of possible values of  $k$ , and plotting these two probabilities against one other. None of these probabilities can be calculated analytically, but they can be found by Monte Carlo simulation. We will do this in a Problem.

### 6.3.3 Generalizations

The task we modeled here was simple: there were only two stimuli. One of the most interesting aspects of visual search is how the hit and false-alarm rates depend on the number of stimuli, also called set size. It is straightforward to extend the derivation above to any number of stimuli. We will also do this in a problem at the end of the chapter.

We have so far assumed that a target always has value  $s_T$  and a distractor always has value  $s_D$ . Especially for the distractors, this is a very restrictive assumption. When you look for

your keys in the room, not all the other objects are identical to each other. An important generalization of this model is therefore to a situation in which the distractors can differ from one other; this is called the case of *heterogeneous* distractors. In the context of orientation stimuli, the distractors could for example be drawn independently from a uniform distribution. In the Bayesian model, the observer then has to marginalize not just over target location but also over distractor orientation. In that case, the expressions for the likelihoods of target presence and absence become more complicated. This is explored in a Problem.

Finally, visual search is not a single task but a class of tasks. These tasks have in common that the observer has to distinguish a target from distractors. We discussed here *target detection*, i.e. determining whether a target is present in a scene. A variation is *target localization*, when a target is always present but the observer decides which of the presented items was the target. A third variation is *target discrimination* (or *classification* or *estimation*), when a target is always present and the observer has to decide which of two or more possible values it has (estimation is usually reserved for reporting a value on a continuum). A great power of the Bayesian framework is that these different tasks are all minor variations of the same inference computation. We will examine some of these variations in the problems at the end of the chapter.

### 6.3.4 Approximating the Bayesian decision rule

The Bayesian integration rule for visual search, Eq. (6.2), seems a bit complicated, and modelers have often used approximations to this rule. One possible approximation is the *maximum-of-outputs* or *max rule*,

$$d = \max_i d_i. \quad (6.2)$$

In other words, the observer uses as the decision variable the largest of the local log likelihood ratios. It can be seen as follows that Eq. (6.2) is an approximation of the Bayesian integration rule: assume that one  $d_i$  is much bigger than all others, for example,  $d_1$  is much bigger than  $d_2, \dots, d_N$ . Then the exponentiated version of  $d_1$  is much much bigger than  $\exp(d_2), \dots, \exp(d_N)$ . The sum in Eq. (6.2) can then be approximated by its largest value,  $\exp(d_1)$ . The outcome of Eq. (6.2) is then  $d \approx d_1 - \log N$ . Since we arbitrarily assumed that  $d_1$  was the largest, the general approximation is  $d \approx \max_i d_i - \log N$ . The final step is that if  $d$  is compared to a varying criterion  $k$ , the constant term  $-\log N$  is irrelevant, so the decision variable is  $d \approx \max_i d_i$ .

When applied specifically to a fixed target and homogeneous distractors, with  $s_T > s_D$ , it can be seen that the Max rule is equivalent to

$$d = \max_i x_i, \quad (6.2)$$

Exercise 6.8: Verify this.

It has been known since Jaarsma (1966) that Eq. (6.2) is a reasonably good approximation to the full Bayesian decision rule, but only when the variance of the noise,  $\sigma_i^2$ , is identical across locations.

The Max rule has both advantages and disadvantages. Taking the largest value is more intuitive than the Bayesian integration rule. In the specific form of Eq. (6.2), the Max rule is extraordinarily simple: unlike the Bayesian rule, it does not involve  $s_T$ ,  $s_D$ , or  $\sigma_i^2$  at all. On the other hand, it is not the optimal rule and sometimes deviates considerably. An empirical test is needed to find out whether humans use the Bayes-optimal rule or the Max rule. We will comment on this in the next section.

## 6.4 Posteriors for general generative models: “following the arrows”

### “Following the arrows”: a general recipe for writing down the posterior based on the graphical model

As we are exploring generative models of increasing complexity, it would be helpful to have a clear and straightforward recipe for deriving an expression for a posterior of our choice based on the information provided by the generative model. Recall that the generative model specifies exactly which distributions are given in the problem. Each variable that does not have any arrows pointing to it follows a regular probability distribution. Each variable that does have arrows pointing to it follows a conditional distribution, where the conditioning is on the variable(s) from which the arrows originate; its distribution does not depend on any other variables in the problem. This gives us the recipe we are looking for:

1. Evaluate the joint distribution over all variables in the generative model by “following the arrows”. Start at the top with the variables that have no arrows pointing to them. Working your way down, write down the prior or conditional probability distribution associated with each node, and multiply all distributions obtained this way.
2. Compute any conditional distribution by writing out its definition and marginalizing the joint distribution (i.e. summing or integrating over the variables not in the conditional).

When a posterior over a state-of-the-world variable is computed, a marginalization is done over every variable in the generative model other than the observations and the state-of-the-world variable. Note that in this recipe, the joint distribution is central, not the likelihood or the prior. In fact, Bayes’ rule, which expresses the joint in terms of the likelihood and the prior, is simply the first step in the recipe of evaluating the joint distribution!

We saw one example in the previous chapter (Fig. 5.10a): The joint distribution is  $p(C,s,x)$ . Following the arrows, we find  $p(C,s,x) = p(C)p(s|C)p(x|s)$ . The posterior we are interested in is  $p(C|x)$ , which we obtain from marginalization:

$$p(C|x) \propto p(C,x) = \sum_s p(C,s,x) = \sum_s p(C)p(s|C)p(x|s) = p(C) \sum_s p(s|C)p(x|s) \quad (6.3)$$

If the generative model looks like a “string” of variables, each of which only receives an arrow from the previous one, such as here  $C \rightarrow s \rightarrow x$ , it is called a *Markov chain*. We will see other Markov chains in Chapter 8.

In the example of Section 6.2.1, the joint distribution is  $p(C,s,\theta)$ . Following the arrows, we find  $p(C,s,\theta) = p(C)p(\theta)p(s|C,\theta)$ . The posterior we are interested in is  $p(C|s)$ , which we obtain from marginalization:

$$\begin{aligned} p(C|s) &= \frac{p(C,s)}{p(s)} \\ &\propto p(C,s) \\ &= \sum_{\theta} p(C,s,\theta) \\ &= \sum_{\theta} p(C)p(\theta)p(s|C,\theta) \\ &= p(C) \sum_{\theta} p(\theta)p(s|C,\theta) \end{aligned} \quad (6.4)$$

which is equivalent to Eq. (6.2). We have used all sums here for convenience, but depending on the nature of the variables in an actual problem, some or all of the sums might have to be replaced by integrals.

We also ignored the  $p(s)$  factor in the denominator of the first line, since it acts, as usual, as a normalization: it is equal to the sum of the final expression over  $C$ . The proportionality sign is understood as saying “what we are going to calculate is the unnormalized posterior distribution; normalize at the end if necessary”. When the MAP estimate is the only quantity of interest, normalization is typically not even necessary, since the normalization factor does not change the MAP estimate. *Any* conditional distribution is proportional to the joint distribution of the variables on both sides of the “|” sign, but it should be kept in mind that the final normalization is over the variable(s) before the “|” sign. So  $p(C|s)$  is proportional to  $p(C,s)$  but must be normalized over  $C$  in the end, and  $p(s|C)$  is also proportional to  $p(C,s)$  but must be normalized over  $s$  in the end.

## 6.5 Applications

The tasks described in this chapter have all been used in psychophysics experiments and Bayesian models have been used successfully to describe their results.

The classification task with overlapping distributions and sensory noise was studied in a recent experiment we did. Sensory noise was manipulated by varying the contrast of the stimulus. At larger values of sensory noise, the decision criterion of the Bayesian observer, stated

in Eq. (6.2) when there is no noise, will move outward. It was found that both humans and monkeys are close to Bayes-optimal in this task. This means that they adjust their criterion based on the noise level, and resulting reliability, on each individual trial.

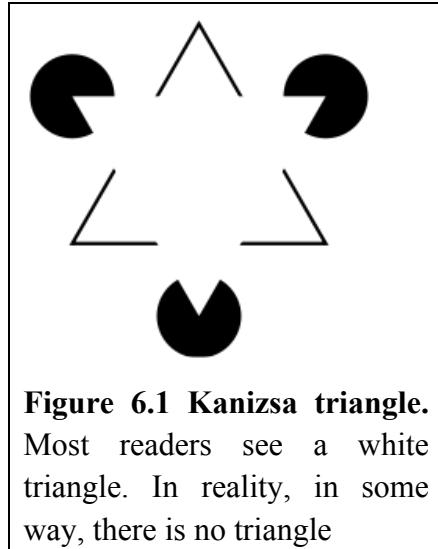
Ma et al. (2011) showed that humans are near-optimal in visual search both with homogeneous and heterogeneous distractors. The theory for heterogeneous distractors will be worked out in a Problem. An important part of optimality is that observations are weighted by their reliabilities, as expressed by Eq. (6.2) (or its analog for heterogeneous distractors). This study tested this prediction by varying either the contrast or the shape of the oriented stimuli unpredictably from item to item and from trial to trial. It was found that human observers do take reliability into account in the way predicted by the Bayesian model. The authors were also able to distinguish the Bayesian integration rule from the Max rule in both its versions (Eq. (6.2) and (6.2)). Humans seem to be using the Bayesian decision rule, not the Max rule.

Rigorously testing whether human observers are Bayes-optimal in more complex generative models is still a growing direction in the field. While most work has been done on visual search, many other interesting tasks remain unexplored from a Bayesian point of view. Certainly, much more work can be done in this area. An important general question is at what level of complexity of the generative model human perception ceases to be approximately optimal.

Object recognition is the ultimate example of Bayesian inference with a very complex generative model. In principle, the idea is the same as we already discussed in Chapter 2. An image is generated by an object, for example see Fig. 2.7. There are several candidate objects that could have produced this image. In this simple example, the likelihood is nonzero for some objects, whereas it is 0 for other objects. There is also a prior over objects: for example, cubes might be more common in the world than dodecahedrons. Likelihood and prior are combined to form a posterior distribution. The observer perceives the object with the highest probability.

The great obstacle to making this work in real scenes is the likelihood function. The visual image generated by an object is influenced not only by the object's size and shape, but as discussed in Section 6.2.1, by many nuisance parameters, such as the viewing angle, the lighting conditions, and potential occluders. On top of that, each object class, for example "dog", is characterized by a probability distribution in a very high-dimensional feature space, since dogs vary in body size, limb length, head size, spine angle, texture, color, etc. In the machine vision literature, the complexity of the object recognition problem is well known, since machines can recognize objects in visual scenes only in very restricted settings: even determining whether there is a human being in a video is far from trivial. In spite of this complexity, significant work has been done in Bayesian modeling of human object recognition.

## 6.6 Bayesian view of structure perception



Extracting information about structure in the world from sensory input is an important part, if not the ultimate goal, of perception. The world is highly structured: the shape of an object consists of a string of small line elements, objects are ordered in depth, and a musical piece consists of delicately arranged sequences of tones. As the famous Kanizsa illusion shows (Fig 7.1), the brain perceives visual structure even if there is only indirect sensory evidence for its existence. Indeed, at some level, all of our perception of structure is based on indirect information. When we walk on a busy street, our brain seemingly effortlessly separates the multiple sound sources from a single continuous stream. Arguably, object recognition, whether visual or auditory, consists of the detection of structures at multiple levels.

Since the birth of cognitive psychology, its practitioners have been intrigued by the ways in which the brain perceives structure among sets of constituent elements. For the most part, the explanations given for structure perception phenomena have been qualitative and descriptive. Most notable are the so-called Gestalt laws, which describe under which circumstances elements are perceived as belonging to a whole (as in the Kanizsa illusion). However, the Bayesian framework can in many cases provide a more principled and precise account of structure perception, and that is the focus of this chapter.

The key idea of Bayesian models of structure perception is that the state of the world of interest is a *relation* between stimuli: for example, are a set of stimuli the same or not, or do two line elements belong to a single contour? In previous chapters, we have focused on estimating or categorizing a single physical stimulus. In Chapter 6, we made a first attempt to move beyond this in our treatment of a visual search task. Visual search can be thought of as categorizing a collection of objects. In fact, the particular search task we described could be solved by answering the question “Are all elements identical to one another?” In that sense, it was a structure perception task. However, in general, visual search is not considered structure perception. To see why, consider the case that the target is always vertical but distractors are drawn from a uniform distribution over orientation space (they can have any orientation). Then

there is no relation between the objects that can distinguish the categories. The categories are then only defined in terms of the presence of the target object.

In this chapter, we examine three fundamental visual structure perception problems. We first discuss the problem of whether a set of stimuli are the same or different. This problem naturally comes with an extension of cue combination as discussed in Chapter 4. We then model how the visual system should decide whether two line elements belong to a single contour. Finally, we illustrate how Bayesian modeling can provide a rational explanation for a typical Gestalt law.

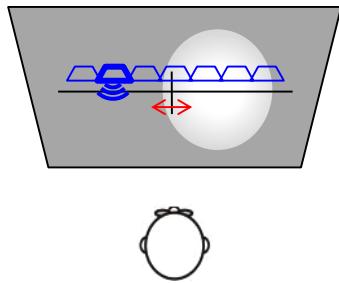
Structure perception often requires marginalization. The reason is that a relation between stimuli is generally not directly observed. However, we have observations of the stimuli. So if we want to say something about the relation, we need to consider all the possible values of the stimuli and, hence, marginalize. We will therefore make frequent use of the techniques encountered in the previous chapter.

Structure perception and ambiguity are closely linked...

## 6.7 Structure perception: causal inference

In cue combination, as discussed in Chapter 4, the observer is given two observations and the knowledge that they share the same source, or cause. In the more general *causal inference* scenario, it is unclear whether the observations have the same cause or different causes. In this case, the observer has to infer the probability that there was a single cause from the observations. This probability will subsequently play a role in estimating the values of the stimuli.

Let us take the performance of a ventriloquist as an example. The ventriloquist will move the mouth of the puppet, producing a visual cue. Simultaneously, the ventriloquist will talk (without perceptibly moving his mouth), producing an auditory cue. The viewer will wonder if the speech comes out of the puppet's mouth. In other words, the viewer will wonder if speech and mouth movements share a common cause (the talking puppet) or two different causes (the performer talks while the puppet's mouth moves). Indeed, if the ventriloquist is good, this will produce a powerful illusion of the puppet talking.

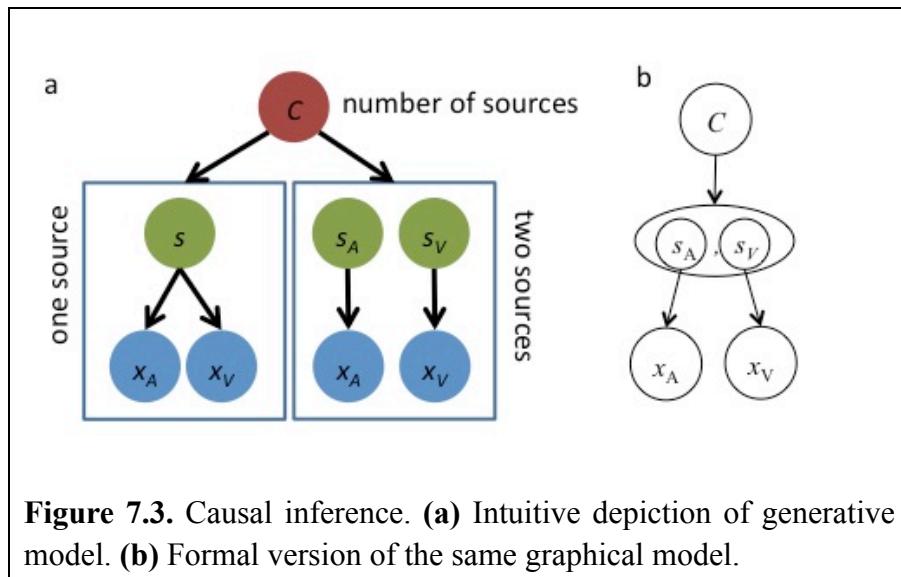


**Fig. 7.2.** Example laboratory experiment for studying causal inference. We present sounds from speakers mounted in a horizontal row behind a screen on which visual flashes are projected. The visual stimulus is a spot of light with its center on the same horizontal line as the speakers. Both auditory and visual stimuli are very brief (e.g. 30 ms). Visual reliability is manipulated through the size of the spot. Trials are either unisensory (auditory or visual) or multisensory. On multisensory trials, a visual and an auditory stimulus are presented simultaneously. In different blocks, the observer either localizes the auditory stimulus using a cursor on the horizontal meridian

(vertical black line), or reports whether the visual and auditory stimuli shared the same location (using a key press).

A more controllable example is the simultaneous presentation of an auditory tone and a visual flash (Fig. 7.2). Intuitively, when the location of the tone  $s_A$  and the location of the flash  $s_V$  are close to one another, observers may conclude they come from the same source; when the two stimuli are farther apart, the conclusion may be that they come from different sources. Using this as an example, we will see how a Bayesian model will come to the same conclusion. The causal inference model was developed in 2007 by two of the authors of this book and collaborators (Kording et al., 2007) as well as by an independent group (Sato, Toyoizumi, & Aihara, 2007).

### 6.7.1 Generative model



**Figure 7.3.** Causal inference. **(a)** Intuitive depiction of generative model. **(b)** Formal version of the same graphical model.

The generative model of causal inference can intuitively be drawn as in Fig. 7.3a. We start with the cause variable,  $C$ . If  $C=1$ , then there is a single cause: both auditory and visual cues are drawn relative to this one event. If  $C=2$ , then there are two independent causes: auditory and visual cues are drawn from independent distributions. An intuitive way of depicting this is by drawing a diagram depicting the switch in generative models. There are two “branches” to this diagram, one corresponding to the value  $C=1$ , and one to  $C=2$ . When  $C=1$ ,  $s$  is the stimulus variable of the single cause; when  $C=2$ ,  $s_A$  and  $s_V$  are the stimulus variables. The two observations are denoted  $x_A$  and  $x_V$ .

#### Box: Generative model of causal inference

Formally, the more correct way of drawing the generative model is shown in Fig. 7.3b. The two stimuli are shown in the same node, since their joint distribution is determined by  $C$ : when  $C=1$ ,

the stimuli are identical, resulting from a single draw from a prior distribution; when  $C=2$ , they are independently drawn from the prior. In equations,

$$\begin{aligned} p(s_A, s_V | C=1) &= \delta(s_A - s_V) p(s_A | C=1) \\ p(s_A, s_V | C=2) &= p(s_A | C=2) p(s_V | C=2) \end{aligned}$$

Sensory noise is described as usual:  $p(x_A|s_A)$  is Gaussian with mean  $s_A$  and variance  $\sigma_A^2$ , and analogously for the visual modality. As in regular cue combination, the noise is assumed independent between the two observations, so  $p(x_A, x_V | s_A, s_V) = p(x_A | s_A) p(x_V | s_V)$ . This generative model thus defines the joint probability distribution across stimuli, observations, and the causality variable  $C$ .

### 6.7.2 Inference

We consider a Bayesian observer trying to infer if an auditory stimulus  $s_A$  and visual stimulus  $s_V$  are from the same place ( $C=1$ ) or not ( $C=2$ ). The posterior of interest is  $p(C|x_A, x_V)$ , the probability over  $C$ , given the auditory and visual measurements  $x_A$  and  $x_V$ . Computing this posterior is similar to the examples of classification in Chapters 5 and 6, except that now there are two observations and two stimuli that must be marginalized over. The likelihood over class is given by:

$$\begin{aligned} p(x_A, x_V | C) &= \iint p(x_A, x_V | s_A, s_V) p(s_A, s_V | C) ds_A ds_V \\ &= \iint p(x_A | s_A) p(x_V | s_V) p(s_A, s_V | C) ds_A ds_V. \end{aligned}$$

For  $C=1$ , this becomes

$$\begin{aligned} p(x_1, x_2 | C=1) &= \iint p(x_1 | s_1) p(x_2 | s_2) \delta(s_1 - s_2) p(s_1 | C=1) ds_1 ds_2 \\ &= \int p(x_1 | s) p(x_2 | s) p(s | C=1) ds \end{aligned}$$

For  $C=2$ , this becomes

$$\begin{aligned} p(x_A, x_V | C=2) &= \iint p(x_A | s_A) p(x_V | s_V) p(s_A | C=2) p(s_V | C=2) ds_A ds_V \\ &= \int p(x_A | s_A) p(s_A | C=2) ds_A \int p(x_V | s_V) p(s_V | C=2) ds_V \end{aligned}$$

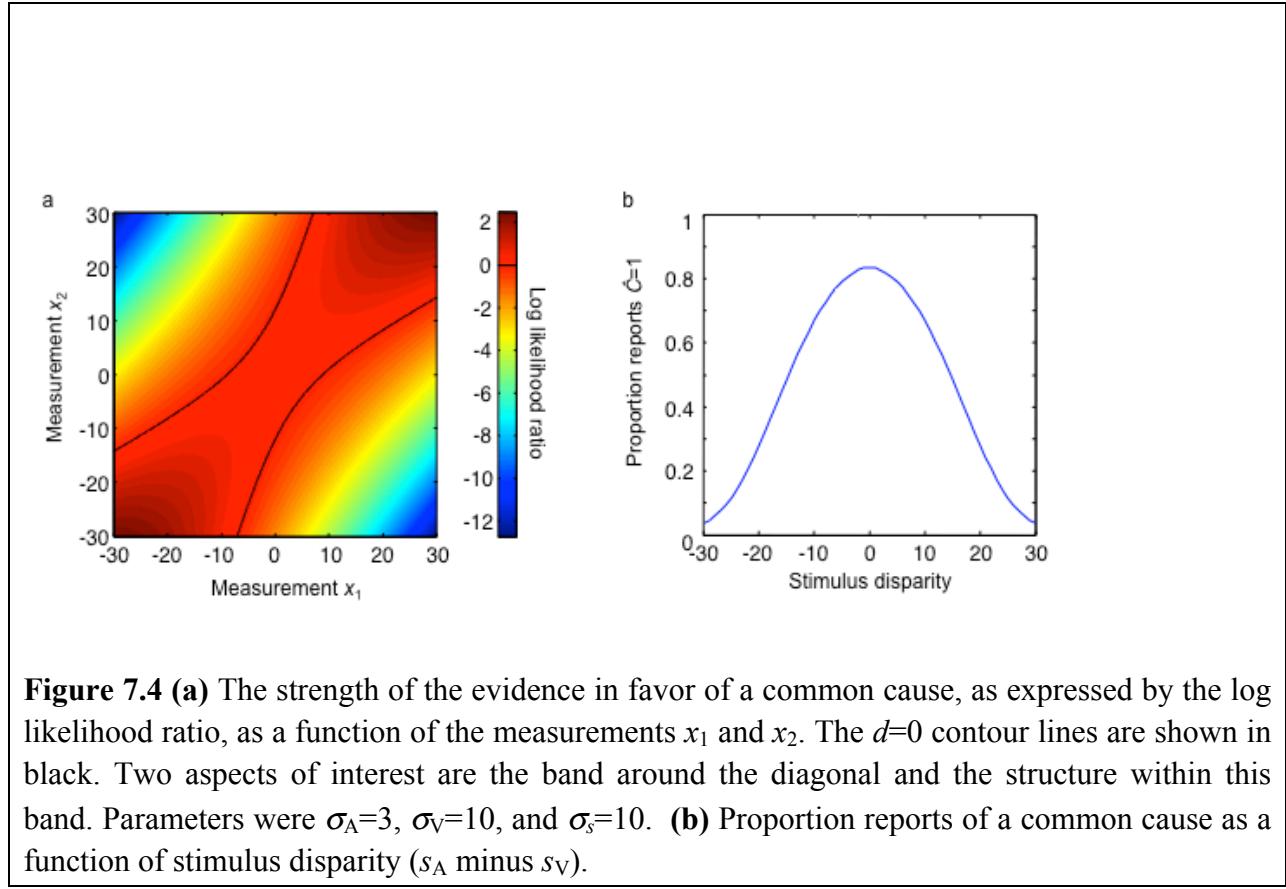
These can now be evaluated under further assumptions, for example that  $p(s|C=1)$ ,  $p(s_A|C=2)$ , and  $p(s_V|C=2)$  are all Gaussian with mean 0 and variance  $\sigma_s^2$ .

Exercise 7.1: Show that under these assumptions, the log likelihood ratio is equal to

$$d = -\frac{1}{2} \frac{J_A J_V}{J_A + J_V + J_s} \left( \frac{J_A}{J_A + J_s} x_A^2 + \frac{J_V}{J_V + J_s} x_V^2 - 2x_A x_V \right) + \frac{1}{2} \log \left( 1 + \frac{J_A J_V}{J_s (J_A + J_V + J_s)} \right),$$

where  $J_A = 1/\sigma_A^2$ ,  $J_V = 1/\sigma_V^2$ , and  $J_s = 1/\sigma_s^2$ .

When faced with a complicated expression, it is always useful to try plotting and interpreting it. This is useful both to detect mistakes and to gain an intuition for the equation.



**Figure 7.4 (a)** The strength of the evidence in favor of a common cause, as expressed by the log likelihood ratio, as a function of the measurements  $x_1$  and  $x_2$ . The  $d=0$  contour lines are shown in black. Two aspects of interest are the band around the diagonal and the structure within this band. Parameters were  $\sigma_A=3$ ,  $\sigma_V=10$ , and  $\sigma_s=10$ . **(b)** Proportion reports of a common cause as a function of stimulus disparity ( $s_A$  minus  $s_V$ ).

In Fig. 7.4, we plotted the log likelihood ratio as a color code against the observations,  $x_1$  and  $x_2$ . We chose  $\sigma_A=3$ ,  $\sigma_V=10$ , and  $\sigma_s=10$ . The diagonal corresponds to trials on which the observations happen to be identical to each other. Generally speaking, the hypothesis  $C=1$  becomes more likely relative to  $C=2$  the closer to the diagonal a set of observations lies. This is intuitive: when two observations are similar, they are likely to have a single cause. Moreover, the farther from 0 such a pair of observations lies, the more likely they have a single cause. This is because we chose a prior that peaks at the origin. Since stimuli are drawn from this prior, even when the causes are different, the two stimuli and therefore the two internal representations tend to lie close to each other near 0. When internal representations lie close to each other but far

from 0, this is harder to explain away as a consequence of the prior, and it is therefore more likely that there was truly a single cause.

### 6.7.3 Estimate distribution and Monte Carlo simulation

Across many trials (Step 3 in Bayesian modeling), we would compute the probability for the MAP observer to report  $\hat{C}=1$  as a function of the true stimuli. This is equal to the probability that  $d$  is positive when  $x_A$  and  $x_V$  are generated by  $s_A$  and  $s_V$ , respectively. After computing the observer's response probabilities as predicted by the model, the experimenter can compare them with empirical data. Unfortunately, these probabilities cannot be easily calculated analytically. In all Bayesian models discussed so far in this book, the estimate distribution (or in the case of a binary state-of-the-world variable, hit and false-alarm rates), could be calculated analytically or expressed in terms of a standard non-elementary function (the error function).

All is not lost, however. The absence of an analytical expression for the estimate distribution is the rule rather than the exception in Bayesian models of perception and action. When no analytical expression is available, numerical simulation is needed. In a simulation, we randomly draw (for instance in Matlab) a large number of pairs of observations  $(x_A, x_V)$  from their respective distributions given  $s_A$  and  $s_V$ . Each pair represents a simulated trial. For each simulated trial, we determine whether or not  $d$  is positive. The fraction of the simulated trials for which the answer is yes is an approximation of the probability that  $d$  is positive given the stimuli. This technique, of approximating a probability distribution by its samples, is a specific case of a method called *Monte Carlo simulation*. In a sense, a Monte Carlo simulation creates an “empirical” distribution, but one from a computer subject. Problem 4 in Chapter 4 (Simulating the ROC in a detection problem) already featured Monte Carlo simulation, even though it was not necessary there (an analytical expression was available).

We performed Monte Carlo simulation for the case when the true stimulus  $s_V$  is equal to 0 and the prior over  $C$  is flat. The resulting probability of reporting that both cues came from the same position:  $\hat{C}=1$  is plotted as a function of  $s_A$  in Fig. 6.7b. We see that the larger the spatial disparity between the two stimuli, the less frequently the observer reports that there is a common cause.

### 6.7.4 Posterior distribution over the stimulus

So far we have discussed inference of the number of causes. One might also be interested in the posterior distribution over  $s_A$  and  $s_V$ , the stimulus values. This posterior can be written as

$$p(s_A, s_V | x_A, x_V) \propto p(x_A | s_A) p(x_V | s_V) p(s_A, s_V). \quad (7.5)$$

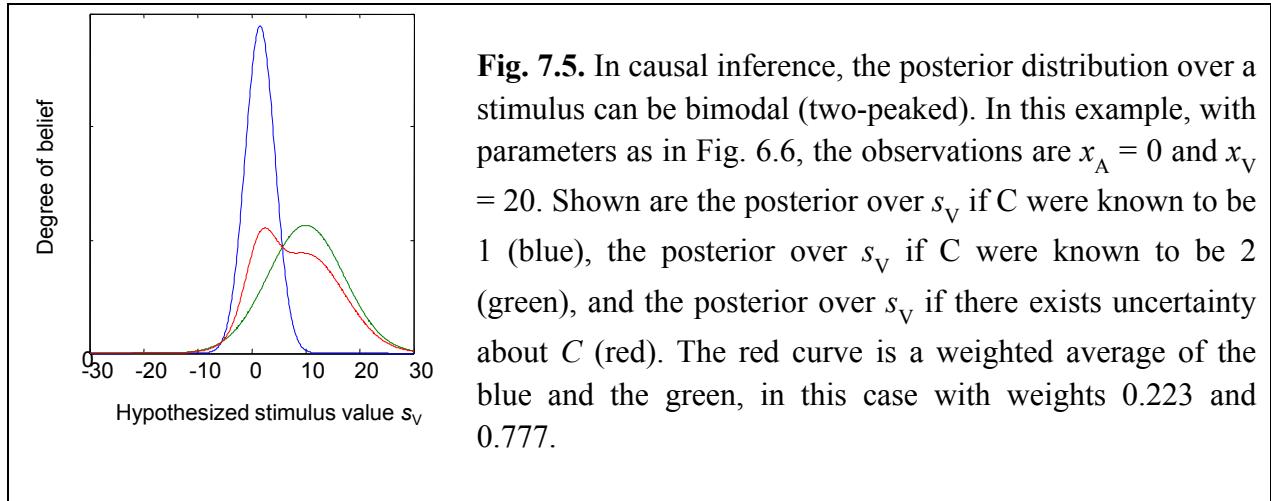
So far, nothing special; we could have done this in Chapter 4. However, in the current generative model, the prior  $p(s_A, s_V)$  is not directly available. All we know is the distribution of  $s_A$  and  $s_V$  conditioned on the number of causes,  $C$ . To find the “prior”  $p(s_A, s_V)$ , we marginalize over  $C$ :

$$p(s_A, s_V) = \sum_C p(s_A, s_V | C) p(C).$$

Substituting in Eq. (7.5), we find

$$\begin{aligned} p(s_A, s_V | x_A, x_V) &\propto p(x_A | s_A) p(x_V | s_V) \sum_C p(s_A, s_V | C) p(C) \\ &= \sum_C p(C) p(s_A, s_V | C) p(x_A | s_A) p(x_V | s_V) \end{aligned}$$

Thus, this posterior before normalization is a weighted average of the likelihood function under the hypothesis  $C=1$ ,  $p(x_A|s_A)p(x_V|s_V)p(s_A,s_V|C=1)$ , and the likelihood function under the hypothesis  $C=2$ ,  $p(x_A|s_A)p(x_V|s_V)p(s_A,s_V|C=2)$ . These likelihoods are weighted by the prior probabilities of  $C=1$  and  $C=2$ , respectively. This type of weighted average always appears when marginalization over a discrete variable (here  $C$ ) is needed.



**Exercise 7.2:** Show that an alternative way to write the posterior over  $s_A$  and  $s_V$  is as a weighted average of the *posterior distributions* conditioned on  $C$ , with weights given by  $p(C|x_A, x_V)$ . This is pictured in Fig. 7.5. This is also known as “Bayesian model averaging”, with the understanding that each value of  $C$  is interpreted as a “model” the observer has about the world.

This is the first time in this book that we encounter a posterior distribution that does not have a single local maximum (unimodal). This posterior has two peaks (is bimodal). For bimodal posteriors, a case can be made that MAP estimation is not the best read-out, since it will essentially ignore the lower of the two peaks, even if it is of almost the same magnitude. Since the estimation method is chosen to optimize a cost function, one might wonder if there are other

cost functions that are sensible in this case. We will explore this in Chapter 9, on cost and reward.

Causal inference is an important generalization of cue combination. Kording et al. (2007) showed that the causal inference accurately describes human data in an auditory-visual localization task. When observers are asked to report whether both stimuli have the same cause, their reports follow the prediction illustrated in Fig. 7.4b. Observers also report the location of the auditory stimulus according to the posterior shown in Fig. 7.5. In fact, it seems they report the mean of this distribution instead of the MAP estimate.

## 6.8 Structure perception: sameness judgment

So far we have discussed judging whether two stimuli are the same or not. This idea can immediately be extended to any number of stimuli, say  $N$ . When the stimuli are the same ( $C=1$ ), their common value  $s$  is drawn from a distribution  $p(s)$ . When the stimuli are different ( $C=2$ ), their values  $s_i$ , where the index  $i$  now takes values from 1 to  $N$ , are drawn independently from the same distribution  $p(s)$ . William James, one of the founding fathers of psychology, called the sense of sameness “the keel and backbone of our thinking”. Judging sameness plays a role in recognizing textures, which tend to consist of elements with the same orientation. Judging sameness is also said to underlie higher cognitive concepts, such as equality and equivalence in mathematics. Many animal species, from honeybees to pigeons to dolphins, can detect sameness at a rather abstract level, suggesting that the concept has had substantial evolutionary importance.

As a concrete example, we consider the case that  $p(s)$  is Gaussian with mean 0 and standard deviation  $\sigma_s$ . An example of a sameness judgment experiment is shown in Fig. 7.6a. Note that the stimuli can have different elongation, chosen randomly. Here, elongation controls the quality of the orientation information: the more elongated the ellipse, the lower the orientation noise will be.

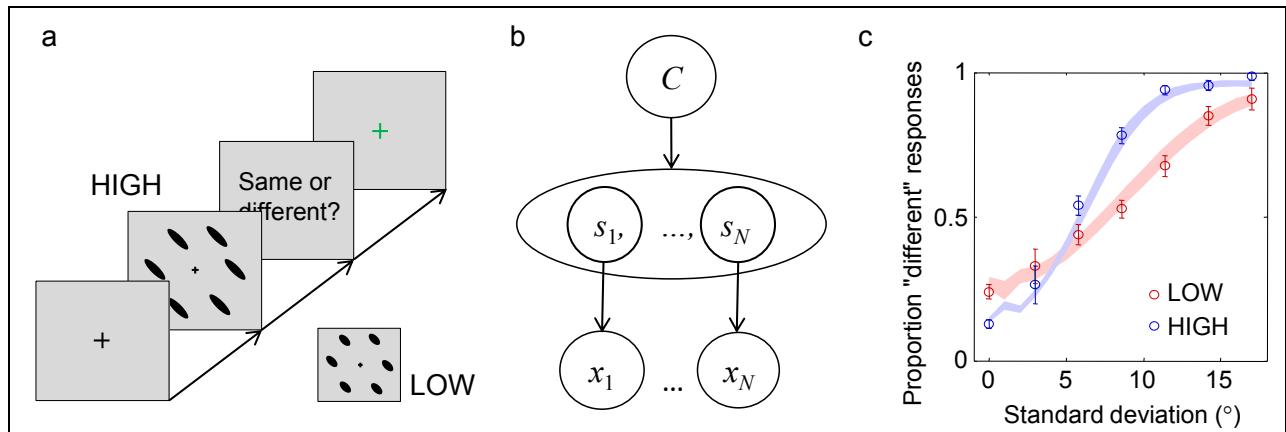


Fig. 7.6. Sameness judgment (Van den Berg, Vogel, Josic, & Ma, 2011). (a) Experimental procedure. Subjects fixated at the cross and a display containing 6 ellipses was shown for 100 ms. Stimulus reliability was controlled by ellipse elongation. In the LOW condition, all ellipses

had low reliability. In the HIGH condition, all had high reliability. (b) Generative model of this task. (c) Proportion “different” responses as a function of the standard deviation of the presented set, for the three conditions. Shaded areas indicate the fit of the Bayesian model.

The generative model of the task is shown in Fig. 7.6b. The variables are as follows:  $C$  is a binary variable that denotes sameness (1 for same, 0 for different),  $\mathbf{s}$  denotes the vector of  $N$  orientations presented, and  $\mathbf{x}$  denotes the corresponding vector of  $N$  measurements.. Each  $x_i$  is drawn from a Gaussian distribution with mean  $s_i$  and standard deviation  $\sigma$ . We refer to  $1/\sigma^2$  as the precision of each observation.

The Bayesian observer bases the decision (same or different) on the posterior probability distribution over  $C$  given the measurements  $\mathbf{x}=(x_1,\dots,x_N)^T$ . Since  $C$  is a binary random variable, we express that posterior as a log posterior ratio:

$$d = \log \frac{p(C=1|\mathbf{x})}{p(C=0|\mathbf{x})} = \log \frac{p(\mathbf{x}|C=1)}{p(\mathbf{x}|C=0)} + \log \frac{p(C=1)}{p(C=0)}. \quad (7.6)$$

Evaluating the likelihoods in this expression,  $p(\mathbf{x}|C)$ , requires marginalization over the stimulus orientations,  $\mathbf{s}=(s_1,\dots,s_N)$ :

$$p(\mathbf{x}|C) = \int p(\mathbf{x}|\mathbf{s}) p(\mathbf{s}|C) d\mathbf{s}. \quad (7.7)$$

As usual, we assume that the standard deviation,  $\sigma$ , of the noise associated with a stimulus is known to the observer for each stimulus and each trial. Therefore, we do not need to marginalize over  $\sigma$ , but can treat it as a known parameter.

When  $C=1$ , all elements of the vector  $\mathbf{s}$  have the same scalar value  $s$ . Then the integral reduces to an integral over this scalar value. Moreover, we assumed that the measurements are conditionally independent, which means that

$$p(\mathbf{x}|\mathbf{s}) = p(x_1|s) p(x_2|s) \dots p(x_N|s) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-s)^2}{2\sigma^2}\right)$$

where  $\Pi$  is notation for a product. Then the likelihood of “same” is

$$p(\mathbf{x}|C=1) = \int \left( \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-s)^2}{2\sigma^2}\right) \right) \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left(-\frac{s^2}{2\sigma_s^2}\right) ds \quad (7.8)$$

Although this integral seems daunting, it can be evaluated using the standard equation for the product of normal distributions. We can similarly evaluate the likelihood of the hypothesis that the stimuli are different, that is,  $C=0$ . In that case, all measurements are completely independent from each other, since they do not share a common  $s$ . Thus, the integral in Eq. (7.7) is a product of integrals, one for each measurement:

$$\begin{aligned} p(\mathbf{x} | C=0) &= \prod_i \int p(x_i | s_i) p(s_i) ds_i \\ &= \prod_i \left( \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - s_i)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left(-\frac{s_i^2}{2\sigma_s^2}\right) ds_i \right) \end{aligned}$$

Using the expressions for  $p(\mathbf{x}|C=1)$  and  $p(\mathbf{x}|C=0)$ , we can evaluate the log posterior ratio in Eq. (7.6) and obtain a decision rule. The decision rule will be a quadratic function of the measurements. We do not complete the derivation here but refer to the Problems.

It has been shown that human subjects judge sameness in a way that is close to the predictions of this model. For an example of a data fit in a closely related task, consider Fig. 7.6c, which shows how often human subjects report that 6 stimuli are different, as a function of the standard deviation of the stimuli. This was done both for low stimulus reliability (high  $\sigma$ ) and high stimulus reliability (low  $\sigma$ ). The procedure for fitting a model to data is explained in Appendix 2.

This example shows how inference of a relatively abstract quality like “sameness” can easily be modeled in a Bayesian way using the exact same procedure we used for inferring a physical stimulus.

## 6.9 Structure perception: contour integration

Objects are defined in part by their boundaries. Therefore, a key part of object recognition is to identify which line elements belong to the same boundary or contour. Contour integration is an example of a task where natural statistics are likely to play an important role in shaping the prior (see Chapter 2, discussion about types of priors).

### 6.9.1 Simple contours

Feldman (Feldman, 2001) had human subjects do the task in Fig. 7.7: they judged whether five dots formed a corner or a smooth curve.

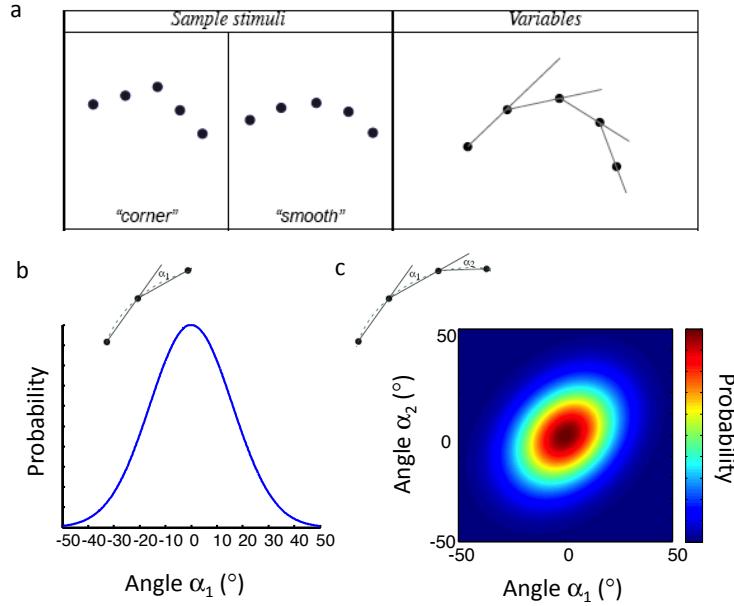


Fig. 7.7: Corner or smooth? From Feldman (2001) (a) Task: subjects judged whether or not a corner was formed by five dots (b) Assumed distribution of angle formed by three dots on a smooth contour. (c) Assumed distribution of pair of angles formed by four dots on a smooth contour. The two angles are correlated.

The top variable in the generative model is a binary variable  $C$ .  $C=1$  indicates that the corner is present,  $C=0$  that it is absent. If a corner is present, there are two smooth contours. If no corner is present, there is one smooth contour. Feldman then parameterized the probability of observing an angle  $\alpha$  among three dots given that they lie on a smooth contour, using a Gaussian distribution about 0 (Fig. 7.7b):

$$p(\alpha \mid \text{smooth contour}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\alpha^2}{2\sigma^2}}$$

Similarly, he defined the probability of a pair of angles,  $\alpha_1$  and  $\alpha_2$ , being formed by four dots on a smooth contour. This was done using a two-dimensional correlated Gaussian distribution as depicted in Fig. 7.7c. The idea is that  $\alpha_1$  and  $\alpha_2$  each have their own variance, but they might also be correlated. For example, if  $\alpha_1$  is positive (rightward deviation), then the smoothness of the contour would make it likely that  $\alpha_2$  is positive. This type of dependence can be captured by a correlation coefficient. Using these building blocks and some additional assumptions, one can compute the likelihood that a corner is present among the five dots. To compute this, one has to marginalize over the location of the corner: it could be at the second, third, or fourth dot.

### 6.9.2 Natural contours

A different approach to the same problem makes use of natural statistics to specify the generative model. Geisler and Perry (Geisler & Perry, 2009) tested the hypothesis that natural statistics combined with a Bayesian model can predict human contour integration judgments. They first formalized the problem by introducing the relevant variables (Fig. 7.8). The state of the world of interest is whether two line elements belong to the same contour. This is a binary variable:  $C=0$  (no) or  $C=1$  (yes). The stimuli are two edge elements, and we assume that only their position relative to each other matters. The parameters used to describe this relative position are shown in Fig. 7.8: distance between midpoints ( $d$ ), the angle between the reference element and the line connecting the midpoints ( $\varphi$ ), the angle between the reference element and the orientation of the other element ( $\theta$ ), and finally the contrast polarity ( $\rho$ ): if one were to connect the two elements using a contour, would which side is darker change between the elements?

The generative model of this task is described by the probability distributions  $p(d, \varphi, \theta, \rho | C=1)$  and  $p(d, \varphi, \theta, \rho | C=0)$ . The authors estimated these probabilities by analyzing natural scenes. A photograph of an outdoors, natural scene, such as leaves lying on the forest floor, was first analyzed automatically by an algorithm that extracted the edges. The image was then presented to a human observer with one pixel marked. The observer indicated which other pixels in the image belonged to the same contour. Humans were highly consistent in making these judgments. In this way, the generative model was estimated. Unlike the previous examples in this book, in this case, the generative model was purely specified numerically, that is, through histograms indicating the frequency of occurrence of every combination of parameters. Another difference with most generative models discussed so far is that sensory noise was assumed to be negligible. All uncertainty in the task derives from ambiguity.

In a subsequent experiment, different human observers judged whether two edge elements passing under an occluder belonged to the same or to different contours. “Same” and “different” each occurred 50% of the time. A Bayesian observer would make this judgment by computing the posterior ratio. When the prior is flat, reflecting the frequencies of “same” and “different”, the posterior ratio is equal to the likelihood ratio.

$$\frac{p(C=1 | d, \varphi, \theta, \rho)}{p(C=0 | d, \varphi, \theta, \rho)} = \frac{p(d, \varphi, \theta, \rho | C=1)}{p(d, \varphi, \theta, \rho | C=0)}$$

The modelers were able to make predictions for human judgments based on the generative model. An illustrative example of these predictions is shown in Fig. 7.8b. As one might expect, this shows that contours tend to be smooth. Human observers performed close to the Bayesian observer, with similar patterns of errors.

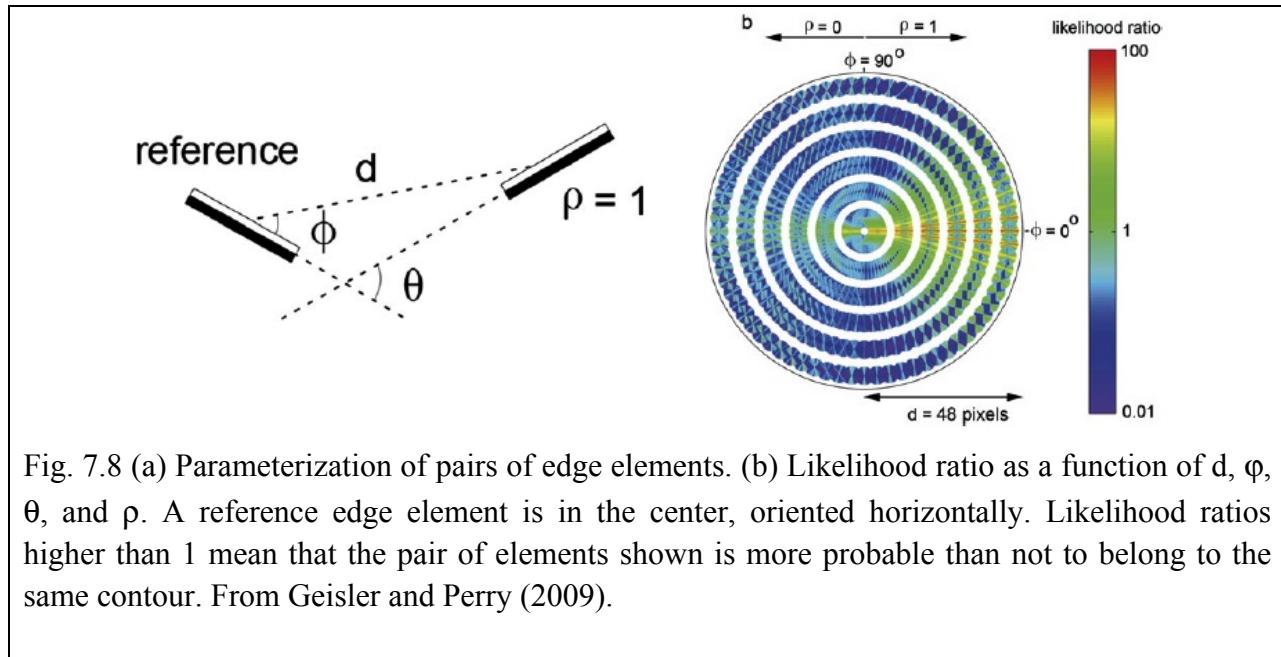


Fig. 7.8 (a) Parameterization of pairs of edge elements. (b) Likelihood ratio as a function of  $d$ ,  $\phi$ ,  $\theta$ , and  $\rho$ . A reference edge element is in the center, oriented horizontally. Likelihood ratios higher than 1 mean that the pair of elements shown is more probable than not to belong to the same contour. From Geisler and Perry (2009).

This study is closely related to the notion of “good continuation” in Gestalt psychology. This Gestalt principle says that elements that suggest a continued visual line will tend to be grouped together. By examining the statistics of natural scenes, this rather vague principle can be quantified: elements are grouped together if they have a higher probability of belonging to the same contour than not. This illustrates how Bayesian models can improve on qualitative observations in psychology.

It is instructive to compare the approaches of these two subsections. Geisler and Perry used a generative model obtained numerically from natural statistics. This is the more constrained approach, as no parametric assumptions are needed. They also modeled the elements of the contour in greater detail: elements had an orientation and a contrast polarity, instead of being dots. On the other hand, simplifying the problem to dots and using analytical expressions for the generative model allows for easier experimental manipulations and a more concise formulation of the model.

## 6.10 Structure perception: Gestalt principles

Gestalt principles or Gestalt laws have been the leading description of structure perception in psychology. We have mentioned the Gestalt principle of good continuation. There are many additional Gestalt principles, including:

- The law of closure: objects such as shapes, letters, pictures, etc., are perceived as being whole even when they are not complete (Fig. 7.9a).
- The law of similarity: elements within an assortment of objects are perceptually grouped together if they are similar to each other (Fig. 7.9b).
- The law of common fate: objects are perceived as lines that move along the smoothest path.

- The law of proximity: objects that are close to each other are perceived as forming a group (Fig. 7.9c-d).
- The law of good gestalt: objects tend to be perceptually grouped together if they form a pattern that is regular, simple and orderly.

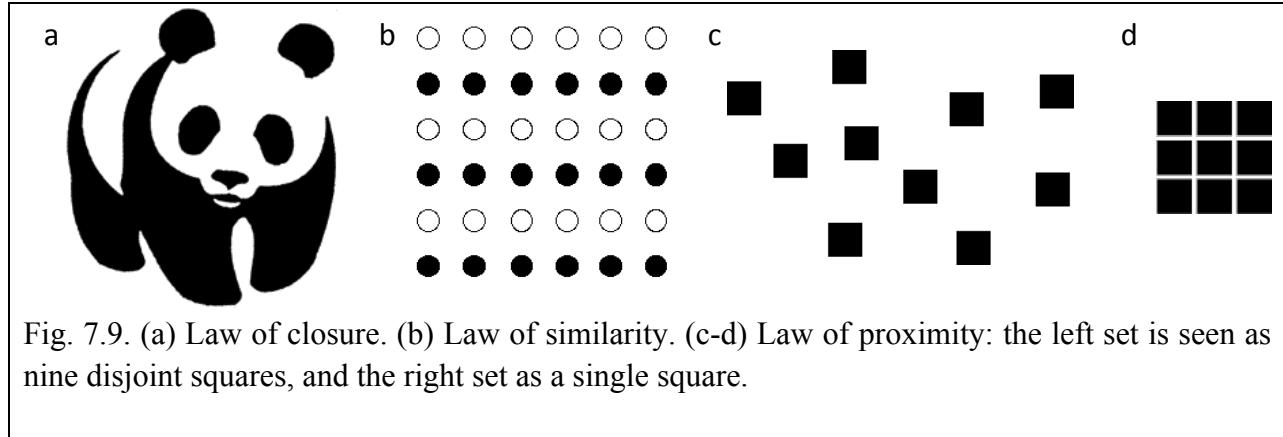


Fig. 7.9. (a) Law of closure. (b) Law of similarity. (c-d) Law of proximity: the left set is seen as nine disjoint squares, and the right set as a single square.

Almost since their conception, Gestalt laws have been criticized for being vague and descriptive, instead of quantitative and explanatory. This is especially evident in the law of good Gestalt, where “regular”, “simple”, and “orderly” are not defined. Bayesian models have the potential to improve on these laws. The basic idea in every case is that the observer considers two hypotheses – for example, the elements belong together or they don’t – and evaluates their posterior probabilities. We now examine a specific case.

Consider the picture in Fig. 7.10a. Most people will see this as two intersecting lines, instead of as two angles touching, though either interpretation is possible (see Fig. 7.10b; there are in fact more possible world states). The law of continuity would state in this case that individuals tend to perceive the two objects as two single uninterrupted entities, because elements tend to be grouped together when they are aligned.

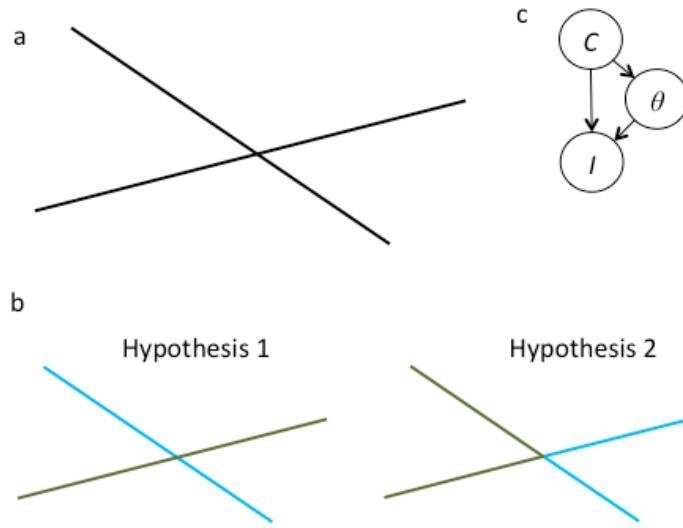


Fig. 7.10. (a) Image. (b) Two possible interpretations of this image ( $C=1$  and  $C=2$ ). (c) Generative model.  $C$  is the world state,  $\theta$  are the parameters,  $I$  is the image.

Here, we don't need to measure natural statistics to be able to make the general argument why Hypothesis 1 is more common. The generative model is shown in Fig. 7.10c. The top variable refers to the world state, corresponding to two intersecting lines ( $C=1$ ) or to two touching angles ( $C=2$ ). We have to parameterize the stimuli. We start with  $C=1$ . A line is parameterized by two numbers, as one can see by writing an equation of a line:  $y = ax + b$ . Thus, four parameters specify the two lines in Hypothesis 1. Now consider  $C=2$ . An angle is parameterized by four numbers: two coordinates for the origin of the angle, one angle for the first leg, and one angle for the second leg. Thus, eight parameters are needed for Hypothesis 2. Finally, in both interpretations, the image is uniquely determined by the parameters.

The Bayesian observer performs inference by computing the posterior ratio of both world states, based on the given image  $I$ :

$$\frac{p(C=1|I)}{p(C=0|I)} = \frac{p(I|C=1)}{p(I|C=0)} \frac{p(C=1)}{p(C=0)}$$

We cannot say much about the prior, but we can evaluate the likelihood ratio. We denote the parameters in each hypothesis by a vector  $\theta$ . Thus,  $\theta$  is four-dimensional when  $C=1$  and eight-dimensional when  $C=2$ . Since  $\theta$  acts as a nuisance parameter, each of the likelihoods is computed by marginalizing over  $\theta$ . To simplify the argument, let's assume that all parameters take on discrete values. The marginalization is then the sum:

$$p(I|C) = \int p(I|C, \mathbf{n}) p(\mathbf{n}|C) d\mathbf{n} \quad p(I|C) = \sum_{\boldsymbol{\theta}} p(I|C, \boldsymbol{\theta}) p(\boldsymbol{\theta}|C) d\boldsymbol{\theta} \quad (7.5)$$

We know that the image is uniquely determined by the parameters and the hypothesis. Therefore,  $p(I|C, \boldsymbol{\theta})$  equals 0 for all parameter combinations  $\boldsymbol{\theta}$  except the one that produces the given image  $I$ . We denote this parameter combination by  $\boldsymbol{\theta}_I$ . For this combination,  $p(I|C, \boldsymbol{\theta})$  equals 1. The integral then simply becomes

$$p(I|C) = p(\boldsymbol{\theta}_I|C)$$

All that remains now is to evaluate the probability of  $\boldsymbol{\theta}_I$  under each hypothesis. For illustration, let's assume that all parameters are independent and each parameter takes on 100 possible values (this and the following argument are due to MacKay (MacKay, 2003)). Then the probability of  $\boldsymbol{\theta}_I$  (or any other parameter combination) under hypothesis 1 is  $\left(\frac{1}{100}\right)^4$ , whereas the probability of  $\boldsymbol{\theta}_I$  (or any other parameter combination) under hypothesis 2 is  $\left(\frac{1}{100}\right)^8$ . That means that the likelihood ratio is  $\left(\frac{1}{100}\right)^4 = 10^8$ . In other words, Hypothesis 1 is 100 million times more likely than hypothesis 2.

This explains why humans observe the image as two intersecting lines rather than as two touching angles. Intuitively, the hypothesis  $C=1$  requires that the opposite angles in the image share a common vertex and be equal, whereas the hypothesis  $C=2$  permits this configuration but also a vast number of other configurations. The fact that the image conforms to the restricted features predicted by  $C=1$  therefore favors that hypothesis.

Of course, the precise numerical value of the likelihood ratio will depend on our assumptions regarding the priors over the parameters within each hypothesis. However, any Bayesian observer who begins with broad prior distributions will favor hypothesis  $C=1$  when shown the image in Fig. 7.10a. The essence of the argument is if two hypotheses can account for the observations equally well, a Bayesian observer will favor the hypothesis that has the lowest number of parameters. In this sense, Occam's razor is an emergent property of Bayesian inference: simpler models are better.

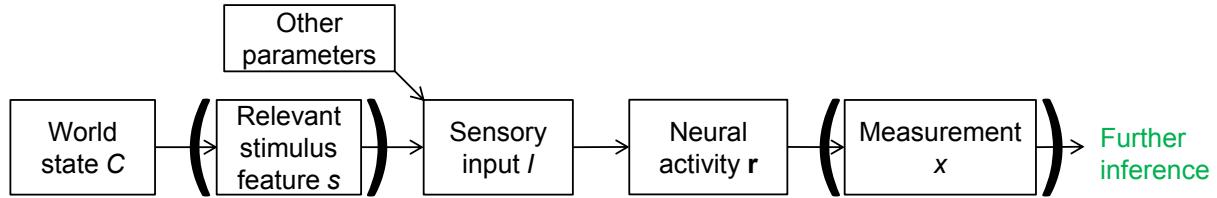
Often, more complex hypotheses (ones with more parameters) can account better for the data. A specific parameter combination within a complex hypothesis may fit the data more precisely than any parameter combination allowed by a simpler hypothesis. Thus, there is a trade-off between complexity and power. This trade-off is also captured in Eq. 7.5 above, since  $p(I|C, \boldsymbol{\theta})$  is an indication of the power of the hypothesis.

Incidentally, the Bayesian observer who selects between two perceptual hypotheses is mathematically identical to a Bayesian experimenter who analyzes data in order to select between two competing models. Thus, Bayesian model comparison follows the same equations that are here discussed to describe the human brain, including the trade-off between complexity and power. This is discussed in detail in Appendix 2.

## 6.11 Concluding remarks

By considering situations where the world state of interest is not a physical stimulus, we have extended the type of tasks we can model. We can modify the diagram in Fig. 3.1 accordingly, giving rise to Fig. 6.11. The classification tasks we discussed in Chapter 5 would be a straight extension by one variable,  $C$ , on the left side of the diagram. Stimulus and measurement are still well defined there, and the only novelty is the introduction of non-trivial class-conditioned stimulus distributions,  $p(s|C)$ .

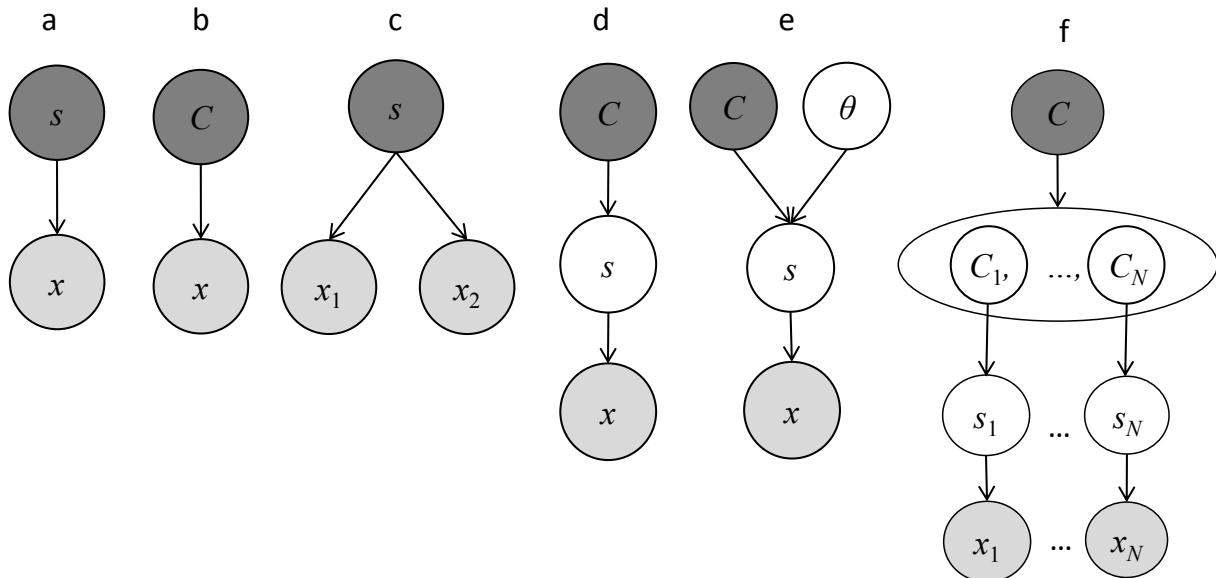
Sometimes, the notion of a relevant stimulus feature is hard to apply. For example, when determining whether an animal in an image is a dog or a cat, or whether a spoken syllable is “light” or “late”, there is no individual physical feature that can be isolated as being solely relevant. In these cases, the diagram may more sensibly be drawn without the boxes for  $s$  and  $x$ .



**Figure 6.11.** Modification of Fig. 3.1 for the more general case where a world state of interest (not necessarily a physical stimulus; can also be a class) produces sensory input. The relevant stimulus feature cannot always be defined (for example in a natural image), which is why we put it in parentheses. As a consequence, the measurement  $x$  is not always well-defined either.

In this chapter we have encountered likelihood functions over variables other than the stimulus, for example likelihoods of the form  $p(x|C)$ , where  $C$  is a higher-level variable that influences the distribution of  $s$  and that is of interest to the observer. This is where the terminology “elementary likelihood” introduced in Chapter 4 might be relevant. An elementary likelihood function is associated with each measurement. Inference consists in large part of computing the likelihood function over the world state of interest from elementary likelihood functions. Other distributions in the generative model are used in the process. For example, Eq. XX expresses the likelihood  $p(x|C)$  in terms of the elementary likelihood  $p(x|s)$  for our classification-under-ambiguity task, and Eqs. (6.2) and (6.2) do the same for visual search with homogeneous distractors.

In this chapter, we have seen the full power of the Bayesian modeling approach. It is completely normative, in the sense that one starts by deriving the *best possible* solution to the problem, and makes all assumptions explicit through the generative model. Importantly, it is possible to write down a Bayesian model of a task before even starting an experiment. Thus, the model has maximal predictive value. Another major advantage is that once the generative model has been formulated, predictions for behavior are obtained using a set recipe without any room for ad-hoc assumptions. Yet, Bayesian models also tend to be easily modifiable. For example, for visual search, we have focused on target detection with no more than one target, a fixed target stimulus, homogeneous distractors, and equal probabilities for all locations to contain the target. With minor modifications that are completely dictated by the experimental design, the model can make predictions for experiments with multiple targets, a target stimulus drawn from a distribution, heterogeneous distractors, unequal probabilities for different locations to contain the target, or even a different response paradigm, such as target localization or target estimation. Some of these variations are examined in the Problems.



**Figure 6.12.** Generative models revisited. The state-of-the-world variable of interest is shaded dark, the observation(s) are shaded light. All non-shaded variables must be marginalized over. (a) Estimation of a continuous variable based on a single observation. (b) Binary decisions. (c) Cue combination. (d) Classification. (e) Nuisance parameter. (g) Visual search.

Figure 6.12 shows all generative models we have discussed so far. Keep in mind that some of these, like “binary decisions” and “nuisance parameter”, refer to broad types of tasks, whereas the generative model of visual search was discussed in the context of a specific experiment. The latter could also be classified broadly.

There are caveats to Bayesian modeling. As we already discussed in Chapter 3, humans might not be correctly learning or incorporating knowledge of the distributions set by the experimenter. For example, even if the experimenter has set a 0.5 probability of target presence,

observers might incorrectly learn this probability as 0.55. To some extent, such incorrect assumptions can be modeled by including them as free parameters in the model. In this example, the prior probability assumed by the observer would be a free parameter and fitted to the data. Furthermore, observers must not only correctly learn the distributions of the generative model; they must also learn the structure of the generative model itself. Humans are very good at detecting patterns, and might have a tendency to believe in spurious patterns. A final caveat is that most Bayesian models don't deal with natural scenes or tasks that are realistic in natural environments. Real-life generative models are very complex and often cannot be unambiguously specified. However, one could argue that a lack of direct applicability to natural environments is a feature of much of experimental psychology in general.

## 6.12 Further reading

- Kersten, D., P. Mamassian, et al. (2004). "Object perception as Bayesian inference." *Annu Rev Psychol* 55: 271-304.
- Nolte, L. W. and D. Jaarsma (1966). "More on the detection of one of  $M$  orthogonal signals." *J Acoust Soc Am* 41(2): 497-505.
- Peterson, W. W., T. G. Birdsall, et al. (1954). "The theory of signal detectability." *Transactions IRE Profession Group on Information Theory, PGIT-4*: 171-212.
- Ma, W. J., V. Navalpakkam, et al. (2011). "Behavior and neural basis of near-optimal visual search." *Nat Neurosci* 14: 783-790.
- Feldman, J. (2001). Bayesian contour integration. *Percept Psychophys*, 63(7), 1171-1182.
- Geisler, W. S., & Perry, J. S. (2009). Contour statistics in natural images: Grouping across occlusions. *Vis Neurosci*, 26, 109-121.
- Kording, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE*, 2(9), e943.
- MacKay, D. (2003). *Information Theory, Inference and Learning Algorithms*: Cambridge University Press.
- Sato, Y., Toyoizumi, T., & Aihara, K. (2007). Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli. *Neural Computation*, 19(12), 3335-3355.
- Van den Berg, R., Vogel, M., Josic, K., & Ma, W. J. (2011). Optimal inference of sameness. *Proc Natl Acad Sci U S A*, 109(8), 3178-3183.

## 6.13 Problems

**Problem 6.1.** In Chapter 4, we considered the task of deciding whether you are holding an orange or a baseball in your hand based on weight, size, and texture. This is a classification problem with two classes. Formally write down the generative model and the inference model in this task.

**Problem 6.4: Classification with overlapping distributions and sensory noise**

An important part of perception is to classify visual information. Is this cloud a rain cloud or not? Is this approaching person my friend or not? Here, we model a simple classification task. An experimenter draws, on each trial, a stimulus  $s$  randomly from one of two classes ( $C=1$  or  $C=2$ ) with equal probability. Each class is characterized by a Gaussian distribution. These distributions have the same mean but different standard deviations,  $\sigma_1=3$  and  $\sigma_2=12$ . An observer's measurement,  $x$ , is drawn from a Gaussian distribution with mean  $s$  and standard deviation  $\sigma$ . The observer decides on each trial which class the stimulus belongs to.

- a) Plot the two distributions  $p(s|C=1)$  and  $p(s|C=2)$  using Matlab. Why can an observer not be 100% correct on this task?
- b) If there would be no noise in the measurement ( $\sigma=0$ ), what would be the optimal rule for making a decision? (No math is needed for this; instead, you can explain the rule using the plot in part (a), but be precise.)
- c) Derive the Bayesian decision rule when there is noise.
- d) Assume that the observer performs MAP estimation. What is the predicted frequency of reporting class 1 when the true stimulus is  $s$ ?
- e) Plot this frequency in Matlab as a function of  $s$  (between  $-30$  and  $30$ ) for  $\sigma=10$ .
- f) Do the same for  $\sigma=1$ . Plot both cases in the same plot. Can you explain the difference between the two?

**Problem 6.5: Camouflage and fish size.**

In the text, we derived likelihoods for the presence and absence of a fish that was 10 units long by 1 unit wide. How could you modify the approach described to treat the case of a species of smaller fish, for instance 3 units long? Hint: you will need to marginalize over both  $x$  and  $y$ . Show that the visibility of these smaller fish is less affected by suboptimal camouflage; i.e., for a given  $b$ -value that is not equal to  $a$ , the likelihood ratios (Fish compared to No-Fish) are on average closer to 1 when the fish is smaller.

**Problem 6.6: Visual search with  $N$  stimuli**

In the text, we discussed visual search with 2 stimuli, homogeneous distractors, and at most one target. Show that for  $N$  stimuli, homogeneous distractors, and at most one target, the log likelihood ratio of target presence is equal to

$$d = \log \left( \frac{1}{N} \sum_{i=1}^N e^{d_i} \right), \text{ with } d_i = \frac{s_T - s_D}{\sigma_i^2} \left( x_i - \frac{s_T + s_D}{2} \right).$$

**Problem 6.7: Visual search with heterogeneous distractors**

An observer detects whether a target, defined by orientation, is present among  $N$  line segments. The target always has orientation  $s_T$ . Each distractor orientation is drawn independently from a

uniform distribution on  $[0, \pi]$ . The observation at the  $i^{\text{th}}$  location,  $x_i$ , is drawn from a Von Mises distribution with circular mean  $s_i$  (the true orientation) and concentration parameter  $\kappa$ :

$$p(x_i | s_i) = \frac{1}{\pi I_0(\kappa)} e^{\kappa \cos 2(x_i - s_i)},$$

where  $I_0$  is the modified Bessel function of the first kind of order 0. The prior probability that the target is present is 0.5.

- a) What is the modified Bessel function doing there? Explain. (You can answer this question even if you don't know the definition of the function.)
- b) Show that in general, when the distractor is drawn from a distribution  $p(s_i | C_i = 0)$ , where  $C_i$  refers to target presence at the  $i^{\text{th}}$  location, then the MAP decision rule is the same as in Problem 6.5, but with the local log likelihood ratio equal to  $d_i = \log \frac{p(x_i | s_i)}{\int p(x_i | s_i) p(s_i | C_i = 0) ds_i}$ . In other words, the observer marginalizes over distractor orientation (in the denominator).
- c) Use this to show in our specific case that the Bayesian MAP observer responds "target present" if

$$\sum_{i=1}^N e^{\kappa \cos 2(x_i - s_T)} > NI_0(\kappa).$$

**Problem 6.8: Ponzo illusion.** You will analyze the simplified Ponzo illusion in the third panel of Fig. 6.7 in a Bayesian way. The key is to interpret the scene as truly three-dimensional, as if you are looking down on a road with two horizontal lines drawn on it.

- a) The relevant variables are true length of the upper horizontal line, its length on the retina, its vertical position, the distance it is away, the true viewing angle with respect to the ground plane, and the context (the angle that each of the two tilted lines makes with the vertical). Give these variables names and put them into a graphical model, indicating influences with arrows.
- b) The observer is interested in the posterior distribution over true length given the length on the retina, the vertical position, and the context. Work out this posterior based on the distributions in the generative model. Make reasonable assumptions as needed.
- c) Use the resulting expression to qualitatively argue why the brain perceives the upper line to be wider.

### Problem 6.9: Visual search variations

Derive the likelihood function (or log likelihood ratio) at set size  $N$  in each of the following visual search scenarios. Assume independent Gaussian or Von Mises noise.

- a) The target is always vertical. Distractors are homogeneous but their value is drawn on each trial from a uniform distribution over orientation. The observer reports whether the target is present.
- b) The target is always vertical. Distractors are drawn from a Gaussian distribution around vertical with variance  $\sigma_D^2$ . The observer reports whether the target is present. (This was done by Benjamin Vincent.)
- c) The target is always vertical. Each distractor is independently chosen to be tilted an amount  $\Delta$  to the left or to the right of vertical. The observer reports whether the target is present.
- d) Distractors are always vertical. The target is drawn on each trial from a Gaussian distribution around vertical with variance  $\sigma_T^2$ . The observer reports whether the target is present.
- e) The target is drawn from a symmetric distribution around 0. Distractors are all vertical. The observer reports whether the target is tilted to the right or left of vertical. (This task has been studied extensively by Stefano Baldassi and colleagues.)
- f) Distractors are homogeneous but their value is drawn on each trial from a uniform distribution over orientation. The target, if present, has a value such that the target-distractor difference is  $\Delta$  on each trial. The observer reports whether the target is present.
- g) The target is always vertical and always present. Distractors are drawn from a uniform distribution. The observer reports which location contained the target.

**Problem 6.10.** In this chapter, the state-of-the-world variable was always discrete. Discuss whether most ecologically relevant tasks have a discrete or a continuous world state variable, and why.

**Problem 6.11 (\* Advanced): Monte Carlo simulation**

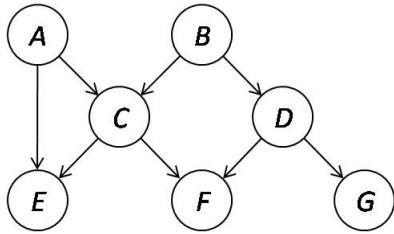
Suppose an observer infers a binary variable  $C$  from an observation  $x$ . We have derived a decision rule  $d(x) > k$  (Bayesian or not does not matter for this exercise). We are now interested in step 3 of Bayesian modeling, determining the probability that the observer will respond  $C=1$ . Explain the following two equalities, of which the second is Monte Carlo approximation.

$$\begin{aligned}
 \Pr_{x|C}(d(x) > k) &= \int 1_{d(x) > k} p(x|C) dx \\
 &\approx \frac{1}{\text{number of samples}} \sum_{\text{samples from } p(x|C)} 1_{d(x) > k}.
 \end{aligned} \tag{6.8}$$

Here, 1 is the *indicator function*: it is 1 when the condition in the subscript is satisfied, and 0 otherwise.

**Problem 6.12. Inference under a complex generative model**

- a) Using the “following the arrows” recipe, derive the posterior for cue combination. The joint distribution is  $p(s, x_A, x_V)$ . Following the arrows, we find  $p(s, x_A, x_V) = p(s) p(x_A|s) p(x_V|s)$ . The posterior we are interested in is  $p(s|x_A, x_V)$ , which is equal to the joint distribution divided by the normalization  $p(x_A, x_V)$ .
- b) For the following generative model, use the principles of marginalization to rewrite  $p(A|E, F)$  in terms of  $p(A)$ ,  $p(B)$ , and the conditional probabilities given by the arrows (such as  $p(C|A, B)$ ):



(From Jensen and Nielsen, *Bayesian Networks and Decision graphs*).

**Problem 6.13.** A musical tone has a pitch (roughly the logarithm of its frequency). An experimenter conducts an auditory oddity detection task, as follows. She draws two values of pitch, denoted  $s$ , independently from a Gaussian distribution  $p(s)$  with standard deviation  $\sigma_s$  (the mean is irrelevant). She then presents to the subject a sequence of three tones, two with the first drawn value of pitch, and one with the second value. The three tones are presented in random order and the subject reports which of the three is the odd one out. Assume that the measured pitch of each tone is independently corrupted by zero-mean Gaussian noise with standard deviation  $\sigma$ .

- Draw a graphical representation of the generative model, and write down expressions for the probability distributions over the variables.
- Derive how the Bayesian MAP observer should estimate the temporal location of the oddball (1, 2, or 3) from the measured pitches.
- Explain why the rule you obtained makes intuitive sense.
- In Matlab, assume  $\sigma_s=2$ . For each value of  $\sigma$  from 0.05 to 5 in steps of 0.05, simulate 100,000 sets of measurements and use those to estimate the probability correct of the optimal observer. Plot this probability as a function of  $\sigma$ .
- What is the value of the asymptote as  $\sigma \rightarrow \infty$ ?
- Repeat step (d) for two ad hoc models. In the first ad-hoc model, the observer determines which measurement lies farthest away from the average of the three measurements, and reports the location of that measurement as the oddball location. In the second ad-hoc model, the observer compares the distances between pairs of measurements, finds which of these three distances is smallest, and chooses as the oddball location the location of the measurement not included in

that pair. Show that these ad-hoc models lead to lower performance by plotting probability correct of the three models as a function of  $\sigma$  in the same plot.

## LAB PROBLEMS

### Problem 6.15: Simulating visual search

Every day, predators look for prey, prey animals look for predators, and countless people look for their mobile phones. All of these are examples of *visual search*: trying to determine whether a target object is present or not. Visual search usually gets harder the more objects are in a scene. We examine that effect in this problem.

In a laboratory visual search experiment, an observer has to detect whether a target orientation is present or absent among  $N$  oriented line segments. The target has orientation  $0^\circ$  and each other item (distractor) has orientation  $10^\circ$ . On each trial, the experimenter chooses whether the target will be present or absent with equal probability. When the target is present, each stimulus is equally likely to be the target. Assume that the measurement at each location is corrupted by Gaussian noise with standard deviation  $5^\circ$ .

- Suppose  $N=2$ . Simulate the measurement on 5000 target-present trials and 5000 target-absent trials. For convenience and without loss of generality, you can assume that when the target is present, it is at the first location.
- On each simulated trial, compute the log posterior ratio using the equation discussed in the lecture.
- Plot the resulting histograms of the log posterior ratio, one for target present and one for target absent, in the same plot.
- Assume the observer performs MAP estimation. Calculate percentage correct.
- Repeat your simulation for set sizes from 1 to 20. Plot percentage correct as a function of set size. At what set size does this observer's percentage correct fall below 60%?

### Problem 6.16: Visual search with heterogeneous distractors

This problem is a follow-up to Problem 6.6 above.

- In Matlab, simulate  $10^5$  trials with  $N=2$ ,  $s_T=0$ , and  $\kappa=10$ . On each trial, draw observations from the generative model (for drawing from a Von Mises distribution, you might want to use `randraw.m`, downloadable from Matlab Central File Exchange) and compute the log posterior ratio of target presence (LPR). Plot the empirical distributions of this LPR when the target is present and when it is absent (using a smooth curve might be better for presentation than using histograms). Vectorize your code as much as you can.
- Plot the receiver operating characteristic (ROC).
- Repeat parts (a) and (b) for two different set sizes,  $N=4$  and  $N=8$ . Plot LPR distributions and ROCs in a way that you can easily compare across  $N$ . Interpret the effects of  $N$ .

**Problem 7.1: Causal inference.** In Section 6.7.2, we derived the log likelihood ratio of the number of causes under certain assumptions. Using the parameters given there,

- Reproduce Fig. 7.4a in Matlab.

- b) Reproduce Fig. 7.4b in Matlab.

**Problem 7.2. Sameness judgment.** In Section 6.8, we made a start with deriving the Bayesian decision rule for a sameness judgment task.

- a) Finish this derivation. The final result should have a quadratic function of  $x_1, \dots, x_N$  on the left-hand side and a constant expression on the right-hand side.
- b) In the special case that  $\sigma_s = \sigma$ , simplify and interpret your decision rule
- c) (\*) Derive the probability of reporting “same” on a “different” trial, when the stimuli are  $\mathbf{s} = (s_1, \dots, s_N)$ .

**Problem 7.3.** Consider sameness judgment in which “same” stimuli are equal to  $\mu$ , and “different” stimuli are drawn independently from a distribution  $p(s|\mu)$ . In both conditions,  $\mu$  is drawn from a distribution  $p(\mu)$ .

- a) Draw the generative model.
- b) Work out the log posterior ratio as far as you can. You do not need to substitute specific distributions for  $p(x|s)$ ,  $p(s|\mu)$ , and  $p(\mu)$ .

**Problem 7.4.** Treat Fig. 7.9c-d (law of proximity) in a quantitative, Bayesian way, analogous to how we discussed Fig. 10 (law of continuity) in Section 7.3.

**Problem 7.5.** Why did we resort to natural statistics for contour integration (Geisler and Perry) while we didn’t need them to explain the percept in Fig. 7.9 (intersecting lines)?

## Table of Contents

8 Recursive Bayesian inference.....	1
8.1 Learning as recursive inference.....	2
8.2 Recursive inference in binary perceptual tasks .....	5
8.3 Systems with dynamics .....	9
8.4 Hidden Markov Models .....	11
8.4.1 Generative model.....	11
8.4.2 Inference .....	14
8.4.3 Using the HMM.....	16
8.5 Kalman filters .....	17
8.5.1 Generative model.....	18
8.5.2 Inference .....	18
8.5.3 Extension to Kalman controllers.....	22
8.5.4 Applications of the Kalman Filter.....	22
8.6 Smoothing.....	24
8.7 Further reading .....	29
8.8 Problems.....	30

## 8 Recursive Bayesian inference and learning

*How can we integrate information over time?*

The previous chapters have focused on scenarios in which information does not need to be integrated over time. Such scenarios are found more frequently in behavioral laboratories than in the real world. In everyday life, we typically integrate information over time, taking in one piece of information after another. Learning typically requires such integration over time. For example, we learn how to move an object by repeatedly moving it. Furthermore, we often observe a changing scene unfolding before us. For instance, we may be following an ongoing conversation, observing the movements of others, or moving ourselves (Fig. 8.1). In such situations, in order to track the changing scene accurately, we must integrate information over time. Temporal integration is crucial in life and in this chapter we will explore its statistical properties.

Plan of the chapter: In this chapter, we apply Bayesian inference recursively to observations taken over time. We will learn how to form estimates of the current world state by combining the likelihood function based on the current observation with our posterior distribution from the previous time step. We analyze progressively more

complex temporal estimation problems. We begin with examples that show how to estimate a fixed world state using successive observations. Next, we explore the ongoing estimation of a changing world state, for which we use sensory information together with knowledge of the world state dynamics. We will describe two classes of filtering algorithms, Hidden Markov Models, which deal with discrete state spaces, and Kalman filters, which deal with continuous state spaces. We will discuss both filtering where we estimate variables based on the past and smoothing where we estimate variables based both on past and present.



**Figure 8.1.** In order to touch down safely on the moon in 1969, the Apollo 11 lunar lander was equipped with an onboard computer that ran a Kalman filter to integrate ongoing sensor readings of speed and altitude. Lunar landing was among the first applications of Kalman filters. In fact, the game Lunar Lander, running on a DEC PDP-8 computer, made history as one of the early computer games. Picture courtesy of NASA.

### 8.1 Learning as recursive inference

Learning is a process by which we update what we believe about the world around us. Whenever we make an observation we will update these beliefs, which will usually get more precise when we make observations (unless the world is changing). When learning is very fast and deals with ongoing changes it is also often called *adaptation*. Much work in cognitive science, perception, and motor control deals with the process of learning.

At some level all of Bayesian statistics is about learning. We calculate beliefs and update them with new information. This change in belief itself can be seen as a form learning. But then the variables that we have beliefs over are of many different kinds. I may have a

belief about a given suitcase being mine, about the general distribution of suitcases or about the structure of physical objects in the world. Many scientists would thus distinguish between immediate estimates, e.g. updating beliefs about a given suitcase being mine, and higher level learning, e.g. updating beliefs about the general distribution of suitcases. However, all of these forms of learning fall under the same Bayesian framework.

Of course we need to be careful about the relation of Bayesian statistics and learning or adaptation. Bayesian statistics is ultimately about the correct use of information while terms like adaptation and learning relate to specific behaviors of animals. Still, Bayesian statistics provides natural normative models for many of these phenomena.

### 8.1.1 An example of learning

Many real-life learning scenarios occur over time with multiple iterations. As a simple example, let's consider a toddler who is learning how to control her limbs. To first approximation, the command signal (e.g., firing rate of motor cortex neurons) sent to the spinal cord adjusts muscle force linearly. But what is the slope of this relationship between motor command and force output for a particular muscle? Without this knowledge, the toddler will be clumsy; as she acquires this knowledge, her motor control will improve. Let's consider the toddler's first attempts to learn this relationship. For simplicity, we assume that the toddler already understands that the relationship between motor command and force output is linear (though, clearly, this too must be learned), but she doesn't know the slope. Her goal is to estimate the slope,  $m$ , relating command signal ( $c$ ) to force output ( $f$ ):

$$f(c) = mc$$

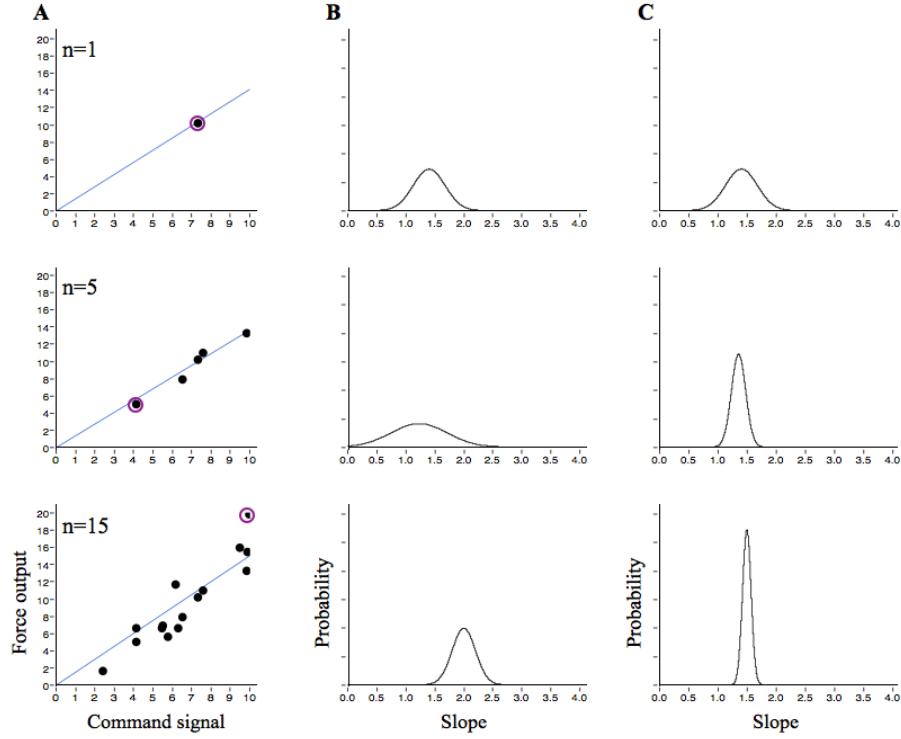
Figure xx shows the results of 10 iterations in which the toddler uses different command signal magnitudes to push against the floor with her arms, while she judges the force she produces. She is able to judge the force based on feedback from her proprioceptors. These sensory signals are noisy, so we model her force measurement,  $x$ , on each trial as a sample drawn from a Gaussian distribution around the actual force produced.

After each push, with her knowledge of the noise distribution, the toddler can construct a likelihood function reflecting the probability of the measurement given the slope. She can then calculate her posterior over slope. To do so, she uses Bayes' rule to combine the likelihood function based on the current trial with the prior over slope:

$$p(m|x_i) \propto p(x_i|m)p(m|x_{i-1})$$

Here,  $p(m|x_i)$  represents the posterior probability over slope, based on all trials up to and including the present trial ( $i$ ), and  $p(m|x_{i-1})$  represents the prior probability of the slope, which is the posterior from the previous iteration, based on all trials up to and

including trial i-1. Notice the recursive nature of this procedure, in which the posterior following each trial is the prior used for the next trial (figure xx).



**Figure xx.** An observer learns the slope relating motor command signal magnitude (x-axis) to muscle force production (y-axis). **(A)** scatterplots showing the data accumulating from trial 1 to 5 to 15. The line in each plot shows the MAP slope estimate based on all trials up to and including the one shown (data point circled in purple). **(B)** Single-measurement likelihood functions from the corresponding trials. **(C)** Posterior distribution over slope. The actual slope value used to generate the data was  $m = 1.5$ , with  $\sigma = 2$ .

## 8.2 Bayesian models of behavior

Within motor control, Bayesian approaches are often used to model learning. For example, during movement we need to know the properties of the muscles in our body and if our muscles get weaker we need to send stronger signals to produce the same force. The force of muscles may vary according to a (potentially multi timescale) random walk. And observations are in good approximation linear. As such, the problem of estimating the properties of the muscles can be formalized by iterative Bayesian updating.

Within motor control, models based on hierarchical Bayesian estimation have been used to explain a broad range of movement phenomena. The way the external world and its

forces change over time has been modeled like this. The way muscles change has been modeled like this. Arguably, models based on iterative Bayesian estimation are the standard model used in motor control.

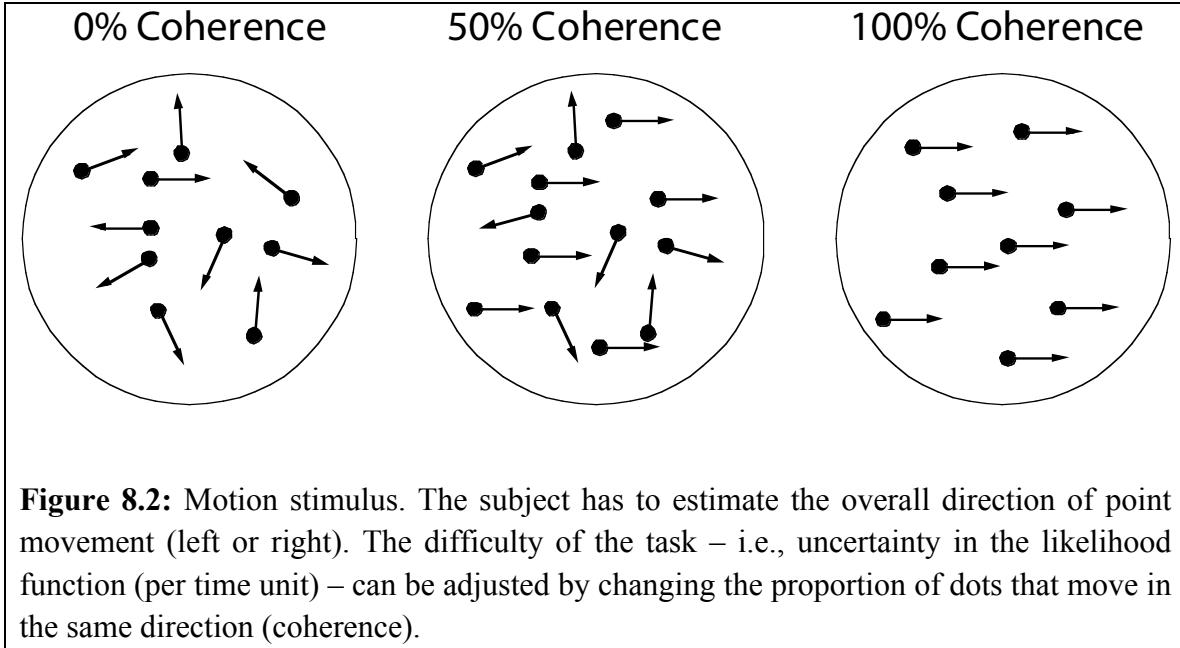
However, Bayesian models of learning are important well beyond motor control. Bayesian learning has been described in the domain of reinforcement learning, perceptual learning, classical and operant conditioning, the estimation of the structure of social groups and many other domains. In all these domains there is evidence that Bayesian models are better at fitting human behavior than simpler non-Bayesian models. Bayesian models allow a simple and straightforward way to model learning.

#### **Box: Learning without prior knowledge is impossible**

If we were to be born into the world without prior knowledge we could never learn anything, and, hence, that science is theoretically impossible. This realization, probably first described by Hume (1739-1740) comes from the fact that past experiences only becomes meaningful through additional assumptions – for example, that similar situations will lead to similar outcomes. In fact, Wolpert (1996) showed in a theorem called *No free lunch for machine learning* that no learning system can be better than any other learning system (across all possible learning problems). However, if a system starts out with prior knowledge about the kind of learning problems to be encountered it can learn and the quality of prior knowledge determines the quality of learning.

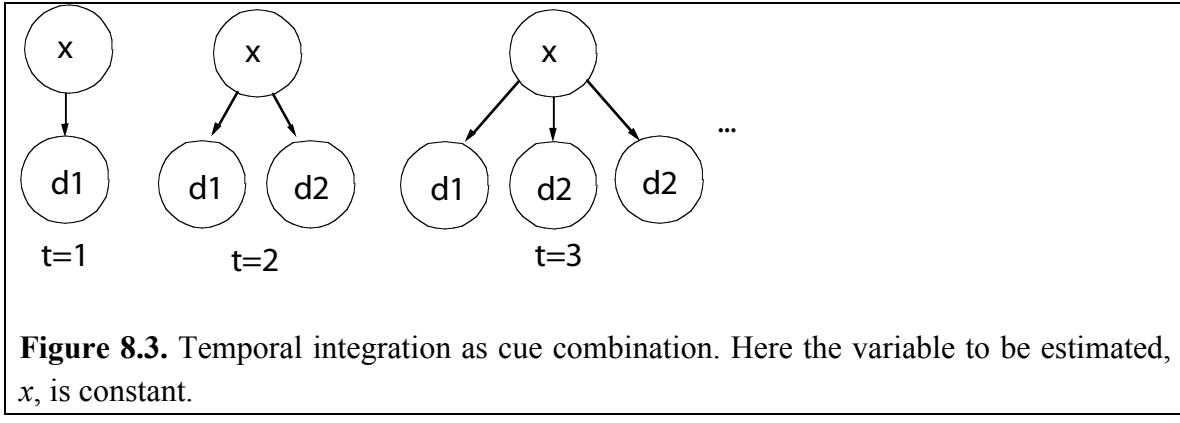
### **8.3 Recursive inference in binary perceptual tasks**

In chapter 5 we discussed binary perceptual tasks in which subjects derived a posterior pdf based on a single observation. In reality, however, we often make multiple observations over time. One example in the domain of neuroscience research is the estimation of the direction of movement (left or right) of a stimulus. In a typical experiment, throughout some time interval subjects observe a stimulus consisting of moving dots (Gold and Shadlen 2001; see Fig. 8.2). The moving-dot stimulus can be varied so that during the observation interval it either produces a rather flat likelihood function (if the dots are moving in nearly random directions) or a sharp likelihood function (if almost all dots are moving in the same direction). Importantly, subjects' perception of motion direction is based on their integration of the sensory information over time.



When the stimulus direction,  $x$ , does not change during the observation interval, it is possible to calculate the posterior probability over  $x$  by viewing the problem as a cue combination problem. Every time-step provides one cue, and we thus have the generative model drawn in Fig 8.3. Under the assumption that the observations at the different time points are conditionally independent given the true stimulus direction, Bayes' rule yields:

$$p(x | d_1 \dots d_N) \propto p(x) p(d_1 \dots d_N | x) = p(x) \prod_{i=1}^N p(d_i | x) \quad (0.0)$$



However, this way of modeling the system does not readily generalize to the estimation of variables that do change over time. To obtain that more general solution, we phrase the estimation problem as a recursive procedure: We start with a belief (prior pdf); when a new data point comes, we use it (likelihood function) to update our belief (posterior pdf); we then use our updated belief, the posterior pdf at the current time, as the prior pdf for

the next time step. To derive this recursive procedure mathematically, we rewrite equation **Error! Reference source not found.** as follows:

$$\begin{aligned}
 p(x | d_1 \cdots d_N) &\propto p(x) \prod_{i=1}^N p(d_i | x) \\
 &= p(x) p(d_N | x) \prod_{i=1}^{N-1} p(d_i | x) \propto p(x | d_1 \cdots d_{N-1}) p(d_N | x)
 \end{aligned} \tag{0.1}$$

Our posterior over  $x$  at time  $N-1$ ,  $p(x | d_1 \cdots d_{N-1})$  is our prior at time  $N$ . This is multiplied by the likelihood at time  $N$ ,  $p(d_N | x)$ , to yield our posterior at time  $N$ ,  $p(x | d_1 \cdots d_N)$ , which will serve as our prior at time  $N+1$ . As we will see below, this iterative conceptualization forms the basis for nearly all temporal integration algorithms.

The logarithm of the posterior ratio Eqn. (0.1) is:

$$d = \frac{\log p(x = \text{right} | d_1 \cdots d_N)}{\log p(x = \text{left} | d_1 \cdots d_N)} = \frac{\log p(x = \text{right})}{\log p(x = \text{left})} + \sum_{i=1}^N \log\left(\frac{p(d_i | x = \text{right})}{p(d_i | x = \text{left})}\right)$$

where  $C$  is an arbitrary constant. Thus, the log posterior evolves as the log likelihood,  $\log(p(d_i | x))$ , is essentially integrated (i.e., summated) over time. Under a broad range of assumptions, the relevant variable to be summed over:  $\log\left(\frac{p(d_i | x = \text{right})}{p(d_i | x = \text{left})}\right)$  has an approximately Gaussian distribution.

Exercise 8.1: Show that for Gaussian likelihoods with fixed  $\mu$  and  $\sigma$  the above approximation is good.

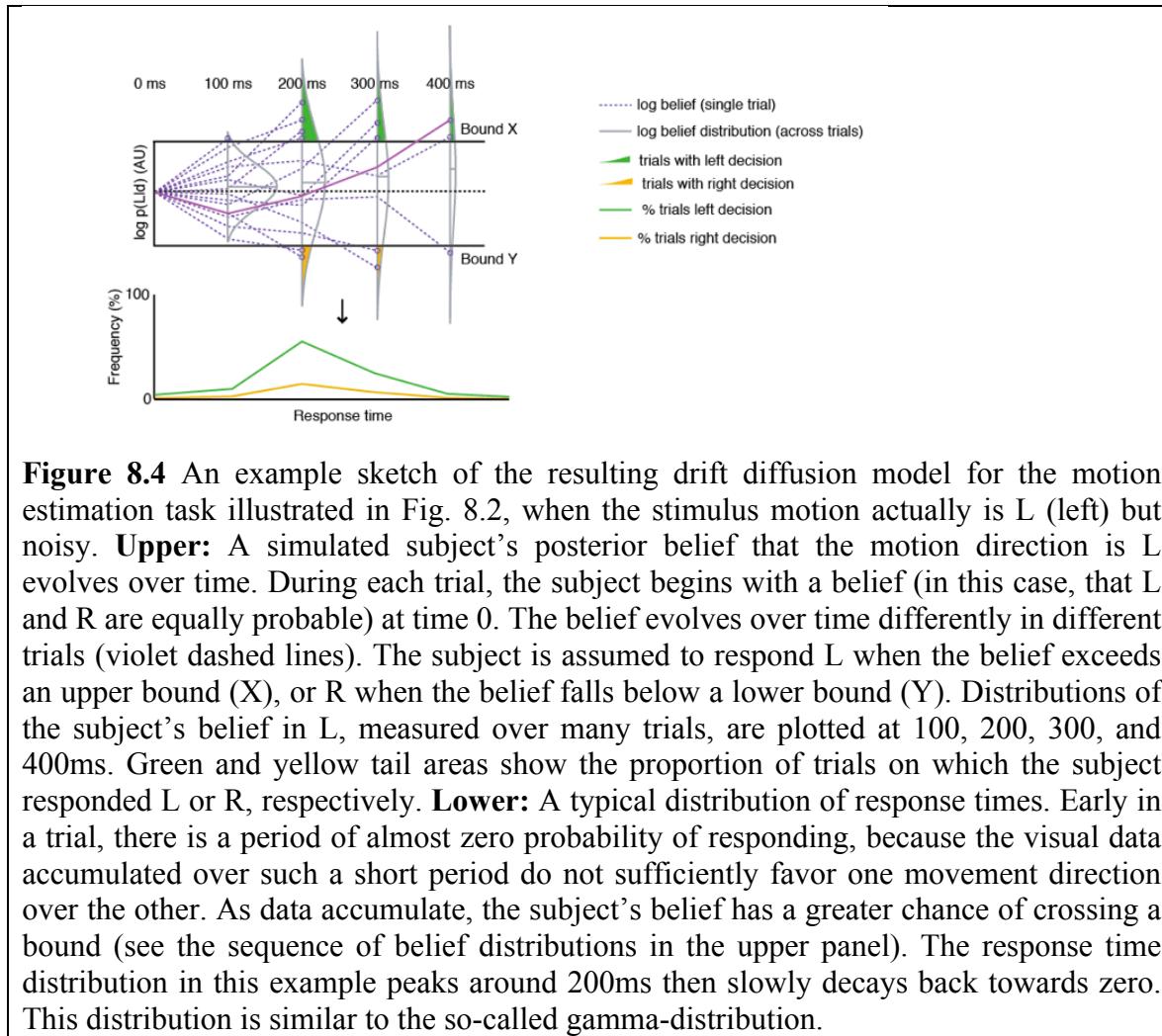
Under those circumstances, the posterior probability ratio evolves as a so-called drift-diffusion process (See Fig 8.4). If people make a decision when the probability of one value of  $x$  crosses a particular threshold value, this represents a diffusion-to-bound process. As soon as the decision variable  $d$  exceeds a certain upper bound (e.g.,  $\log 100$ ) or falls below a lower bound (e.g.,  $\log 0.01$ ), a decision is taken. Importantly this process predicts a gamma-like distributions of reaction times, which has been observed in experiments.

We want to go through an example of this process. Lets say we start with the assumption that leftward and rightward motion is equally probable:  $p(X = \text{left}) = p(X = \text{right}) = .5$ . Without any observations, each of the two potential stimuli are thus equally probable  $\log p(X = \text{left} | \{\}) = -\log\sqrt{2}$ .

We assume that  $\log\left(\frac{p(d_i | x = \text{right})}{p(d_i | x = \text{left})}\right)$  has a Gaussian distribution with mean .5 and std of

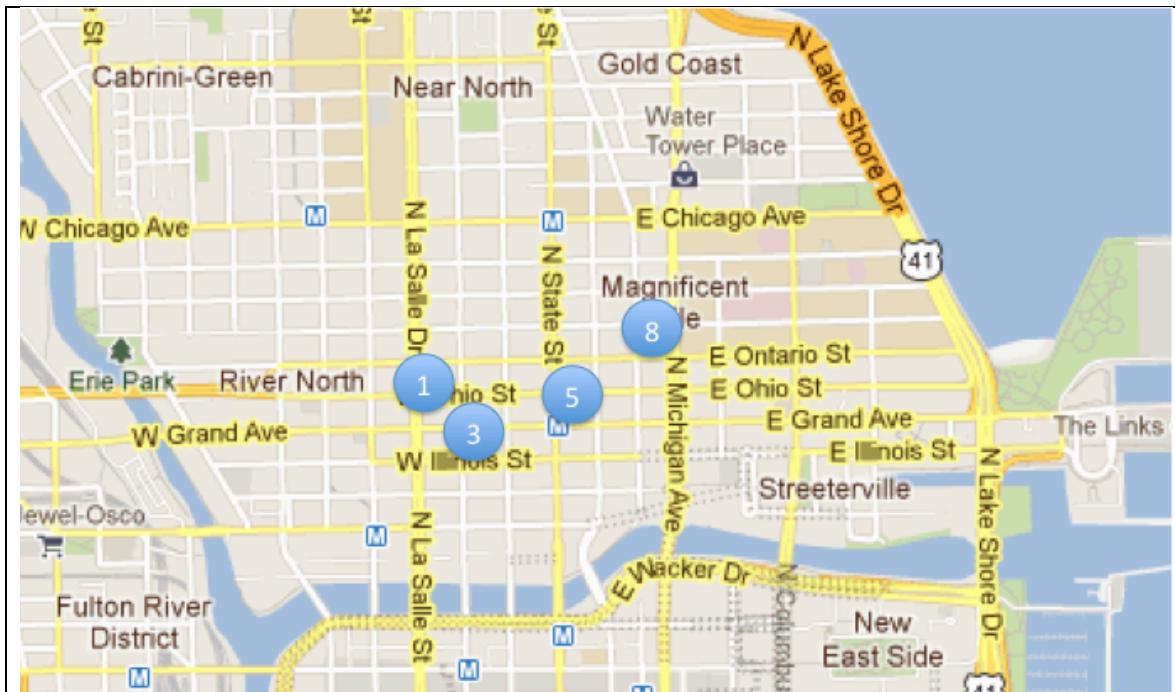
2. Trials end when the log posterior probability is either above  $\log(100)=4.6$  or below  $\log(.01)=-4.6$

In a first trial we might first obtain a ratio of 1.3 (right is more probable). Next we might observe a -1.7 (left is more probable) which would lead us with a cumulative -.4 (left marginally more probable). This could be followed by 2.9 (right is more probable) and another 2.9 (right is more probable). After these two observations we would have a cumulative 5.4 allowing us to conclude with  $p<.01$  that the stimulus is moving to the right. However, on another trial we might observe (-2.8 -1.5 -2) getting us an integral of -6.3 getting us to come up with the wrong answer of left movement. On other trials, we might obtain the correct answer much faster, e.g. we might obtain a value of +5 at the first time point. The cumulative sum of the likelihood ratio along with the prior defines the decision variable.



## 8.4 Recursive inference in systems with dynamics

Many real-life situations require us to integrate information while the state of the world changes. For example, suppose we are walking through the downtown of a large city. We occasionally observe a reading on a handheld GPS device, revealing our location on Google Maps with some positional uncertainty (Fig. 8.5). However, due to interference from skyscrapers, much of the time we have no GPS reading. As we are on the move, and we have only intermittent sensor readings, how can we best estimate our location at any given time? The answer is that we can combine the information from our recent sensor readings with knowledge of our dynamics. For example, we may have a general understanding of the direction in which we are walking, which constitutes a predictable part of our dynamics. Dynamics and observations jointly define the estimation problem.



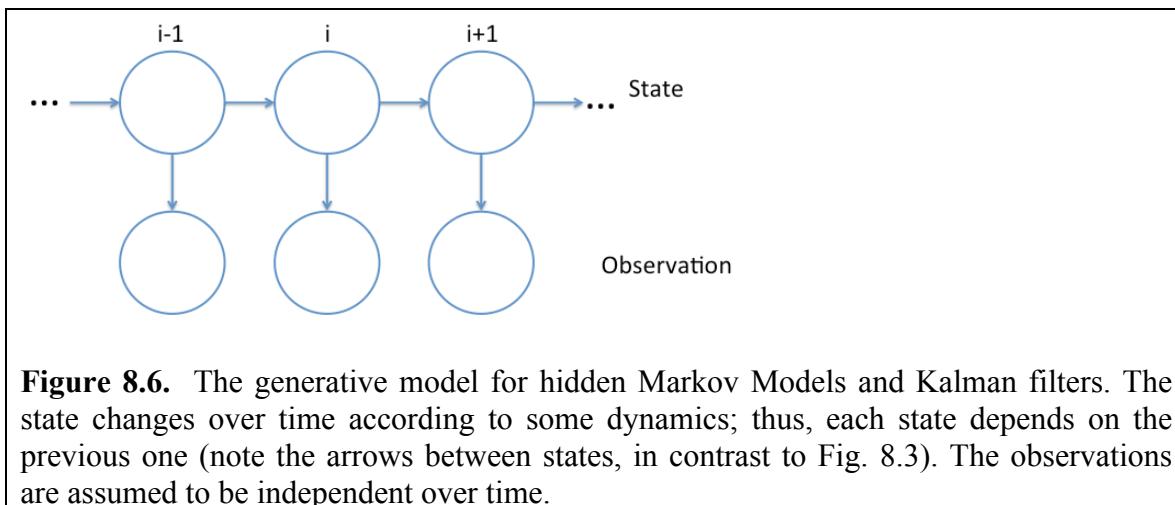
**Figure 8.5.** A person walking through a city receives intermittent readings from her GPS device. Using Bayesian filtering, she estimates her location from knowledge of her dynamics and information obtained from her sensors.

Obviously, we will use our observations to update our estimate of the world state (our location). Each GPS reading is a maximum likelihood estimate. Most GPS receivers also report their uncertainty; thus, we know the width of the likelihood function. Through observations, our estimates generally become less uncertain, and the higher the quality of the observation (i.e., the sharper the likelihood function), the more our uncertainty will decrease.

However, as we walk through the city, our position is continually changing, so we also have to contend with dynamics. In fact, our walking has two components: a predictable component - we may roughly know the direction we are moving -, and an unpredictable component - we may have some noise in the way we walk that produces random displacements of our position. Through dynamics the world state generally becomes less predictable with the passage of time.

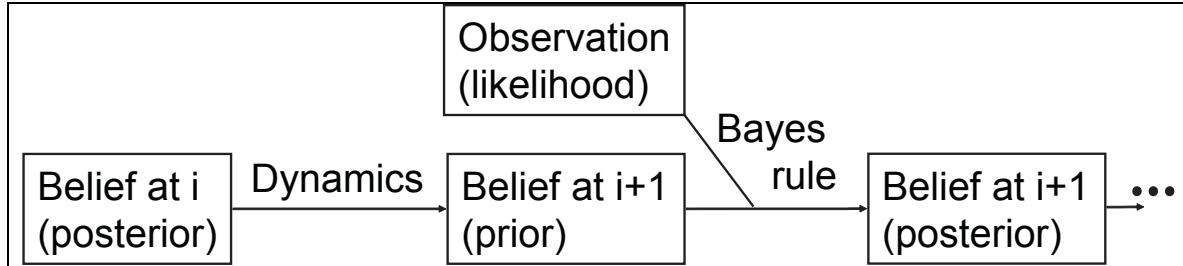
There are a great many real-life situations involving dynamics, similar to our city walking example, that require integration over time. As we move our arm to reach for a glass of water, we can estimate where our hand is in space based on the known dynamics of our hand along with our proprioceptive and visual observations of hand position. In higher-level cognition, we may estimate what other people believe about us given our understanding of the dynamics of such beliefs and the communication we have with those people. In the laboratory, neuroscientists also encounter estimation problems involving dynamics. For example, we may want to estimate the state of neurons over time, based on neuronal dynamics and electrophysiological observations.

In the following sections, we describe two Bayesian approaches used to address such problems: hidden Markov models and Kalman filters (Fig. 8.6).



**Figure 8.6.** The generative model for hidden Markov Models and Kalman filters. The state changes over time according to some dynamics; thus, each state depends on the previous one (note the arrows between states, in contrast to Fig. 8.3). The observations are assumed to be independent over time.

The Bayesian strategy – known generally as recursive Bayesian filtering – is to use our understanding of the dynamics to convert our posterior belief at each time point into a prior belief for the next time point. The observation (likelihood function) is then combined with this prior using Bayes' rule, to generate a new posterior, and the process repeats (Figure 8.7).



**Figure 8.7** Bayesian filtering algorithms alternate between two steps: the application of dynamics and the application of Bayes rule as new information comes in.

## 8.5 Hidden Markov Models

If the world could change in arbitrary ways from one moment to the next, then temporal integration would be ineffective. Fortunately, however, the world in which we live exhibits a high degree of statistical regularity across time. For example, many time-varying processes in everyday life conform approximately to the so-called Markov assumption. A Markov process is a time-varying process in which the future value of the world state variable depends only on its current value; thus, the future is conditionally independent of the past, given the present:  $p(\text{future} \mid \text{past}, \text{present}) = p(\text{future} \mid \text{present})$ . Many physical processes are Markovian. For example, if we know the current position and velocity of a ball thrown in the air, we can predict its future trajectory. Any information we have about its trajectory leading up to the current moment does not contribute to our ability to predict its future flight. Here, we describe how to estimate the current state of a changing world that is characterized by Markovian dynamics.

### 8.5.1 Generative model

#### *Observation*

Let us return to the problem of estimating the direction of a moving stimulus, for instance the direction of movement of an animal observed from a distance, or the direction of movement of the random dot display (Fig. 8.2). For the sake of simplicity, we will suppose that the movement can occur in only two directions, left (L) or right (R), so that the world state variable is binary. We will assume that leftward and rightward movements are equally probable a priori. We will further suppose that the sensory input is noisy, so that whichever direction the stimulus is actually moving, we have a 20% probability of wrongly observing it to be moving the opposite way. Thus, using L and R to represent the latent world state variable (true movement direction), and arrows to indicate observed movement to the left ( $d = \leftarrow$ ) or right ( $d = \rightarrow$ ), we have:

$$p(R) = p(L) = 0.5$$

$$p(\leftarrow \mid L) = 80\%$$

$$p(\leftarrow \mid R) = 20\%$$

### Dynamics

Let us say that the direction of movement changes occasionally. At any given point of time, motion is in one direction, but on average the direction switches every 5 time steps. These dynamics can be summarized by a transition distribution that characterizes the state at time  $i+1$  conditioned on the state at time  $i$ .

$$p(R_{i+1} | R_i) = 0.8$$

$$p(L_{i+1} | R_i) = 0.2$$

$$p(R_{i+1} | L_i) = 0.2$$

$$p(L_{i+1} | L_i) = 0.8$$

These equations define the dynamics. They define how the state of the system changes, independent of any observation.

Now we wish to use our knowledge of the dynamics to generate a prior pdf for time step  $i+1$ , given our belief about the world state (posterior pdf) at time step  $i$ . If at the  $i^{\text{th}}$  time point the probability of the two states are  $p(L_i)$  and  $p(R_i) = 1 - p(L_i)$  then we can use our knowledge of the dynamics to find the corresponding probabilities at the next time point, by marginalizing:

$$p(R_{i+1}) = p(R_{i+1} | R_i)p(R_i) + p(R_{i+1} | L_i)p(L_i)$$

$$p(L_{i+1}) = p(L_{i+1} | L_i)p(L_i) + p(L_{i+1} | R_i)p(R_i)$$

Thus, we have calculated the probability of each possible state at the next time step, before we make an observation at that time step. This is our prior pdf at time step  $i+1$ .

#### Box: Marginalization as part of dynamics.

*Marginalization* is the statistical procedure that considers all possible values of a variable that we are not interested in when calculating the probability distribution of a variable that we are interested in (see Chapter 6). In doing Bayesian inference with dynamical systems, we encounter a typical case of such an irrelevant variable that gets marginalized out.

If we knew the world state at the previous time step, the dynamics would allow us to easily specify our beliefs about the world state at the current time step. For instance, if we knew that the previous state was L, then we would know that the current state could be R with 20% probability or L with 80% probability. However, we do not know, but rather have a probability distribution over the previous world state. To take this uncertainty into account, we must marginalize over all possible values of the previous state in order to compute our beliefs (prior to the observation) at the current time:

$$p(state_{now}) = \sum_{\text{previous states}} p(state_{now} | state_{before}) p(state_{before})$$

This equation sums the probabilities of all ways that a particular current state could arise. For instance, the current state could be “L” if the past state was L AND the state remained L, OR if the past state was R AND the state changed to L. Following the rules of probability, we represent AND by multiplication, and OR by addition.

If we are dealing with continuous states, the marginalization sum becomes an integral:

$$p(state_{now}) = \int p(state_{now} | state_{before}) p(state_{before}) dstate_{before}$$

When states are high-dimensional, these sums and integrals require exponential effort, numerically. This is arguably the most complicating element in Bayesian statistics, and many Bayesian methods boil down to finding good approximate solutions to the above equations. Fortunately, for finite dimensional discrete distributions, Poisson distributions, Gaussian distributions, and others, the marginalization equations can be solved analytically.

Conveniently, in the case of the HMM it is possible to rewrite the dynamics using linear algebra. We can define the belief state (before the observation) as a vector

$$\mathbf{P}_{i, \text{before observation}} = \begin{bmatrix} p(L_i | d_1 \dots d_{i-1}) \\ p(R_i | d_1 \dots d_{i-1}) \end{bmatrix}$$

We can define the dynamics matrix  $\mathbf{M}$  as

$$\mathbf{M} = \begin{bmatrix} p(L_{i+1} | L_i) & p(L_{i+1} | R_i) \\ p(R_{i+1} | L_i) & p(R_{i+1} | R_i) \end{bmatrix}$$

For our example,  $\mathbf{M}$  is a constant matrix:

$$\mathbf{M} = \begin{pmatrix} .8 & .2 \\ .2 & .8 \end{pmatrix}$$

We can then rewrite the dynamics part:

$$\mathbf{P}_{i+1, \text{before observation}} = \mathbf{M} \mathbf{P}_i \tag{0.1}$$

Thus, the HMM dynamics can be viewed as a matrix multiplication.

### 8.5.2 Inference

The fundamental idea behind the HMM is that at any point of time we have a belief: we can assign a probability to every possible state the system may be in. The belief at the next time step, before we make any additional observations, will then be determined by the dynamics of the system. As we assume the Markov property, the belief at the beginning of the next time step, before the observation, will only depend on the belief at the last time step and not on the beliefs at any previous time step.

After obtaining a posterior pdf at one time step, we update our belief, using the dynamics equation, to calculate a prior pdf for the next time step. At each time step, we thus have a prior derived from the dynamics and the posterior at the previous time step. Then, given the current data (e.g., a rightward observation), we update using Bayes rule:

$$P(R_i \rightarrow_i, d_1 \cdots d_{i-1}) = p(\rightarrow_i | R_i, d_1 \cdots d_{i-1}) p(R_i | d_1 \cdots d_{i-1}) / p(\rightarrow_i | d_1 \cdots d_{i-1})$$

Because the observation depends only on the current world state, we can write  $p(\rightarrow_i | R_i, d_1 \cdots d_{i-1}) = p(\rightarrow_i | R_i)$  so that:

$$P(R_i \rightarrow_i, d_1 \cdots d_{i-1}) = p(\rightarrow_i | R_i) p(R_i | d_1 \cdots d_{i-1}) / p(\rightarrow_i | d_1 \cdots d_{i-1})$$

And through normalization:

$$P(R_i \rightarrow_i, d_1 \cdots d_{i-1}) = \frac{p(\rightarrow_i | R_i) p(R_i | d_1 \cdots d_{i-1})}{p(\rightarrow_i | R_i) p(R_i | d_1 \cdots d_{i-1}) + p(\rightarrow_i | L_i) p(L_i | d_1 \cdots d_{i-1})}$$

with an analogous term for  $p(L_i \rightarrow_i, d_1 \cdots d_{i-1})$ .

Again we can write this in matrix form by defining:

$$\mathbf{V}_i = \begin{bmatrix} p(\rightarrow_i | L_i) \\ p(\rightarrow_i | R_i) \end{bmatrix}$$

We then get (where  $*$  is the pointwise product):

$$\mathbf{P}_i = \frac{\mathbf{P}_{i \text{ before observation}} * \mathbf{V}_i}{\sum_k (\mathbf{P}_{i \text{ before observation}} * \mathbf{V}_i)_k} \quad (0.1)$$

Here, the sum in the denominator is over the two array elements; this is simply the denominator of Bayes' rule.

Alternating between application of the equations for dynamics (8.3) and observation (8.4) allows us to calculate the optimal Bayesian estimates as they evolve over time for the HMM.

### *Concrete example*

Here we solve the problem step by step for our example, assuming that we observe the sequence:  $\leftarrow \leftarrow \leftarrow \rightarrow \rightarrow \rightarrow$

We start with

$$\mathbf{P}_0 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}.$$

We do not know anything about the direction of movement. After applying the dynamics we still have

$$\mathbf{P}_{1, \text{before}} = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}.$$

We observe the first leftward movement  $\leftarrow_1$ , and obtain

$$\mathbf{P}_1 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} * \begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix} / 0.5 = \begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix}.$$

So, our best estimate after the first observation is exactly what is given by the likelihood function. We should expect that because thus far, dynamics has only provided us with a flat prior.

### *Next time step*

We now continue the time-dependent estimation. Using the dynamics we obtain

$$\mathbf{P}_{2, \text{before}} = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix} \begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix} = \begin{bmatrix} 0.68 \\ 0.32 \end{bmatrix}$$

What we see is that the dynamics made our beliefs more uncertain. This should naturally be expected, as it could be that direction of movement has just switched. We observe another leftward movement  $\leftarrow$  and obtain (rounding to two significant digits).

$$\mathbf{P}_2 = \begin{bmatrix} 0.68 \\ 0.32 \end{bmatrix} * \begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix} / 0.61 = \begin{bmatrix} 0.89 \\ 0.11 \end{bmatrix}$$

After observing the second leftward movement we are more certain that the state is leftwards.

*And the next time step*

Again we calculate the dynamics

$$\mathbf{P}_{3, \text{before}} = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix} \begin{bmatrix} 0.89 \\ 0.11 \end{bmatrix} = \begin{bmatrix} 0.74 \\ 0.26 \end{bmatrix}$$

and  $\mathbf{P}_3 = \begin{bmatrix} 0.92 \\ 0.08 \end{bmatrix}$

We can see that there is a certain element of saturation. As we observe more and more of the same data we continue becoming more certain about the (leftwards) state. However, because it is always possible that the stimulus will turn rightwards this belief rapidly saturates.

*Reversal of evidence*

And, indeed, after the next observation of a rightward movement we obtain

$$\mathbf{P}_{4, \text{before}} = \begin{bmatrix} 0.75 \\ 0.25 \end{bmatrix} \text{ and } \mathbf{P}_4 = \begin{bmatrix} 0.43 \\ 0.57 \end{bmatrix}.$$

In other words, because of the ongoing dynamics, the system, even after observing much data in favor of one interpretation will readily switch its beliefs. And now, it will rapidly converge to the opposite interpretation with

$$\mathbf{P}_5 = \begin{bmatrix} 0.83 \\ 0.17 \end{bmatrix} \text{ and } \mathbf{P}_6 = \begin{bmatrix} 0.9 \\ 0.1 \end{bmatrix}.$$

Exercise 1: Show that the above values for  $\mathbf{P}_3$ ,  $\mathbf{P}_4$ ,  $\mathbf{P}_5$ , and  $\mathbf{P}_6$  are correct.

### 8.5.3 Using the HMM

The HMM is a frequently used model because it balances in an optimal way the use of previous with current information. The nature of the dynamics along with the nature of the observations determines how much weight is given to each of the two sources. HMM can also model any kind of natural dynamics, as long as the involved variables are discrete. HMMs are more commonly used to solve inference problems than to model how humans perceive. However, in some cases HMMs are used to model human behavior. The HMM model is simple in its interpretation and mathematical details and yet, it is being used in many domains in neuroscience, cognitive science and computer science. For example, there are some neurons that switch between two states, a highly excitable “up” state where the neurons will fire even in response to weak inputs, and a less excitable “down” state where neurons will only fire rarely. Observed spike data are noisy

as there are occasional spikes in the off state and occasional periods of silence in the on state. HMMs are also used for speech recognition. Phonemes, the basic constituents of spoken words, can only noisily be observed; hence, HMMs are often used to model words. In language processing HMMs are frequently used, even in the domain of word learning.

**Myth:** Hidden Markov Models (HMMs) are the only models that describe hidden states characterized by the Markov property.

**Truth:** Virtually all time-varying models and many others include latent variables that have the Markov property. The Markov property exclusively denotes that the future is conditionally independent of the past given the present. This feature characterizes many models. For instance, Kalman filters are really Hidden Markov Models. Nevertheless, by convention, the term HMM has come to be used only for discrete-state models.

## 8.6 Kalman filters

Many real world tasks that the nervous system has to solve involve the estimation of continuous variables. We need to estimate where our hand is in space, where a sound is coming from, how fast a car is moving, or how big a potential predator is. In these and a multitude of other scenarios, the brain must use estimation algorithms that work in continuous spaces.

How can we solve such estimation problems? A first thought might be to simply use an HMM with a very large number of states. For example, we could model our hand's location by drawing an imaginary grid to divide space into a large number of discrete cubes. As the number of dimensions increases, however, this strategy would require an exponential number of states. Furthermore, for each of these potential states the nervous system would need to update beliefs and learn the transition probabilities. This would be infeasible in many cases.

Fortunately, it turns out that filtering algorithms are still possible with continuous states. As we will see, the Kalman filter follows exactly the same basic idea as the hidden Markov Model, but applied to continuous states. The observer has a distribution of beliefs, now described by a mean and covariance over a continuous space. The observer uses knowledge of the dynamics to transform the posterior belief distribution at each time point into a prior belief distribution for the next time point. Upon receiving the current observation, the observer multiplies the likelihood function with the prior belief

distribution, using Bayes' rule to create a new posterior belief distribution, and the process repeats.

### 8.6.1 Generative model

The Kalman filter is based on Gaussian probability distributions, along with linear noisy dynamics. Let's consider the example of estimating the position of the hand in space. At any point of time, we can summarize our belief distribution over the hand's position as a three-dimensional Gaussian distribution with mean  $\mu_t$  and covariance matrix  $C_t$ :

$$P(x_t | \mu_t, C_t) = \frac{1}{(2\pi)^{N/2} |C_t|^{1/2}} \exp(-(\mathbf{x}_t - \mu_t)^T C_t^{-1} (\mathbf{x}_t - \mu_t))$$

Exercise 3: Show that the above equation is properly normalized.

For the purpose of illustration, we will consider a one-dimensional estimation problem here, so that our Gaussian belief distribution reduces to the simpler form:

$$P(x_t | \mu_t, C_t) = \frac{1}{\sqrt{2\pi C_t}} \exp(-(x_t - \mu_t)^2 / (2C_t^2)) = N(\mu_t, C_t)$$

We consider multi-dimensional estimation in the Problems.

The generative model assumes linear dynamics with noise:

$$x_{t+1} = Ax_t + \eta \tag{0.1}$$

where  $\eta$  is drawn from a Gaussian distribution with zero mean and  $\sigma_\eta$  standard deviation. This generative model has the property that the distribution of  $x_{t+1}$  given  $x_t$  again has a Gaussian form:

$$P(x_{t+1} | x_t) = \frac{1}{\sqrt{2\pi\sigma_\eta}} \exp(-(x_{t+1} - Ax_t)^2 / (2\sigma_\eta^2)) = N(Ax_t | \sigma_\eta)$$

### 8.6.2 Inference

As an example, suppose we want to estimate the (one-dimensional) position  $x$  of our hand, which we are trying to keep still in space. In order to introduce noisier measurements, we will perform the experiment with our eyes closed, so that the sensory feedback is only proprioceptive and not visual. Let's say that we want to make a position estimate every 100 ms. Normally, one would want to use a state space that contains both position and velocity of the hand, as this is generally necessary for the Markov property to hold, but here, for simplicity's sake we assume position to be Markov. Incidentally, this would be a good approximation if we were to move our hand in a highly viscous medium (e.g. honey).

Now, let's say we start with the prior belief  $P(x|\mu_0, C_0) = N(0, \sqrt{2}) \text{ cm}$ ; that is, we believe our hand is around 0 cm (along a reference axis that extends horizontally from left to right in front of us, with 0 cm indicating straight-ahead), but we have some uncertainty as expressed by a variance of 2 cm<sup>2</sup>. Let's further say that we have dynamics that are given by equation (0.1) with  $A=0.9$  and  $\sigma_\eta = 1 \text{ cm}$ . Because  $A < 1$ ,  $x_{i+1}$  tends to be smaller than  $x_i$ , so that our hand tends to approach zero. Thus, our understanding of the dynamics is that our hand will on average drift towards the middle of the workspace, straight ahead of the body. There is some evidence that this may actually be a reasonable model for the task.

Now, we will consider the situation in which our initial belief about our hand's location is in fact wrong, and our hand actually starts out – unknown to us, of course – at 5 cm. Suppose we then observe our hand at 100ms, 200ms and 300ms, based on noisy proprioceptive input, to be at 4, 6, and 4 cm, respectively.

Here, we go through the inference calculations step by step (all calculations are precise to 2 significant digits). We start with the belief:

$$p(x_0) = N(\mu_0, \sigma_0) = N(0, \sqrt{2}).$$

We then apply the dynamics. To calculate where we will be at the next point in time, we need to marginalize over all possible values of the latent variable:

$$P_{\text{before}}(x_{100ms} | \mu_0, \sigma_0) = \int_x p(x | \mu_0, \sigma_0) p_\eta(x_{100ms} | x) dx \quad (0.1)$$

In other words, to calculate how probable any particular current hand position is, before taking an observation, we have to integrate over all possible positions that the hand could have taken at the previous time, considering for each of those positions the probability that it would transition to the current position.

Equation (0.1) describes the convolution of two Gaussians, which turns out to be another Gaussian whose mean is the difference of the means of the two Gaussians, and whose variance is the sum of the individual variances. Thus, we can rewrite  $P_{\text{before}}$  as:

$$P_{\text{before}}(x_{t+100ms}) = N(0.9\mu_0, \sqrt{\sigma_t^2 + \sigma_\eta^2}) \quad (0.1)$$

Next, we include the new observation using Bayes rule. As the Kalman filter uses Gaussian beliefs and Gaussian likelihoods, we can use the equation for the product of two Gaussians that we derived in chapter 2:

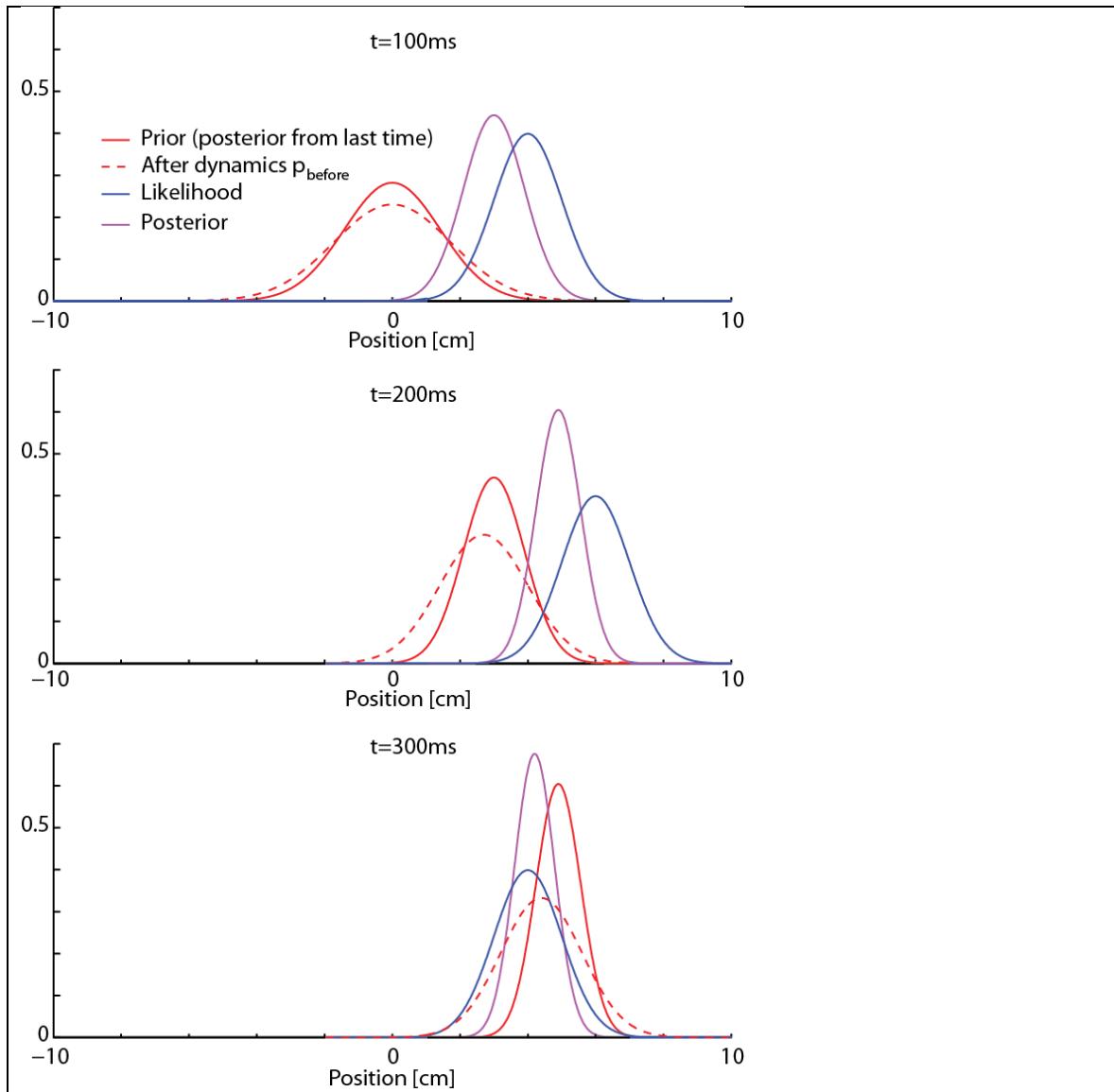
$$\mu_{t+100} = 0.9\mu_t(1 - W) + W * \mu_{\text{observe}}$$

Where  $\mu_{\text{observe}}$  is the mean of the observation and  $W = \sigma_t^2 / (\sigma_{\text{observe}}^2 + \sigma_t^2)$ . For the sake of this example we assume  $\sigma_{\text{observe}} = 1$ .

We obtain

$$\sigma_{t+100\text{ms}} = \sqrt{\frac{1}{1/\sigma_{\text{observe}}^2 + 1/(\sigma_t^2 + \sigma_\eta^2)}}$$

Thus, we have calculated the new mean and variance of our belief based on observations of the hand position, consideration of the dynamics, and the belief we had at the previous time step. In full analogy to the HMM, we can now continue the example.



**Figure 8.6.** The Kalman filter position estimation example.

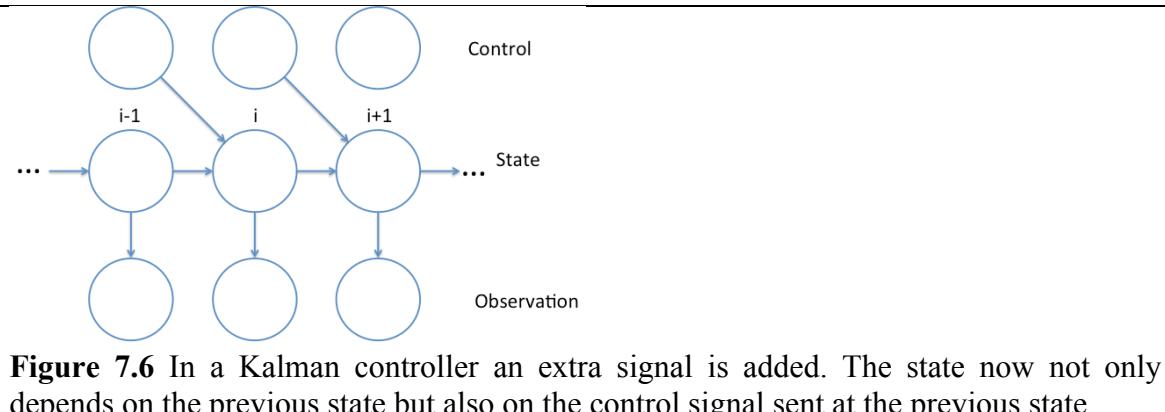
We start with  $p_0 = N(0, \sqrt{2})$ . After applying the Dynamics we obtain  $p_{t+100ms, \text{before}} = N(0 * 0.9, \sqrt{0.9^2 * 2 + 1}) \approx N(0, 1.61)$ . We then want to use the first observation of 4 cm. We calculate the weight as  $W = 1.61^2 / (1^2 + 1.61^2) \approx 0.72$ . We thus obtain  $\mu_t + 100ms = 0.72 * 4 + 0 = 2.88$  cm, and  $\sigma_t = 1 / \sqrt{(1/1.61^2 + 1/1^2)} \approx 0.72$  cm. As we should expect, we are both more certain than our prior uncertainty with  $\sigma = \sqrt{2}$  cm and also more certain than our observation with  $\sigma = 1$  cm. We thus obtain our full estimate after the first 100ms have passed:  $p_{t+100ms} = N(2.88\text{cm}, 0.72\text{cm})$

After applying the Dynamics we obtain  $p_{t+200ms, \text{before}} = N(2.88 * 0.9, \sqrt{0.9^2 * 0.72^2 + 1}) \approx N(2.59, 1.19)$ . We then want to use the second observation of 6 cm. This gives us the weight  $W = 1.19^2 / (1.19^2 + 1^2) \approx 0.59$ . We thus obtain  $\mu_{t+100ms} = 0.59 * 6 + 0.41 * 3 = 4.6$  cm, and  $\sigma_t = 1 / (1/1.19^2 + 1/1^2) = 0.59$  cm and  $p_{t+200ms} = N(4.6\text{cm}, 0.59\text{cm})$ .

At the next time step we obtain  $p_{t+300ms, \text{before}} = N(4.9 * 0.9, \sqrt{0.9^2 * 0.59^2 + 1}) = N(4.1, 1.1)$ . The third observation of 4 cm gives us the weight  $W = 1.1^2 / (1.1^2 + 1^2) \approx 0.55$ . We thus obtain  $\mu_{t+100ms} = 0.55 * 4 + 0.45 * 4.1 \approx 4.0$  cm, and  $\sigma_t = 1 / (1/1.1^2 + 1/1^2) \approx 0.55$  cm and  $p_{t+300ms} = N(4.0\text{cm}, 0.55\text{cm})$ .

We can see here that over time our knowledge about our position improved and hence the weight we put onto our observation decreased. This is a hallmark feature of the Kalman filter: as more information arrives we become more certain about our estimates and assign lower weights to feedback. Within the Kalman filtering field, the weight  $W = \sigma_{\text{state}}^2 / (\sigma_{\text{state}}^2 + \sigma_{\text{feedback}}^2)$  is called the Kalman gain, and is usually abbreviated as **K**. Importantly, nothing in the derivation of the Kalman gain uses the actual size of the observations: our best estimate of the uncertainty about the state is unaffected by the observations.

### 8.6.3 Extension to Kalman controllers

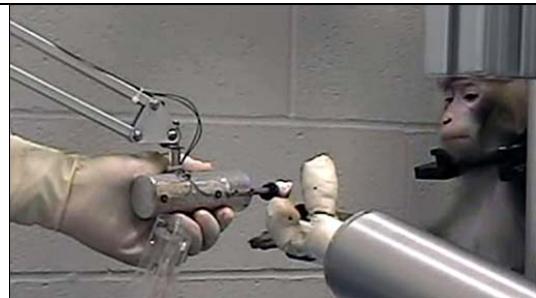


**Figure 7.6** In a Kalman controller an extra signal is added. The state now not only depends on the previous state but also on our own choice as expressed by a control signal  $u$ . In this case the dynamics equation is replaced by:

$$x_{t+1} = Ax_t + u_t + \eta$$

All other aspects of the inference remain unchanged. The Kalman controller is a regularly used extension of Kalman filters.

### 8.6.4 Applications of the Kalman Filter



**Figure 7.5** In brain-machine-interfaces, recordings from the brain are used to steer a machine. In this case, when the monkey thinks of moving, the pattern of action potentials recorded from multiple neurons in its motor cortex is interpreted by a filtering algorithm that steers a robotic arm. By controlling the robotic arm in this way, the monkey can feed itself. Image courtesy Schwartz lab.

There are two uses of the Kalman filter: It is being used to solve practical technical problems, and it is being used to model human behavior. When the Kalman filter is applied to solve technical problems, it is enough to build the model so that it works optimally, and there is no reason to compare it to actually measured distributions. One such practical application is in the context of brain machine interfaces. Suppose we want to learn how to use recorded neural activity from the motor cortex to drive a robotic arm (see Figure 7.5); the potential application for paralyzed humans is obvious. A necessary

step in developing such a system is to estimate how an animal moves its limbs through space, given recordings from neurons in its primary motor cortex. To do so, we can combine a model of the dynamics of hand movement with a model of how movement relates to neural activities. In this case, the neurophysiological recordings are our observations, and our goal is to produce optimal estimates of the time-varying trajectory of the hand based on these observations and on our understanding of limb dynamics. We can use a Kalman filter to solve this technical problem. In the exercises we will see some applications of this general idea.

However, if we want to ask how well a Kalman filter fits human behavior we could follow several different approaches. (A) We could calculate for every point in time how likely each decision would be as a function of the past decisions and stimuli. This would properly model the fact that subsequent choices are not independent from one another. (B) We could calculate the marginal choice distribution at every point of time fully modeling the noise in the system. (C) We could simply try to model mean behaviors by effectively setting observation noise to be zero in the simulation. This actually leads to unbiased estimates of the mean behavior but wrongly underestimates the width of the response distribution. The reason why the estimator is unbiased is because Kalman filter estimates are linear in the observations and hence unbiased noise has no effect on those estimates. In most cases where Kalman filters are fit to human behavior, only averages and not distribution are fit. This is because in many cases, the variance is affected by the factor that is modeled (e.g. human estimation) but also affected by other factors (e.g. motor noise) Therefore, typical models that are fit to human behavior generally use strategy (C).

The Kalman filter example that we worked through above dealt with linear dynamics, full observability and a one dimensional problem, but of course many real-world examples are considerably more complicated. Nevertheless, the Kalman filtering framework works equally well when the problem is phrased in multiple dimensions and not everything is observed at all times. There are many applications of Kalman filtering, and here we are only able to go through a small number of illustrative examples.

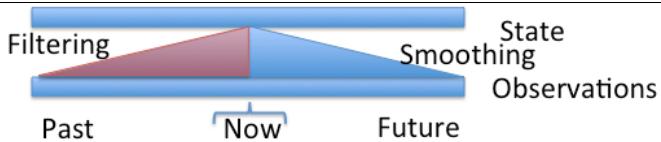
There are many models that hypothesize that the brain chooses movements optimally by estimating relevant variables in the environment in an optimal way. Many of these models use Kalman filters. For example, recent papers by the Shadmehr group have shown that when subjects estimate their hand position they put a higher weight on visual information when they have more uncertainty about their state and less if they have more uncertainty about their feedback, consistent with their use of a Kalman filter. These models set up the generative model (1), calculate the optimal inference (2), and compare that to human behavior using a step (3). According to simple Kalman filter models,

human subjects should converge more rapidly to a target when the information about the target is better, as then  $\sigma_{\text{observe}}$  is smaller and hence the Kalman gain  $\mathbf{K}$  is larger. Indeed, a clever experiment has shown people to have exactly this tendency (Izawa and Shadmehr 2010).

Another use of Kalman filters is in the context of multi-timescale phenomena. Many aspects in the world evolve over multiple timescales. For example, my muscles get weaker and stronger on a fast timescale due to fatigue and recover, on a medium timescale due to working out, and on a very long timescale through development and aging. Kalman filters can model such simulations by utilizing a statespace where each timescale is represented by a latent variable. There is ample evidence that behavior exhibits such multi-timescale properties.

In summary, filtering algorithms can generally be used when we want to make estimates based only on the past. Filtering algorithms can be divided into three parts: a prior that summarizes what we believe initially, dynamics that describes how the system changes over time, and observations. Dynamics generally make us less certain with the passage of time, while observations make us more certain. Filtering can be phrased as the simple alternation of dynamics, a marginalization process, and observations, via the application of Bayes rule.

## 8.7 Smoothing

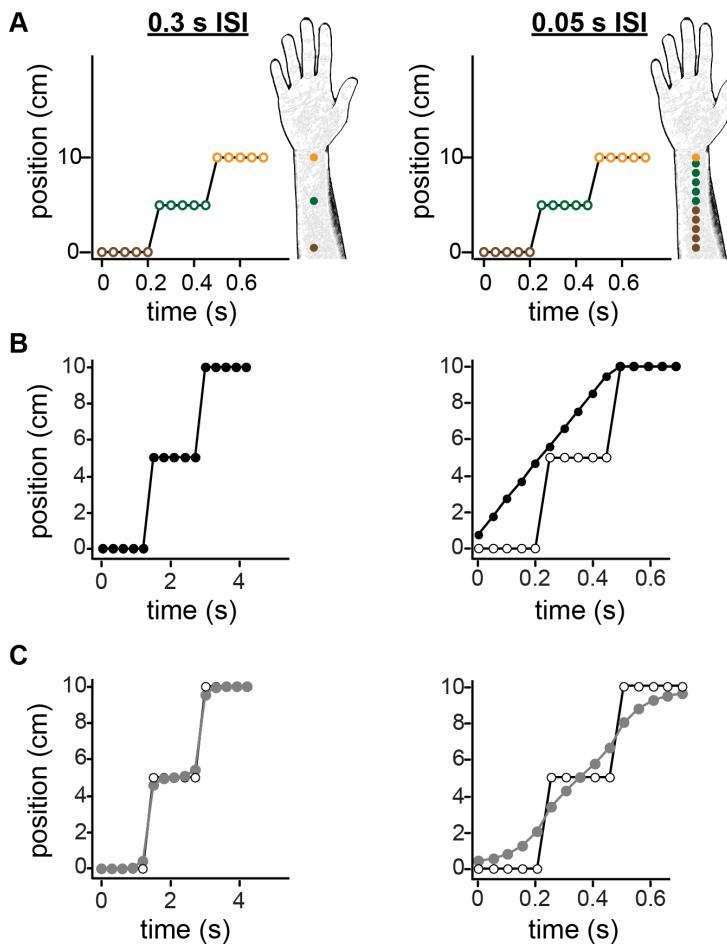


**Figure 7.2.** Filtering and Smoothing. In filtering, estimation of the current state is made using just information from the past. In smoothing, the estimation is based on both past and future observations, relative to the time point of interest.

As we have seen, filtering algorithms infer the present state of the world by considering not only current but also past observations and knowledge of the dynamics. Thus, filtering involves *prediction* from past to present. A procedure known as *smoothing* additionally incorporates information from the future relative to the time point in question. When we base our inference about the state of the world at a particular time point in part on observations that occurred after that time point, we are engaging in *postdiction* (Eagleman and Sejnowski, 2000). Thus, smoothing involves both prediction and postdiction.

Smoothing has been implicated in both visual and tactile motion perception. The brain's perception of the trajectory taken by a moving target is well-described by a smoothed

version of the observed trajectory (Rao et al., 2001; Goldreich and Tong, 2013). Specifically, smoothing has been implicated in the flash-lag illusion, in which a brief visual flash that occurs at the same location of a smoothly moving target is seen to lag behind the target (Rao et al., 2001; Soga et al., 2009). Smoothing has also been hypothesized to produce the perception of rapid sequences of taps to the skin. A Bayesian smoothing model with the expectation for low-speed movement qualitatively replicates illusions that occur when sequences of taps are delivered to the skin in rapid succession (Goldreich and Tong, 2013) (Figure XX).



**Figure xx.** Sequences of rapid taps to the skin are misperceived in the cutaneous rabbit illusion. **(A)** Qualitative reports of human subjects from Geldard (1982). At 0.3 s inter-stimulus interval (ISI), a sequence of 15 taps to the forearm, 5 at each of three skin positions, is perceived veridically. In contrast, at 0.05 s ISI, the first 10 taps are misperceived as progressively marching up the arm. Space-time plots: delivered stimuli. Forearm drawings: perception. **(B)** Plots showing the same results as in (A). Open circles: actual stimulus sequence; black filled circles: perception. **(C)**. Performance of a Kalman

smoothing model with an expectation for low-speed movement. Open circles: actual stimulus sequence; grey filled circles: model perception (MAP estimate) (Adapted from Goldreich and Tong, 2013).

Smoothing has many other applications. One of these is to facilitate learning. Let's say we made a mistake, thinking that a lion was in fact a stone. Suppose that some time after we concluded the lion was a stone, we observe that it is in fact a lion. If we survive this episode, we can use this "future information" to learn how to better interpret such scenes when they occur again. Concretely, the next time we observe a stone-like entity, we may realize, correctly, that it is a lion. Indeed, many computer learning algorithms propagate information from the future back in time to use it for learning. For learning, we want to use the best information we can obtain, which is generally the combination of past, present and future information.

Smoothing is also used by researchers in analyzing neural and behavioral data. Suppose we have recorded from an animal's brain while it is moving its arm, and we wish to make the best possible inference regarding the movement that the animal has been making. In this case, there is no reason for us to use only the neural activity that occurred before and during the movement, when information is also present in the activity that followed it. For example, neurons that represent the animal's perception of having moved might be very informative. In order to obtain accurate estimates of the animal's movement, we will generally be better off using data from both the past and the future.

Importantly, in smoothing the first step of modeling is identical to the first step in filtering. The generative model, our assumption of how the state of the world has dynamics, and how its state evolves over time, are the same in the two cases. The only difference is that instead of calculating  $p(\text{state}|\text{past,present})$ , as in filtering, we want to calculate  $p(\text{state}|\text{past, present, future})$ .

Before we continue we briefly want to mention the reasons that smoothing is called smoothing and filtering is called filtering. In traditional filtering, the value of the filtered signal is a weighted combination of past signals. As the Kalman filter does exactly that, albeit with potentially changing weights due to changes in our state uncertainty, it is called filtering. When smoothing signals, one generally considers both future and past signals. As Kalman smoothing does that it is called smoothing.

The second step of Bayesian modeling is now somewhat more complicated. After all, the estimate depends on the past and the future. So at first, one might be worried – is it necessary for every point of time to consider the future and the past, and wouldn't that lead to a very slow algorithm? It turns out that, aided by the Markov property, it is possible to calculate the solution to smoothing for all times by passing through time once

forward, and once backward. We will see below that this is a special case of belief propagation and that this can be generalized to a wide range of other problems.

For smoothing, we want to calculate  $p(state|past, present, future)$ , the probability distribution about the current state given the current sensory observation, sensory information from the past, and the future. Using Bayes' formula, and abbreviating all future sensory observations as *future* and all past sensory observations as *past* we can rewrite this as:

$$\begin{aligned} p(state|past, present, future) &\propto p(state)p(past, present, future|state) \\ &= p(state)p(present|state)p(past|present, state)p(future|present, state, past) \end{aligned} \quad (0.1)$$

And using the Markov property we can rewrite this as:

$$= p(state)p(present|state)p(past|state)p(future|state) \quad (0.1)$$

The inference thus factorizes into three terms, one depending on the present, one on the past, and one on the future. The basic idea behind the algorithm we will describe uses this factorization. One can think of the entire strategy as an implicit divide-and-conquer approach. We can further rewrite:

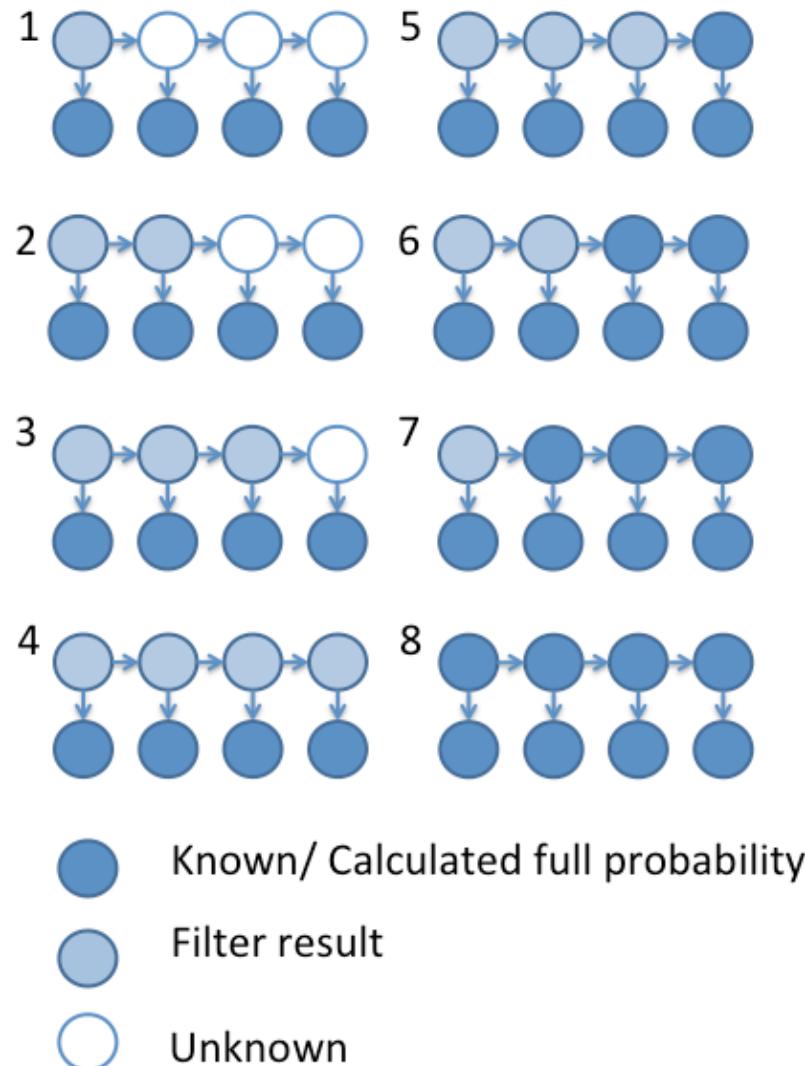
$$= p(state|present, past)p(future|state) \quad (0.1)$$

We have seen that we can calculate the first term for every point in time by going through the data just once by using a HMM or a Kalman filter for discrete and Gaussian cases, respectively. The question thus is just how we can calculate  $p(future|state)$ .

For the last point in time we can calculate  $p(future|state)=1$ . Hence we have no problem at that point of time. For the second but last time we observe that

$$\begin{aligned} P(state_{T-1}|present, past, future) &= \sum p(state_T|o_1 \dots o_T)p(state_{T-1}|present, past, state_T) \\ &= \sum p(state_T|o_1 \dots o_T)p(state_{T-1}|present, past)p(state|state_T) \end{aligned} \quad (0.1)$$

Now, the first term here, we just calculated and importantly, in the HMM case it is a discrete distribution and in the Kalman case it is a continuous Gaussian distribution. The second term, we calculated when we did our forward pass using a normal filtering algorithm, and the last term is the same class of transitions with inverse dynamics. Hence, the calculation takes on exactly the same form as in the forward case. In one of the problems, you will implement such an algorithm. Now, the same strategy that we used to use our belief at  $t$  along with the forward algorithm to calculate our belief using past and future at  $t-1$  can be iterated. For each point of time  $t$  we can calculate marginal distributions using this algorithm.



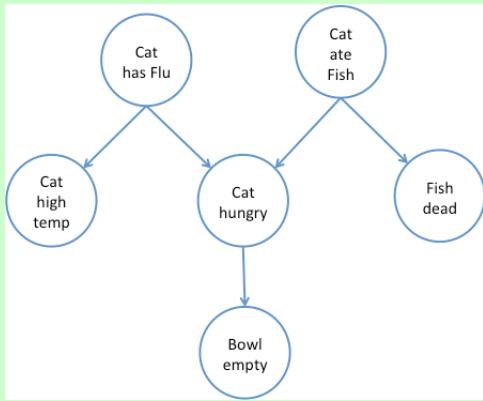
**Figure 7.6.** During smoothing, information is passed from the start of the chain forward, and then from the end of the chain backward. In the forward pass, filtering is performed; in the backward pass, the filtering results are adapted to take into account the future observations. Run-time is just twice that of the forward pass.

For cases where all distributions are either Gaussians or discrete distributions, smoothing just takes twice as long as filtering. Importantly, both algorithms for efficient filtering and algorithms for efficient smoothing are driven by the Markov assumption. This allows an effective divide-and-conquer strategy. Solve one part (past), then solve the future and iterate.

#### Box: Belief propagation (Experts)

It turns out that the basic idea of the forward backward algorithm used for smoothing can be generalized much further, leading to a class of algorithms called *belief propagation*. These algorithms allow an arbitrary graphical network that has the property that it has no

loops. We then order the nodes in a way that starts with one of them. Then we do the same kinds of updates as in the Kalman smoothing case. The information sent from one neuron to the next is now called a message. Again, the algorithm only requires a single forward pass through all the nodes where all messages are passed forward, and one backward pass where all messages are passed backwards.



Of course, if there were loops in the system then the messages would never end and would propagate forever. However, it turns out that for some classes of problems with circles the algorithm nonetheless converges to the correct solution (cite).

Algorithms based on Belief propagation are successfully used in many domains in computer science. For example, the best algorithms for sending data through noisy channels, e.g. a telephone modems, use belief propagation (cite turbocodes).

Belief propagation is an expert area but important because many problems in neuroscience and cognitive science can be phrased as belief propagation problems. The technical details go beyond the level of this book, but there are multiple specialized books that deal with the general topic (Jordan 1998; Bishop 2006)

## 8.8 Further reading

- Bishop, C. (2006). Pattern Recognition and Machine Learning, Springer.
- Eagleman, D.M., and Sejnowski, T.J. (2000). Motion integration and postdiction in visual awareness. *Science* 287, 2036-2038.
- Geldard, F.A. (1982). Saltation in somesthesia. *Psychol Bull* 92, 136-175.
- Gold, J. I. and M. N. Shadlen (2001). "Neural computations that underlie decisions about sensory stimuli." *Trends Cogn Sci* 5(1): 10-16.

- Goldreich D, Tong J (2013) Prediction, postdiction, and perceptual length contraction: a Bayesian low-speed prior captures the cutaneous rabbit and related illusions. *Front Psychol* 4:221 doi: 10.3389/fpsyg.2013.00221.
- Izawa, J. and R. Shadmehr (2010). "Online processing of uncertain information in visuomotor control." *Journal of Neuroscience*.
- Jordan, M. I. (1998). *Learning in Graphical Models*. Cambridge, MA, MIT Press.
- Rao, R.P., Eagleman, D.M., and Sejnowski, T.J. (2001). Optimal smoothing in visual motion perception. *Neural Comput* 13, 1243-1253.
- Soga, R., Akaishi, R., and Sakai, K. (2009). Predictive and postdictive mechanisms jointly contribute to visual awareness. *Conscious Cogn* 18, 578-592.

## 8.9 Problems

**Problem 8.1.** In the text above we have derived the equations of the Kalman Filter, where each latent variable is estimated based only on the information obtained so far. In analogy to the derivation of smoothing in the context of HMMs derive the equations and procedures for a Kalman smoother that makes estimates of Gaussian latent variables based on both past and future observations.

**Problem 8.2.** The drift-diffusion model describes how the belief of an optimal observer evolves over time. It is assumed that when the probability of the variable given the observation exceeds a certain threshold or is below a different threshold that the subject makes the decision. Derive the drift diffusion model, which may be easiest when framing it in terms of probability ratios. Is it necessary to assume that the log likelihood has a Gaussian distribution?

**Problem 8.3.** Can there be a Kalman filter with discrete or binary observations? What changes relative to the regular Kalman filter? Implement a simple version of such a Kalman filter in Matlab. Under what circumstances would such a strategy be a good idea and what kind of real world applications could you see?

**Problem 8.4.** Implement a Kalman filter on paper (as in the text) assuming Dynamics matrix  $A=1$ . Observation matrix  $C=0.5$ , Process noise  $Q=1$ , and observation noise  $R=1$ . Prior= $N(0,2)$ , Observations  $(2,4,2,1)$ . Which steps take how much time for you when doing it on paper? How long would a computer take for each of the steps?

**Problem 8.5.** Implement a simple one dimensional Kalman Filter in Matlab. Run it on a few simple scenarios. Now choose the constant  $A$  that characterizes the dynamics to be greater than 1. Describe the resulting estimates. What happens when you choose  $A$  to be much larger than 1? What happens if you choose very good observations?

**Problem 8.6.** Lets say we have a monkey moving and we obtain data from 100 neurons. We can record from each of the neurons. We can now assume a generative model where

the movement of the monkey makes neurons fire through a linear model. Is this model close to the way neurons actually work? Do you think it would work?

**Problem 8.7.** Implement the model sketched in (6) and apply it to data from the movement database (<http://crcns.org/data-sets/movements/cadre>, Stevenson dataset). If you want you can download a Kalman Filter toolbox to facilitate the implementation (<http://www.cs.ubc.ca/~murphyk/Software/Kalman/kalman.html>).

**Problem 8.8.** In many calculations we have to use convolutions. The convolution of two signals  $f(t)$  and  $g(t)$  is defined as:

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t - \tau)d\tau$$

Show that the convolution of two Gaussians, again is a Gaussian. Does that make Gaussians special? Are there other functions that have this property?

## Table of Contents

<b>8 Cost and reward .....</b>	<b>1</b>
8.1 Choosing among two actions .....	2
8.2 Choosing among several actions .....	4
8.3 Continuous cost functions.....	6
8.4 Choosing along one dimension .....	7
8.4.1 <i>The delta cost function and the posterior mode (MAP) readout .....</i>	9
8.4.2 <i>The quadratic cost function and the posterior mean readout .....</i>	9
8.4.3 <i>What it means to decide optimally.....</i>	10
8.5 Cost functions for uncertain outcomes .....	12
8.6 Evidence for optimal integration of outcome uncertainty and rewards in humans.....	13
8.7 Measuring cost functions.....	14
8.8 Optimal control: Continuous decision making over time.....	14
8.9 Further reading .....	16
8.10 Problems .....	16

## 8 Cost and reward

*How can we make optimal decisions when potential rewards are involved?*

In this chapter, we explore Bayesian decision theory, a framework for understanding optimal decision making. So far, this book has framed perception as a process in which the observer generates a posterior distribution over world states then reads out the most probable state (MAP estimate). We have taken the wisdom of the MAP readout largely for granted. In reality, however, the observer faces a *decision* in choosing which quantity to read out. Here, we use Bayesian decision theory to derive the optimal choice of posterior readout. We will show that a crucial function of the nervous system is not just to generate posterior distributions over world states but – based on those distributions and our preferences – to take actions in order to achieve beneficial outcomes. Thus, the generation of a posterior probability distribution is a first step in a decision process. We will illustrate Bayesian decision theory by considering daily decision-making examples, and we will then reframe perception itself as a decision-making process.

The scientific fields of optimal control, economics and decision theory each make use of cost functions in one form or another. In economics and decision theory, researchers often specify a *utility function*. This may be related to economics' traditional focus on goods versus effort: the higher the utility, the better for the actor. In optimal control, researchers usually specify a *cost* or *loss function*. This may be related to the field's focus on minimizing the energetic cost of producing movements or the fuel costs of rockets. However, in each of the

fields, there are positive and negative factors contributing to the function. Ultimately the two formulations are equivalent, as cost can be thought of as negative utility.

Plan of the Chapter: We begin the chapter by showing how the optimal decision depends on our probability distribution over world states and on our preferences over outcomes. We first consider how to make an optimally binary decision (e.g., should I take my umbrella or not). We next consider how to choose among several actions (e.g., where should I search for my lost keys?) or even a continuum of actions. Finally, we use Bayesian decision theory to revisit the deceptively simple process of deciding which value to read-out from a posterior distribution. We will learn that taking the MAP estimate (mode of the posterior) is not generally the optimal decision. Depending on our cost function – i.e., the penalty or reward associated with different outcomes – it may be optimal to read out the posterior mean, median, mode, or another value.

## 8.1 Choosing among two actions

Suppose that, as you prepare to leave home, you wonder whether it will rain. Combining a quick assessment of the cloudy sky (visual data) with knowledge of weather patterns in your area, you estimate the posterior probability of rain at 30%. Your posterior distribution --  $p(\text{it will rain} \mid \text{visual data, background knowledge}) = 0.3$ ,  $p(\text{it will not rain} \mid \text{visual data, background knowledge}) = 0.7$  -- represents your belief about the world state of interest, but it does not dictate how you should *behave*. You need to make a decision: should you carry an umbrella or not?

It might at first seem that, since you believe that an upcoming rainfall is less likely, you should simply leave your umbrella at home. However, it should become clear upon reflection that your decision whether to carry an umbrella will be based not only on your estimate of the chance of rain but also on the value you attach to different possible *outcomes* that could result from your choice of action. If you decide not to carry the umbrella, and it rains, then you will suffer the undesirable consequence of becoming wet. On the other hand, if you decide to carry the umbrella, and it does not rain, then you may feel inconvenienced by holding the unnecessary umbrella. An outcome may be undesirable, in which case we associate it with a *cost*, or desirable, in which case we associate it with a *utility*. Figure 9.1A illustrates the costs, specified by one individual, of the four possible outcomes in the umbrella problem.

		World State			
		Rain probability 0.3	No Rain probability 0.7		
Action	Do not take umbrella	wet & cold cost: 90	dry & happy cost: 0	Action	Do not take umbrella
	Take umbrella	slightly damp cost: 10	dry & encumbered cost: 5		Take umbrella
		Rain probability 0.3	No Rain probability 0.7		
		wet & cold cost: 40	dry & happy cost: 0		
	Do not take umbrella	slightly damp cost: 20	dry & encumbered cost: 10		
	Take umbrella				

**Figure 9.1.** On a 100-point scale, two people rank the unpleasantness (cost) of four possible outcomes. **A**, Walking in the rain without an umbrella results in a wet-and-cold outcome of cost 90; walking in the rain with an umbrella leaves one only slightly damp (cost 10); and so on. The optimal decision for this individual is to carry his umbrella. **B**, Another individual assigns different costs. The optimal action for this individual is to leave her umbrella at home.

Under the framework of Bayesian decision theory, the optimal behavior is the decision that minimizes expected cost, or, equivalently, maximizes expected utility. The expected cost is the cost associated with each possible outcome multiplied with the probability of that outcome. Referring to Figure 9.1A, we see that the expected cost of carrying the umbrella is:

$$\text{cost(umbrella)} = (0.3)(10) + (0.7)(5) = 6.5$$

That is, if we choose to carry the umbrella, we have a 30% chance of incurring a cost of 10, and a 70% chance of incurring a cost of 5. The expected cost of 6.5 can be thought of as the average cost that would result, if we chose to take the umbrella each day, over many days with weather identical to the current day.

In contrast, the expected cost of not carrying the umbrella is:

$$\text{cost(no umbrella)} = (0.3)(90) + (0.7)(0) = 27$$

If we choose not to carry the umbrella, we have a 30% chance of incurring a cost of 90, and a 70% chance of incurring a cost of 0. The expected cost of 27 can be thought of as the average cost that would result, if we chose not to take the umbrella each day, over many days with weather identical to the current day. Because  $6.5 < 27$ , the action that minimizes expected cost is to carry the umbrella; that is the optimal action for this individual.

Importantly, the cost or utility placed on an outcome reflects a personal preference; it is inherently subjective. Indeed, two people with the identical posterior distribution over the world state may choose opposite courses of action, because of the distinct values that they place on particular outcomes. To illustrate this point, consider a second individual who agrees that the chance of rain is 30%, but for whom the costs of the outcomes are different (Figure 9.1B). Note that the two people agree in their *ranking* of the outcomes from highest to lowest in cost, but they assign different numerical costs to the outcomes. For this second person, the optimal action is to leave the umbrella at home.

Exercise 9.1: Show this.

Mathematically, the optimal action  $a_{\text{optimal}}$  is defined as the action out of all possible actions  $a$  that minimizes the expected cost,  $EC(a)$ :

$$a_{\text{optimal}} = \underset{a}{\operatorname{argmin}} EC(a) = \underset{a}{\operatorname{argmin}} \sum_s \text{Cost}(s, a) p(s | x) \quad (8.1)$$

Here the world state,  $s$ , is any one of the things that may happen (e.g., it rains) and  $p(s|x)$  is the posterior probability of that world state happening, given the observation,  $x$  (e.g., a cloudy sky). The outcome results from the world state and the action taken ( $s,a$ ). For instance, we get wet (outcome) if it rains (world state) and we have not taken our umbrella (action).

## 8.2 Choosing among several actions

The procedure we have illustrated for selecting one of two actions applies equally to situations that require the selection of one of several actions. As an illustration, suppose that, upon getting a ride home from a picnic in the park with a group of friends, you realize that your house keys are not in your right pants' pocket, where you customarily keep them. Based on knowledge of your recent activities, you quickly generate a probability distribution over the location of your lost keys. They might have fallen out in your friend's car; you might simply have placed them in a different pocket; or you might have lost them at the park. Depending on where your keys actually are, and where you decide to search for them, nine outcomes are possible, and each of these has a particular cost to you (Figure 9.2). The decision you need to make is: where should you search first?

World State (location of keys)				
		your friend's car probability 0.1	another pocket probability 0.1	the park probability 0.8
Action	Search car	inconvenience friend, find keys! cost: -75	inconvenience friend, disappointing result cost: 15	inconvenience friend, disappointing result cost: 15
	Search pockets	easy to do, disappointing result cost: 11	easy, find keys! cost: -79	easy to do, disappointing result cost: 11
	Search park	big inconvenience, disappointing result cost: 90	big inconvenience, disappointing result cost: 90	big inconvenience, find keys! cost: 0
			<b>Costs</b> find keys: -80 easy to do: 1 inconvenience friend: 5 disappointing result: 10 big inconvenience: 80	

**Figure 9.2.** On a -100 to +100 point scale, the costs of different outcomes in the key search problem. Costs for each outcome were calculated by addition of the costs associated with each feature of the outcome (inset). If you decide to the search the car, you will need to call your friend and ask him to do that for you, which imposes an inconvenience on your friend and you do not like that (cost 5); if you decide to search the grass at the park where you picnicked, you will need to take a long trip back to the park, and probably spend lot of time searching there, a big inconvenience (cost 80). Thoroughly searching your other pants' pockets and the pockets in your coat is easy to do (cost 1). Searching and not finding your keys would be disappointing (cost 10). Finally, finding your keys would be very rewarding (cost -80).

The optimal decision will be the one associated with minimal expected cost. Computing the cost of each action, we have:

$$cost(car) = (0.1)(-75) + (0.1)(15) + (0.8)(15) = 6$$

$$cost(pockets) = (0.1)(11) + (0.1)(-79) + (0.8)(11) = 2$$

$$cost(park) = (0.1)(90) + (0.1)(90) + (0.8)(0) = 18$$

Thus, despite the fact that you consider it most probable that the keys are in the park, it is optimal to search first in your pockets.

### The drunkard and the lamppost

The tale of the drunkard who loses his keys walking from the bar to his car, but decides to search for them under a lamppost, presents an amusing case of suboptimal decision making. From a Bayesian decision theoretic perspective, the drunkard has correctly assigned low cost (high reward value) to the outcome (search under lamppost, find keys), because if his keys are there he is likely to encounter them easily and quickly given the light. However, he has failed to take into consideration that the probability is zero that his keys are in that location, since he never was near the lamppost to begin with.



Exercise 9.2: Suppose that, upon searching your pockets, you fail to find the key. Where would you search next? To find out, use Bayes' rule to update your probability distribution over location, given the new data that your keys are not in your pockets, then once again find the action that minimizes expected cost.

### Bayesian search

The example we have been considering is one of Bayesian search. Bayesian search has been used successfully to discover valuable lost objects; it is commonly used to search for ships lost

at sea. A famous example was the loss of the U.S. Navy's nuclear submarine, USS Scorpion, which disappeared during a voyage in the Atlantic in May, 1968, with 99 crew aboard. Interviews with Navy experts were used to assign probabilities to different scenarios that might have caused the sinking of the sub, and computer simulations were then run to define a prior probability distribution for the sub's location on the ocean floor. A search grid was constructed, with a prior probability assigned to each square on the map. Each square was associated with a particular cost, based in part on the difficulty of finding the sub if it were at that location; the seafloor differs in depth, and in some areas has narrow canyons that would increase the difficulty of the search. After a grid square was searched and the sub not found, the probability distribution over the map was updated, and the optimal search location recalculated. The USS Scorpion was found in October, 1968, in about 10,000 feet of water, approximately 400 miles southwest of the Azores islands.



### 8.3 Continuous cost functions

In the examples discussed above, we somewhat arbitrarily assigned a cost to each possible outcome. We were able to do this without too much work, because those situations involved only a limited number (e.g., 4 or 9) of possible outcomes. This was the case because both the world state variable and the action choice could take on only a limited number of discrete values. In contrast, many daily examples involve continuous world states and/or choices among a continuum of possible actions. We can no longer apply costs individually to every outcome in such cases; rather, we need to use a *cost function* over the space of outcomes.

For instance, returning to our umbrella example, we could consider a more nuanced scenario in which the world state can take a continuum of values reflecting the rate of the rainfall, from 0 (no rain) to 10 (extremely heavy rainfall). This results in a continuum of possible outcomes, each with a specific cost, even when we consider only two actions (to take or not to take the umbrella). If we tried to create an outcome table to represent this situation, we would have two rows (as in Figure 9.1 A or B) but an infinite number of columns. Clearly, another approach is needed here.

In such cases, we need to construct a function that reflects cost over the continuous space of outcomes for each action. Perhaps the cost to us of being in the rain grows linearly with the amount of water that lands on us. The cost function could then be expressed as  $cost(s, a) = (A + Ba)s$ , where  $s$  represents the rainfall rate,  $a$  represents our action (0 for carrying, and 1 for not carrying the umbrella), and  $A$  and  $B$  are constants. Thus, whether we have our umbrella or not, we would be increasingly displeased by greater rainfall, but the rise in our displeasure, as a function of rainfall rate, is greater when we lack the umbrella. We would then replace the sum in Eqn. (8.1) with an integral:

$$a_{\text{optimal}} = \underset{a}{\operatorname{argmin}} EC(a) = \underset{a}{\operatorname{argmin}} \int \text{Cost}(s, a) p(s | x) ds \quad (9.2)$$

To make the problem even more realistic, we could consider not just a continuum of rainfall rates, but also a continuum of possible actions reflecting not just our decision to take or leave the umbrella, but also our foot speed. We might walk slowly or attempt to minimize our exposure time to a possible rainfall by sprinting to our location (or go at any speed in-between). If we tried to create an outcome table to represent this situation (e.g., Figure 9.1), we would have an infinite number of rows and columns. Again, to deal with this situation we would need to specify a cost function over the space of outcomes.

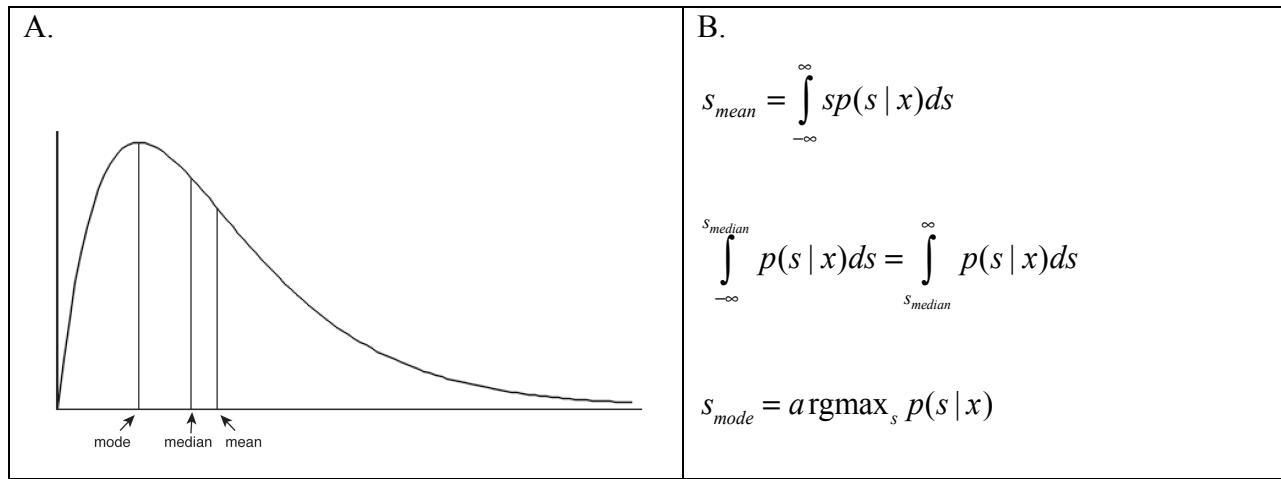
The specification of a cost function for real-life decision problems is a difficult task, although there are multiple obvious ingredients of the cost function. We want to satisfy our immediate needs and desires as well as progress towards more distal goals. This implies that obtaining food, drink, shelter, mating, the wellbeing of our family and other factors will have utility. At the same time, we do not want to overly exert ourselves, we want to minimize our energy consumption, the risk of damage to our body and other factors that are negative to our wellbeing.

## 8.4 Choosing along one dimension

Throughout the book we dealt with many situations in which one dimensional variables needed to be estimated: distances and positions, weights and speeds. Implicitly we had assumed that subjects just do Maximum-A-Posteriori (MAP) estimates of variables – that they go with the estimate that is most probable. However, taking the idea of cost functions seriously we can now come up with cleaner definitions of the objectives of estimation.

Suppose you want to estimate the location of a sound source. You indicate your estimate of the source location with a laser pointer (see Chapter 2). Clearly your goal is to point accurately, so you would feel more pleased if you succeed in pointing precisely to the source than if you point to any other location. Mathematically, this means that, if we represent the outcome as the difference between your the true source location and your estimate,  $outcome = s - \hat{s}$ , the minimum of your cost function (i.e., mode of your utility function) occurs at an outcome of zero. However, we still do not know the cost you associate with other outcomes: we do not yet know the shape of your cost function.

We can try to glean your cost function shape by asking you a series of questions, such as: How displeased would you be if you did not point exactly to the correct location? Would you grow more displeased with increasing error, and if so, how much more displeased? In the sections that follow, we consider three different answers to these questions, and derive the optimal posterior readout in each case. For the three cost functions that we consider, we will see that a different measure of central tendency of the posterior – the mode, the mean, or the median – is the optimal readout. Before we continue, we remind the reader of the definitions of these three measures of central tendency (Figure 9.3).



**Figure 9.3. A,** An asymmetrical probability density; this might represent an observer's posterior probability distribution over sound source location on a particular trial. The three lines, from left to right, represent the mean, median, and mode of the distribution. **B,** Definitions of mean, median, and mode. The mean is the center-of-mass of the distribution; the median divides the distribution into two halves of equal area; the mode is the point at which the distribution is highest.

### 8.4.1 The delta cost function and the posterior mode (MAP) readout

If you are a perfectionist, you might answer that you would be extremely unhappy to miss the correct sound location by any amount at all, no matter how small. In that case, your cost function could be approximated by a negative delta function centered on the outcome  $\hat{s} - s = 0$ :

$$C(s, \hat{s}) = -\delta(s - \hat{s}) \quad (9.3)$$

That is, the cost associated with being off by any amount is greater than the cost of being correct and once you are wrong it does not matter how wrong you are (For a description of the delta function, see Appendix 1). Equivalently, you just get a reward if you are exactly correct and we ignore the fact that the probability of being exactly correct is arbitrarily small. We can show mathematically that the optimal decision given this cost function is the posterior mode. We write the expected cost (EC) for the readout  $\hat{s}$  as

$$EC(\hat{s}) = \int_s C(s, \hat{s}) p(s | x) ds = \int_s -\delta(s - \hat{s}) p(s | x) ds = -p(\hat{s} | x)$$

This is because the delta function is zero unless  $s = \hat{s}$ . The expected cost of reading out a particular  $\hat{s}$  is the negative of the posterior probability over  $\hat{s}$ . The expected cost is therefore minimized when we choose to read out the mode of the posterior. Interestingly, according to this cost function, there never is a success as the probability of estimating perfectly is zero. And arguably, there are no real world tasks that need you to be perfect. Still, the MAP estimate is the optimal readout given the negative delta cost function.

#### The zero-one cost function

When the world state variable is discrete rather than continuous, the negative delta cost function can be represented equivalently as a *zero-one* cost function, in which only a single outcome (the estimate equals the true world state) has zero cost, and the others have a cost of 1 or any other positive value (the actual positive value is unimportant).

### 8.4.2 The quadratic cost function and the posterior mean readout

The MAP estimate is an intuitive choice of posterior readout, but as we have seen it results from a perfectionist cost function. After all, in the sound localization problem, the probability of pointing to precisely the correct location is infinitesimally small ( $p(\hat{s} = s) \sim 0$ ); in reality, we can at best hope that our estimate is close to the true location. We might be forgiven, then, for

admitting to be relatively pleased if we miss the target by only a small amount, with of course decreasing satisfaction the greater our error. Indeed, in many estimation problems, we can make rapid corrections if we are only a little off in our initial estimate, whereas larger deviations will require greater correction times. Such is the case, for example, in many motor tasks require pointing to targets or grasping objects. In such situations in movement research, the cost function is often assumed to be quadratic in the error:

$$C(s, \hat{s}) = (s - \hat{s})^2 \quad (9.4)$$

We can show that the optimal decision given this cost function is the posterior mean. We write the expected cost (EC) for the readout  $\hat{s}$  as

$$EC(\hat{s}) = \int_s C(s, \hat{s}) p(s | x) ds = \int_s (s - \hat{s})^2 p(s | x) ds$$

To find the value of the estimate that minimizes the expected squared error, we recall that where a function reaches the maximum its derivative is zero. The derivative of the expected cost with respect to  $\hat{s}$  is

$$\frac{\partial EC}{\partial \hat{s}} = \int_s 2(s - \hat{s}) p(s | x) ds = \int_s 2s p(s | x) ds - \int_s 2\hat{s} p(s | x) ds = 2s_{mean} - 2\hat{s}$$

This is zero when  $\hat{s} = s_{mean}$ . This cost function is often used in practice as the mean is easy to calculate, statistically reliable, and the cost function often approximates what subjects care about (be close to the target). The posterior mean is the optimal readout given the quadratic cost function.

### 8.4.3 What it means to decide optimally

We have seen that even for perception there are different plausible cost functions, and the procedure of minimizing expected cost will lead to a different read-out rule for each cost function. In the context of perceptual estimation problems, the choice of cost functions is

particularly important in the context of asymmetric posterior distributions (e.g., Figure 9.3A). For symmetric unimodal distributions such as the Gaussian distribution, the mean, median and mode are identical. For asymmetric distributions, however, they are distinct. Furthermore, some distributions, while symmetric, are bimodal. In Chapter 7, when discussing causal inference, we encountered bimodal (two-peaked) posterior distributions. For such distributions, the mean and mode are generally distinct and research has suggested that human subjects report closer to the mean of such a bimodal posterior (Kording et al., 2007).

We note that the cost functions we have considered here – e.g. delta and quadratic cost – are all symmetrical functions; in some situations humans will naturally adopt asymmetrical cost functions. Suppose you are hiking in foggy weather on a mountain trail with a cliff on your right side. From your limited visual information, you can estimate your distance from the cliff with some uncertainty. You then need to decide where along the trail to walk, choosing from among a continuum of possible headings. The center of the path may be most comfortable on your feet. Making an error by veering off towards the left is relatively harmless; perhaps the path becomes rockier. An error to the right, in contrast, could be fatal. Because the costs are asymmetric, the optimal decision – the one that minimizes expected cost – will be to bias your position towards the side of the path that is farther from the cliff.

Under any cost function, the posterior distribution is what needs to be used to calculate optimal behavior. So far, we have argued that given the posterior, the cost function dictates which readout of the posterior is optimal. However, why would the posterior itself be needed for optimality? Why could the observer not use some other probability distribution over the stimulus computed from the observations, say  $q(s|x)$ ? A situation where this might happen is that the observer does not know the correct structure of the generative model. For example, in the sound localization task, the observer might use an incorrect prior over the position of the sound source. This would give rise to a different posterior,  $q(s|x)$ . Such a wrongly estimated posterior distribution could be used for decision making.

Using a  $q$  that differs from  $p$  will in general not minimize cost across trials. Formulated very precisely: say  $s$  and  $x$  follow a joint distribution  $p(s,x)$ . Suppose the observer is free to follow ANY estimation strategy to go from  $x$  to  $\hat{s}$ , let us say some function  $F$ . The cost incurred on a single trial is then  $C(s,F(x))$ , and the expected cost incurred across trials is

$$\text{TotalCost} = \sum_{s,x} C(s,F(x)) p(s,x)$$

Then the  $F$  that minimizes total cost is the one that minimizes expected cost (EC) under the true posterior,  $p(s|x)$  (and not under some other distribution  $q(s|x)$ ). To see this, observe that for this  $F$ , the quantity  $\sum_s C(s,F(x)) p(s|x)$  is as low as it can possibly be. Since this statement is true for any  $x$ , it follows that it is true when we average over  $x$ . Thus,  $F$  is such that

$\sum_x p(x) \sum_s C(s, F(x)) p(s|x)$  is as low as it can possibly be. But this quantity is exactly equal to TotalCost above. As a particular consequence, using the “true” posterior  $p(s|x)$  (i.e. based on the true generative model) rather than some substitute  $q(s|x)$ , will guarantee minimization of the total cost, i.e. optimality. In this sense, the computation of the Bayesian posterior is the essence of optimal inference.

### Data Processing Inequality

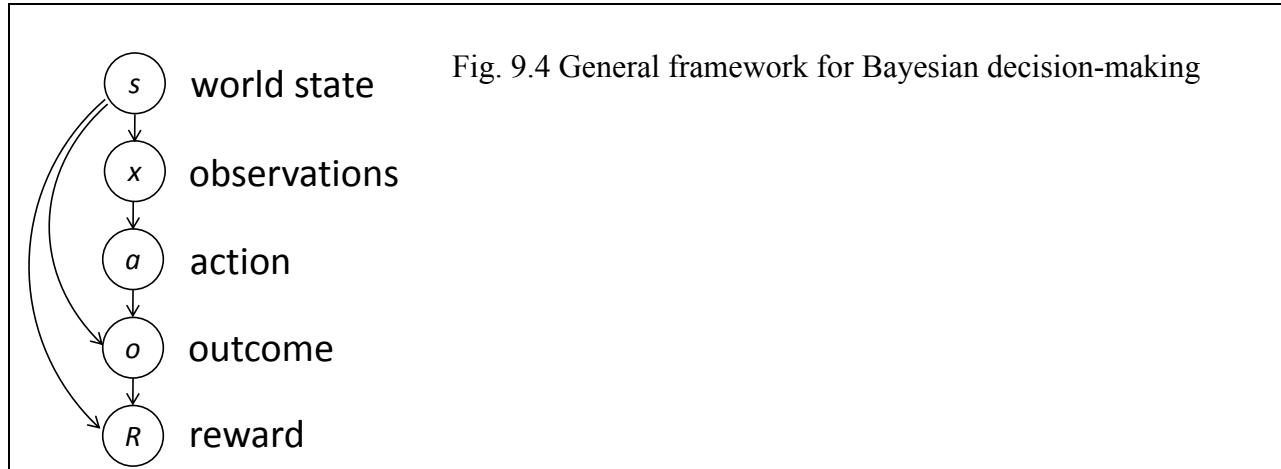
Suppose I have a measurement  $x$ . Is there any way I could transform  $x$  to some other representation  $y$ , such that based on  $y$  I could achieve systematically better performance (lower expected cost) than based on  $x$ ? Intuitively, this is not possible: you cannot create information “out of nowhere”. Formally, this is known as a data processing inequality. The original data processing inequality comes from information theory, but in the previous subsection we have essentially given a proof in a Bayesian context. In the Bayesian decision theory context, information refers to the lowest achievable expected cost. We saw that given a representation  $x$  of  $s$ , no strategy can achieve a lower expected cost than the Bayesian strategy of minimizing expected cost. First processing  $x$  to  $y$  and then applying some rule to  $y$  is just another strategy and it can thus not be better than the Bayesian strategy. In terms of the brain, this means that in subsequent layers of processing, information can only be lost, not gained: based on the cortical representation of a visual stimulus, for example, it is not possible to systematically outperform the optimal Bayesian estimator based on the activity in the retina.

## 8.5 Cost functions for uncertain outcomes

In the examples we have considered so far, each combination of world state and action has mapped deterministically onto a single outcome, to which we assigned a cost. In reality, however, it is often the case that many outcomes could result – with different probabilities – from a particular combination of action and world state. For example, in Figure 9.1 we assumed that, if we were to leave the house without our umbrella (action) and it were to rain (world state), then we would get wet (outcome). In reality, however, a range of possible outcomes might result from this combination of action and world state, depending probabilistically on the availability of places along our path under which we could take cover from the rain. In such situations, given a particular combination of action and world state, one can define a probability distribution across outcomes:  $p(o|a,s)$  (Fig. 9.4). As usual, each outcome ( $o$ ) is associated with a cost, and the optimal decision-maker would choose the action  $a$  that minimizes expected cost. Thus, the formula for the optimal action (Eqn. (9.2)) generalizes to:

$$a_{\text{optimal}} = \underset{a}{\operatorname{argmin}} EC(a) = \underset{a}{\operatorname{argmin}} \left( \int_s \left( \int_o C(o) p(o | s, a) do \right) p(s | x) ds \right) \quad (9.5)$$

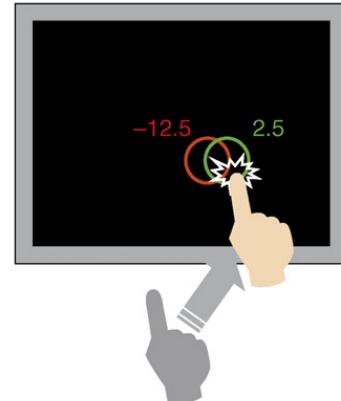
In other words, you average your cost function both over possible outcomes given your hypothesized action and over possible states of the world.



## 8.6 Evidence for optimal integration of outcome uncertainty and rewards in humans

Often, when we are examining the behaviors of a human subject or animal, we want to compare them against what is optimal. We can compare predictions from Bayesian Decision theory with actual behavior. Yet, there is a fundamental hurdle in this comparison. For this type of analysis to be constructive, we must know beforehand what it is the experimental subject is trying to achieve, i.e., what their cost function is. With this in mind, researchers have designed experiments where the cost of the task is relatively explicit. In a set of reaching studies, it was observed that people are remarkably close to the optimal choices prescribed by decision theory when final positions translated into known monetary gains and losses (Trommershäuser et al., 2003, Trommershäuser et al., 2005, Maloney et al., 2006). Similar experiments have studied visual tasks (see Whiteley and Sahani, 2008) and force-producing tasks (Kording et al., 2004, Kording and Wolpert, 2004a; see also, Section 2: Computational neuroscience models). This further demonstrates people's abilities to integrate statistical information in a Bayes-optimal manner, not just for estimation, but also for action selection.

Julia Trommershäuser and colleagues used the following experiment to study movement under uncertainty. Imagine you have to make a pointing movement to inside a small green circle on a screen, but avoid the inside of a small red circle. If you hit inside the green circle but not inside the red, you earn 2.5 points. If you hit inside the red circle but not inside the green, you lose 12.5 points. If you hit in the intersection of both circles, you lose 10 points. If you



hit outside both circles, you earn 0 points. If you take too long, you earn 0 points as well, so you are forced to move fast. Assume the radius of the circles is 1 and the distance between the centers is  $D$ . Also assume that the place where you hit the screen is not exactly where you aim, because it is corrupted by movement noise. This noise has a two-dimensional Gaussian distribution with standard deviation  $\sigma$ . There is a point where you should aim to maximize the expected number of points you earn. This point will depend on the costs and rewards, as well as on the separation of the circles and your own motor noise (see Problems). Humans learn to optimally take into account these quantities to decide on their optimal reaching point.

### 8.7 Measuring cost functions

Cost functions are hypothesized functions that human subjects are supposed to optimize. There are some known properties of cost functions that are important. Of primary importance is the fact that cost functions cannot be uniquely measured. Suppose we have a cost function  $C(o)$  that a given subject is trying to optimize through a choice of action. Then  $\theta(C(o))$ , where  $\theta$  is any monotonic function, leaves the preference of one action over another action invariant. Hence, a person using  $C(o)$  and another using  $\theta(C(o))$  will have the same set of preferences over actions. As the cost function is only observable indirectly through preferences over actions, it is impossible to distinguish cost functions from their monotonic transformations. Nevertheless, while it is not possible to measure cost functions directly, it is possible to measure indifference curves. If we have a multidimensional cost function we can produce games where we ask about which pairs of outcomes subjects are indifferent to. Through many such measurements, cost functions can be reasonably well characterized. Using inverse decision theory, recent studies have examined implicit cost functions subjects use when performing motor tasks and when penalizing target errors (Kording and Wolpert, 2004b, see Fig. 4). These inferred cost functions have highly nonlinear and nontrivial forms (Todorov, 2004). These findings highlight a crucial problem in decision theory: good fits to behavior may be obtained with incorrect cost functions. Inverse decision theory can thus be used as a means of searching for violations in the assumptions we make using the Bayesian approach to decision making.

### 8.8 Optimal control: Continuous decision making over time

The central objective of the nervous system is to continuously change the state of the world around us so that it is more favorable to us – and arguably this is the only task of the nervous system. This is somewhat complicated because we have perceptual uncertainty about the world and because the world itself is a dynamical system. We cannot choose its state but just locally change its dynamics – hopefully changing the world’s future to suit us. In the context of optimal control there are two classes of costs: the accumulating costs that generally depend on what we do and the task related costs and rewards. For example, moving our hand costs energy but bringing food to our lips is rewarding. We can express this when  $TC$  is the total cost:

$$TC = \int_{t_0}^{t_{max}} Cost(x(t), u(t)) dt$$

where  $x(t)$  is the time-varying state and  $u(t)$  the time varying control. Importantly the control signal  $u(t)$  itself depends on the observations that were made. The optimal control problem thus boils down to choosing a  $\$u(t)\$$  which depends on sensory information such that it minimizes the resulting total cost.

The general problem problem is known to be very difficult and the theory of optimal control fills many books(Bertsekas, 2001). Important to know for students of Bayesian statistics is that there are classes of problems that can actually be solved very efficiently. For example, as long as all cost functions are quadratic, the noise is Gaussian and the dynamics are linear there exist analytical solutions to the relevant problems (Kalman, 1960). There are classes of problems that can be efficiently solved by estimating the cost to go (the expected cost that will accumulate over the rest of the trial), usually by solving the Hamilton-Jacobi-Bellman equations. And lastly there are many approximate methods for various special cases (Stengel, 1994, Todorov and Tassa, 2009). While optimal control squarely falls into the domain of Bayesian decision theory its details go beyond the focus of this book.

### History of Utility theory

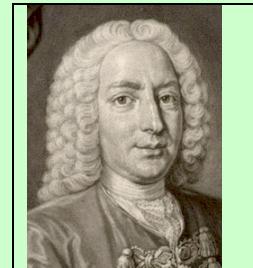
Daniel Bernoulli and before him Gabriel Cramer worked on the St. Petersburg paradox. This paradox relates to a simple game of chance: on every round, a fair coin is thrown, and the game ends as soon as the coin lands tails (T). The subject then gets  $2^h$  dollars, where  $h$  is the number of heads (H) observed. For example, if the sequence of landings observed was T, the player would win \$1; HT, \$2; HHT, \$4; HHHT, \$8; and so on. It turns out that the expected payoff of this game is infinite. The expected payoff is the sum over all possible outcomes of the payoff times the probability of the outcome. Representing each possible outcome by its number of heads, we have:

$$\text{Expected payoff} = \sum_{h=0}^{\infty} p(h)2^h = \sum_{h=0}^{\infty} \left(\frac{1}{2}\right)^{h+1} 2^h = \sum_{h=0}^{\infty} \frac{1}{2} = \infty$$

The paradox is that, although the expected payoff is infinite, people would not pay more than a few dollars to enter the game. A way to resolve this paradox is through the introduction of a utility function reflecting the decreasing marginal utility of money: 1,001,000 dollars is only slightly more valuable than 1,000,000 dollars, whereas 1,000 dollars is much more valuable than nothing.

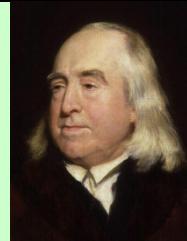


Gabriel Cramer



Daniel Bernoulli

Jeremy Bentham, another inventor of the idea of utility, applied it more directly to the pleasures and pains of humans. He used these ideas to derive how society should be organized – namely by maximizing the utilities of all the citizens, the philosophy of utilitarianism. Bentham saw all moral and legal norms as derivable from this simple principle using methods from logic and experimentation.



Jeremy Bentham

## 8.9 Further reading

- Bertsekas D (2001) Dynamic Programming and Optimal Control Bellmont, MA: Athena Scientific.
- Kalman RE (1960) Contributions to the theory of optimal control. *Bol Soc Mat Mexicana* 5:102-119.
- Kording KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L (2007) Causal inference in multisensory perception. *PLoS ONE* 2:e943.
- Kording KP, Fukunaga I, Howard IS, Ingram JN, Wolpert DM (2004) A neuroeconomics approach to inferring utility functions in sensorimotor control. *PLoS Biol* 2:e330.
- Kording KP, Wolpert DM (2004a) The loss function of sensorimotor learning. *Proc Natl Acad Sci U S A* 101:9839-9842.
- Kording KP, Wolpert DM (2004b) The loss function of sensorimotor learning. *Proc Natl Acad Sci U S A* 101:9839-9842.
- Maloney LT, Trommershäuser J, Landy MS (2006) Questions without words: A comparison between decision making under risk and movement planning under risk. In: *Integrated Models of Cognitive Systems* (Gray, W., ed) New York , NY: Oxford University Press.
- Stengel RF (1994) *Optimal Control and Estimation*: Dover Publications.
- Todorov E (2004) Optimality principles in sensorimotor control. *Nat Neurosci* 7:907-915.
- Todorov E, Tassa Y (2009) Iterative local dynamic programming. In: *Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, 2009, pp 90-95: IEEE.
- Trommershäuser J, Gepshtein S, Maloney LT, Landy MS, Banks MS (2005) Optimal compensation for changes in task-relevant movement variability. *J Neurosci* 25:7169-7178.
- Trommershäuser J, Maloney LT, Landy MS (2003) Statistical decision theory and the selection of rapid, goal-directed movements. *Journal of the Optical Society of America A*, 20:1419-1433.
- Whiteley L, Sahani M (2008) Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. *J Vis* 8:2 1-15.

## 8.10 Problems

**Problem 9.1.** Suppose that you are walking in the dark in an area that is occasionally inhabited by lions. You hear a suspicious noise that may indicate the presence of a lion. In deciding whether to run away or not, you apply the following cost structure to each of 4 possible outcomes:

**World State**

		Lion		No Lion	
		Run	Stay	Run	Stay
Action	Lion	physical effort no injury cost: 2	physical effort no injury cost: 2	no effort severe injury cost: 100	no effort no injury cost: 0
	No Lion	physical effort no injury cost: 2	no effort no injury cost: 0	no effort no injury cost: 0	no effort no injury cost: 0

Lion or just noise?

- a) Given the cost structure shown, if you believe that there is only a 30% chance that the lion is present, should you run or stay?
- b) How low would your belief in the lion's presence have to be before you decided to stay rather than to run?
- c) The outcomes listed in the table above represent a simplistic view of the problem. To make the problem more realistic, consider the following: In reality, if the lion is present, it may catch you even if you run; if you stay, the lion might (with some probability) decide not to attack you; if you run, you have some probability of injuring yourself by falling or colliding with objects (trees, boulders). Try to modify your solution to the problem, taking into account these realistic considerations (In your calculations, you may use what you believe to be realistic probabilities for each of these contingencies).

**Problem 9.2.** Prove that the posterior median is the optimal readout given the absolute error cost function,  $C(s, \hat{s}) = |s - \hat{s}|$ .

Keep in mind, that at the maximum of a function, the first derivative must vanish.

**Problem 9.3.** This problem refers to the Trommershauser outcome uncertainty task in Section 8.6.

- a) Derive an equation for the optimal point of aiming. (This is not a closed-form expression.) For simplicity, you can regard this as a one-dimensional problem, even though it is really a two-dimensional problem.
- b) In Matlab, choose a reasonable range of values for  $D$  and a reasonable range of values for  $\sigma$ . Then numerically compute the optimal point of aiming as a function of the separation  $D$  and as a function of  $\sigma$ . Plot this optimum as a function of  $D$  and  $\sigma$  using a color plot ("imagesc" in Matlab). Interpret your plot.

**Problem 9.4.** The woodchuck does not actually chuck wood, it rather lives of grass, herbs and insects. Every day, again and again, it has to allocate its resources between various possible activities. Lets say you have records what a woodchuck does, e.g. where it walks, how fast, and what it eats and when. How would you build a normative model of why the woodchuck behaves the way it does and how would you test that with data?

**Problem 9.5 (Advanced).** Implement a linear quadratic regulator in matlab that stabilizes a simulated inverted pendulum.

**Problem 9.6.** Show that for any symmetrical posterior, if there exists only a single optimal solution that the cost function has no influence on the best estimate.

**Problem 9.7.** Certain cost functions (e.g. quadratic loss) lead to behavior that is less variable than other cost functions (e.g. absolute error loss). Calculate the expected variance in one dimensional estimates for these two estimators. Simulate the same system in MATLAB. Do the theoretical estimates of variance match up with the empirical ones?

### Problem 6.X

we introduced stimulus variables that take values on the circle, such as motion direction. Assume a circular variable  $s$ . The posterior is  $p(s|\text{data})$ .

- a) For estimation of a circular variable, using the squared error between the estimate,  $\hat{s}$ , and the true value of the variable as a cost function does not make sense. Why not? Explain as concretely as possible.
- b) A sensible cost utility function is the cosine of the estimation error,  $U(\hat{s}, s) = \cos(\hat{s} - s)$ . Show that the estimate that maximizes expected utility on a given trial is the circular mean of the posterior, denoted  $\mu_p$ , which is defined by the equations

$$\cos \mu_p = \langle \cos s \rangle$$

$$\sin \mu_p = \langle \sin s \rangle$$

where  $\langle K \rangle$  denotes the expected value under the posterior.

**Problem 2.11. THE FIRST PART OF THIS PROBLEM WILL GO INTO CHAPTER 3. HERE REFER AND DO PARTS D and E** Assume the stimulus and measurement distributions of Chapter 2, but an observer who uses a Gaussian prior with standard deviation  $\sigma_p \neq \sigma_s$  (but the correct  $\mu$ ). This observer's MAP estimate can be thought of as being based on an incorrect generative model.

- a) When the measurement is  $x$ , what is the MAP estimate?
- b) What are the mean and variance of the MAP estimate for given  $s$ ?
- c) What are the mean and variance of the MAP estimate across all trials in the experiment? (Hint: the expression should contain both  $\sigma_p$  and  $\sigma_s$ .)
- d) What is the mean squared error of this observer?
- e) Is this mean squared error always larger than that of the MAP observer using the correct prior, always smaller, or does it depend? Is this answer expected, and if so, why?

**Problem 13.1.** In the birds-on-a-wire example from Section 1.5.1, suppose the five birds sing with frequencies  $p_1$  through  $p_5$ . You get 1 point for each time you correctly guess which bird is singing. Show that MAP estimation (under the correct generative model, which means incorporating knowledge of  $p_1$  through  $p_5$ ) maximizes the number of points you receive over many trials.

**Problem 13.7.** One could calculate the correlation coefficient between the estimate  $\hat{s}$  and the stimulus  $s$  across all trials in the experiment. In the example of Chapter 2, is this correlation coefficient maximized by the MAP estimator?

## Table of Contents

9	Basics of neural coding.....	
9.1	A focus on generative models of neurons.....	3
9.2	Neurons as mappings from an input to an output .....	4
9.3	Tuning curves .....	4
9.3.1	Bell-shaped tuning curves.....	6
9.3.2	Monotonic tuning curves.....	7
9.4	Variability.....	8
9.4.1	Poisson variability.....	9
9.4.2	More realistic models.....	11
9.4.3	The origin of variability .....	12
9.5	Population codes.....	12
9.5.1	Independent Poisson variability in a population .....	15
9.5.2	Numerical example .....	15
9.6	Further reading.....	16
9.7	Problems .....	18

## 9 Basics of neural coding

*How do neurons represent the world?*

At first glance, there seems to be a world of difference between fundamental physiological features of neurons, such as firing rates and tuning curves, and the quantitative measurements of perception and behavior that we have described in the previous chapters. Yet we know that somehow, neuronal processes must underlie all of perception and behavior. The goal in this chapter and the following is to indicate how this gap can be bridged. We will discuss, if somewhat speculatively, how neurons could implement Bayesian computations.

So far, we have treated the brain as a black box that performs probabilistic inference. This suffices if one is primarily interested in modeling behavior.

However, in systems neuroscience, elucidating the link between biological and psychological states is a central objective. In this chapter and the next, we explore the connection between behavior and neural activity from the perspective of our normative framework. As we have seen that abundant evidence exists for Bayesian optimality at the behavioral level, at least in simple tasks, we will ask the following questions:

- How do neurons represent states of the world?
- How do neurons represent likelihood functions?
- How do neurons use these representations to calculate posterior distributions?

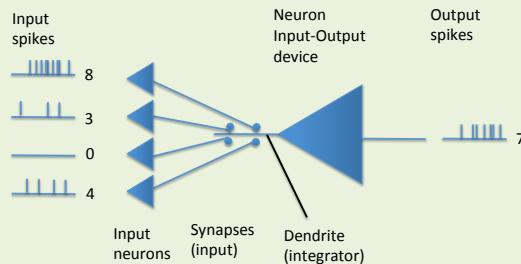
The Bayesian normative framework offers a means of addressing these questions that runs counter to the bulk of neural modeling work. Modelers often construct neural networks out of simulated, more or less biophysically plausible elements, and examine their emergent dynamics. In many cases, this “bottom-up” approach has the disadvantage that the link of these networks to behavior is tenuous or only qualitative. In contrast, it is possible to take a top-down approach to neural computation, in which the construction of a neural model is guided by a normative behavioral model. In this approach, the search for a neural implementation of a perceptual phenomenon is guided simultaneously by behavioral and physiological constraints.

Plan of the chapter: We will first computationally describe neurons, formalizing them as a generative model that produces outputs in response to either direct inputs or to a stimulus given to the brain. We will introduce the concepts of a neural population and neural variability. Using these concepts, we can understand how neurons can encode a probability distribution and therefore can carry implicit knowledge of uncertainty. As an introduction to neural modeling and population coding, the present chapter is limited in scope. Our main goal here is to provide the necessary background information for an understanding of the representation of probability at the neural level.

**Box 9.1: Biology of a neuron:** Neurons as input-output devices. Neurons transmit electrical impulses (action potentials, aka *spikes*) away from their cell bodies, along output structures (axons). In a majority of cases in the mammalian nervous system (the so-called chemical synapse), when the spike reaches the end of an axon (axon terminal), it induces the release of neurotransmitter molecules that diffuse across a narrow synaptic cleft and bind to receptors on the input structure (dendrites) of the postsynaptic neuron. The effect of the transmitter released by a given input neuron may be inhibitory (reducing the probability that

the postsynaptic neuron will fire spikes of its own) or excitatory (increasing the probability that the postsynaptic neuron will fire), depending on the transmitter released and on the receptor that binds it.

Each neuron receives input from a (typically large) number of other neurons. This figure shows a single postsynaptic neuron receiving just four inputs for simplicity. Each of the input neurons fires a number of spikes over a relevant time interval. These result in neurotransmitter release onto our neuron of interest. The postsynaptic neuron integrates these inputs and produces an output spike train of its own. In this book, we will simplify the modeling of the neuron to modeling the *number* of output spikes, either in response to a stimulus or in response to the numbers of input spikes it receives from other neurons.



## 9.1 A focus on generative models of neurons

In earlier chapters, when we defined the generative model with respect to world states and sensory observations, we conveniently represented the sensory observation as a measurement that lived in the same space as the stimulus. For instance, we conceived a sound stimulus at a particular location as producing an observation drawn from a Gaussian distribution centered at that location. At the neurobiological level, however, the sensory observation is not such a measurement, but rather neuronal activity evoked by the sensory stimulus: neurons encode the physical stimulus as a pattern of spiking activity, and the brain must somehow decode this activity in order to infer the world state. Here we take a first look at the neuronal level, and we consider the mapping from sensory stimuli to the spiking activity produced in sensory neurons. Once we fully specify how sensory stimuli give rise, in a probabilistic way, to neural activities, we will be in a position to formulate how neurons may encode uncertain stimuli and how the brain can infer the state of the world from neuronal activity.

The brain does not have direct access to the sensory input,  $I$ . Rather, the sensory input activates receptor cells such as auditory hair cells or photoreceptors, which in turn activate nerve fibers (axons), causing electrical impulses (action potentials or spikes) to travel into the central nervous system. These impulses are

the data upon which the brain makes inferences about the world. The activity of neurons in a relevant brain area in response to a stimulus, denoted  $\mathbf{r}$ , constitutes the internal representation or observation of that stimulus. Neural activity is variable: when the same sensory input  $I$  is presented repeatedly,  $\mathbf{r}$  will be different every time. This is due to stochastic processes that inject variability: photon noise, stochastic neurotransmitter release, stochastic opening and closing of ion channels, etc. Thus, a probability distribution  $p(\mathbf{r}|I)$  is needed to describe the neural activity.

Since activity  $\mathbf{r}$  is variable even when the input  $I$  is kept fixed, and  $I$  is variable even when the stimulus  $s$  is kept fixed, it follows that  $\mathbf{r}$  is variable when  $s$  is kept fixed, even if the nuisance parameters are not variable. This neural variability is captured in a probability distribution  $p(\mathbf{r}|s)$ . The main goal of this chapter is to define and motivate mathematical descriptions of  $p(\mathbf{r}|s)$ . We would like you to think of  $\mathbf{r}$  as the spiking activity in early sensory cortex, such as primary visual area V1. The stimulus is a basic stimulus feature such as position, orientation, etc. – similar to all preceding chapters.

## 9.2 Neurons as mappings from an input to an output

There are many ways of modeling neurons, ranging from detailed biophysical models of the structure and function of individual ion channels, to highly abstract models of neurons as information processing units. For the purpose of this book, we treat neurons as simple input-output systems. A neuron receives inputs from a group of other neurons. Over the relevant time scale it receives a number of spikes from each of the input neurons and produces (or emits) a number of output spikes. A neuron is thus characterized by its transfer function, spikes =  $f(\text{input spikes})$ .

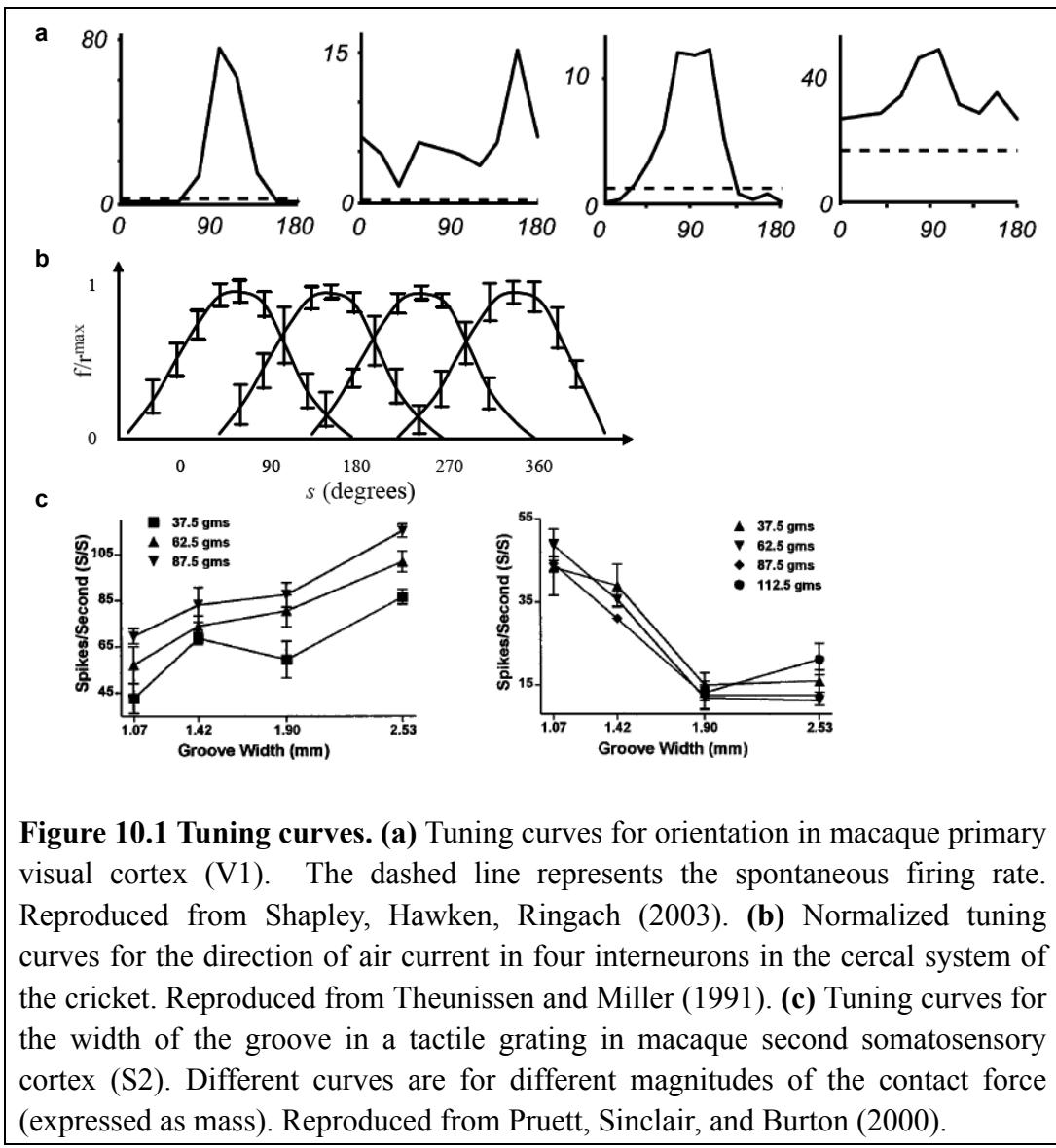
We will work towards asking the question “How could a given computation be implemented?” and not so much “Which specific neural circuits actually implements that computation?” That being said, for making testable physiological predictions, it is clearly important to focus on a particular species and brain area. However, even so, the localization of the computation is of secondary interest to the mechanisms of the computation. Our goal is to understand potentially ubiquitous neural processing mechanisms, rather than to model a specific circuit.

## 9.3 Tuning curves

The concept of tuning curves became popular with the pioneering experiments of Hubel and Wiesel in the mid-1960s. They recorded from primary visual cortex (V1) in cat while stimulating with illuminated oriented bars (see Fig 1a). They found that the response of a cortical neuron was systematically related to the orientation of the stimulus. There exists one orientation of the stimulus where the

neuron fires most rapidly: the neuron's preferred orientation. For other orientations, the activity decreases with increasing angle relative to the preferred orientation. A plot of the mean firing rate (e.g., spikes per second) as a function of angle describes the neuron's tuning curve. In the case of many visual neurons, this is a unimodal function (See Fig 1b).

*tuning curve*

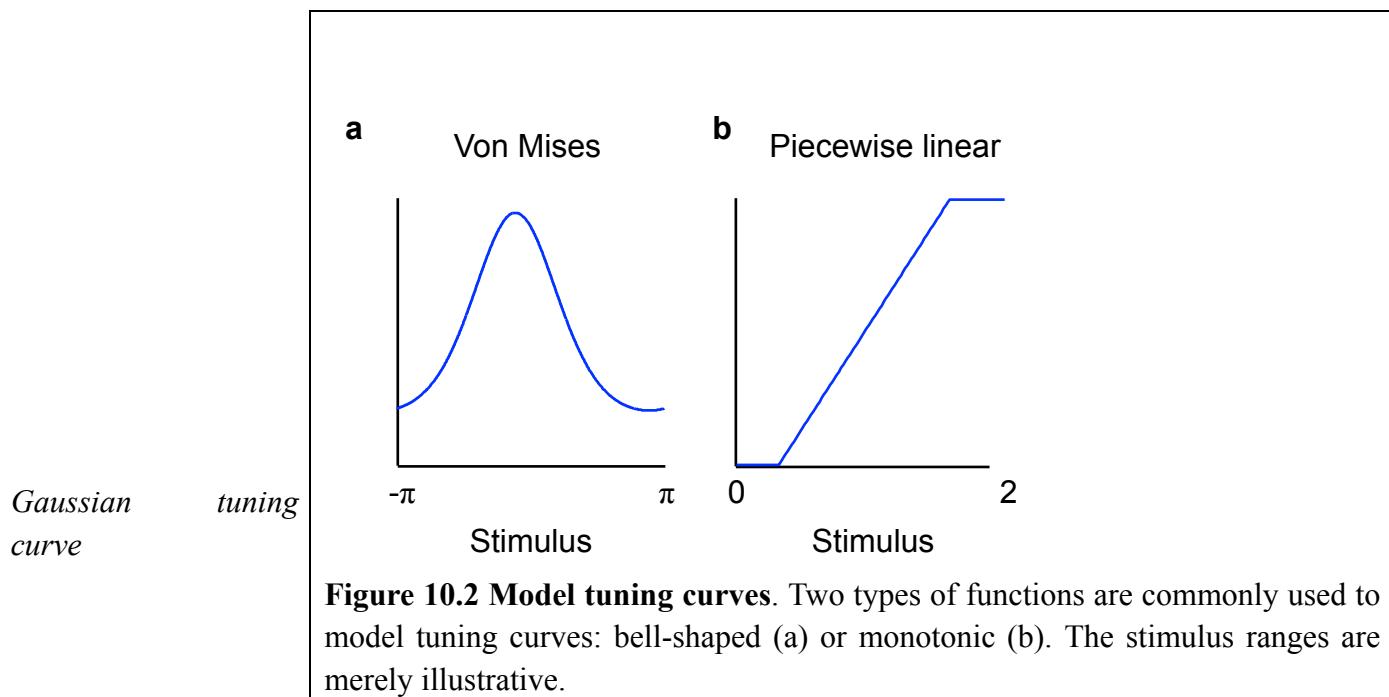


Tuning curves can have a wide variety of shapes, depending on the species, the brain area, and stimulus feature. For example, in motor cortex, we find that neural responses influence the direction of movement of the hand of a monkey. Instead of narrow unimodal functions we usually find very broad tuning curves. In auditory cortex, the frequency of the sounds stimulus affects the firing

rate of the neuron in a complex tuning curve. And in the hippocampus, a region of the mammalian brain involved in memory acquisition and navigation, there is a two dimensional representation of positions. In experiments with rats, firing rates of hippocampal neurons depend on both x and y position. The important thing in all these cases is that reasonably simple tuning curves characterize the mapping from sensory stimuli to the activity of neurons.

### 9.3.1 Bell-shaped tuning curves

When modeling tuning curves, scientists usually use simplified functions. When we model tuning curves like those found for neurons in primary visual cortex, we typically use *bell-shaped* tuning curves. Because rotating a bar by 180 degrees leads to the same response, we usually use circular Gaussian functions (Von Mises function, Fig 1A). By contrast, when scientists deal with auditory stimuli of varying amplitude, tuning curves typically show increasing activities. Piecewise linear functions can be used for such scenarios (red in Fig 1B). We will now consider some tuning curve functions in more detail.



**Key point:** even though the tuning curve might look bell-shaped or even be described by a Gaussian function, it is **NOT A PROBABILITY DISTRIBUTION!** It is not normalized and has nothing to do with probability distributions.

## Von Mises tuning curve

In primary visual cortex, neurons are tuned to the orientation of a visual stimulus, such as a bar. The tuning curve is typically unimodal and symmetric around the preferred orientation. Furthermore, the mean spike rate is the same at any angle and that angle plus or minus 180-degrees (as the bar stimulus is identical when rotated by 180 degrees). A common way to describe such a curve is to use a *Von Mises function* (also called a circular Gaussian) (Fig 10.2b):

$$f_i(s) = g e^{\kappa(\cos(s - s_i) - 1)} + b. \quad (10.1)$$

Here,  $\kappa$  is called the concentration parameter. The higher  $\kappa$ , the more narrowly the neuron is tuned. This function has been used to fit tuning curves over orientation, such as those in Fig 10.1a.

### 9.3.2 Monotonic tuning curves

A non-bell-shaped tuning curve occurs in some neurons. An example is a monotonic tuning curve such as those shown in Fig 10.1b. Several possibilities can be considered. The simplest form is a rectified linear function:

$$f_i(s) = [gs + b]_+, \quad (10.2)$$

where  $g$  is positive for monotonically increasing, and negative for monotonically decreasing tuning curves. Note that these neurons do not truly have a preferred stimulus (for the monotonically decreasing tuning curves, you could say it is 0, but that does not help much). A clear problem of a rectified linear function is that it is unbounded: as  $s$  increases,  $f(s)$  does not stay below any maximum value. This is unrealistic, since neurons have a limited dynamic range and cannot fire more than a certain number of spikes per second, no matter how large  $s$  becomes.

It may come as a surprise that the exact shape of the tuning curve is not critical to most of the theory we will discuss here. The theory will be general and work for tuning curves of any shape. However, the shape is of great practical relevance, since it is the starting point of any implementation of neural population activity. Moreover, we will occasionally use a specific functional form to allow for analytical calculations.

#### Key points

The tuning curve describes a neuron's mean activity as a function of the presented stimulus. Tuning curves are usually bell-shaped or monotonic. Various mathematical functions have been used to model them.

#### Box: More detailed models for the tuning curve

It is possible to describe the mapping from stimulus  $s$  to the output of a V1 neuron in much more detail. Instead of describing the world state as a scalar orientation, we can describe the entire two-dimensional image as a vector  $\mathbf{I}=(I_1, I_2, \dots, I_m)$ , where  $m$  is the number of pixels. Fig X shows examples of images of an oriented bar similar to the ones used by Hubel and Wiesel. As the orientation of the bar changes, the entire image changes, so we can consider  $\mathbf{I}$  a function of  $s$  and write it as  $\mathbf{I}(s)$ .

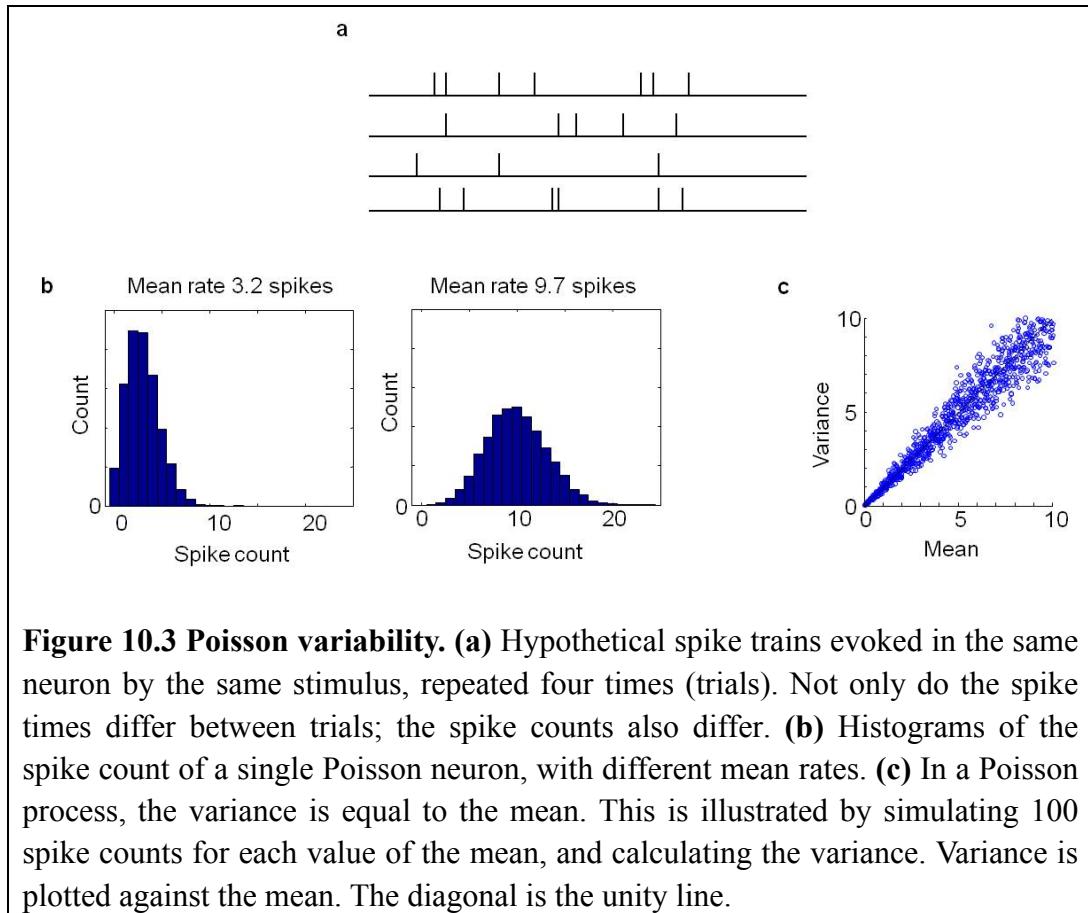
Each neuron has a spatial filter, which means that it will respond positively to light in certain locations in the image and negatively to light in other locations. In other words, the neuron associates a weight, positive or negative, with every pixel in the image. We call this filter or weight vector  $\mathbf{w}$ , and it can itself be visualized as an image. A typical V1 filter is shown in Fig. X [image of Gabor]; this is built so that if the image contains an orientation at the location of the filter, then the neuron will respond strongly. The neuron's average spike count in response to the image is then a sum of the pixel intensities multiplied by the corresponding weights:

$$f(s) = w_1 I_1(s) + w_2 I_2(s) + \dots + w_m I_m(s) = \mathbf{w} \cdot \mathbf{I}(s) . \quad (10.3)$$

As orientation is varied, this produces a tuning curve similar to the one in Fig. X. Spike counts would still be generated from a Poisson variability with mean  $f(s)$ , just like in the main text. The model of tuning curves in Eq. (10.3) is called a *linear model*, because  $f(s)$  is a linear combination of image intensities; it is not a linear function of orientation  $s$ . There are many ways to extend this model. For instance, one can postulate that  $f(s)$  is a *nonlinear* (but monotonically increasing) function of  $\mathbf{w} \cdot \mathbf{I}(s)$ . The resulting family of models is called LNP models, where L stands for “linear” (Eq. (10.3)), N for nonlinear, and P for Poisson.

## 9.4 Variability

So far, we have discussed a neuron's *selectivity*: which stimuli it spikes to or “likes”, as described by its tuning curve. However, if we view neuronal activity as resulting from a statistical generative model, we need to specify both their stimulus-dependent activity as well as their (also potentially stimulus-dependent) variability. For an identical, repeated stimulus, how much variation exists in the response from trial to trial? This variability will be critical for doing inference at the neural level. Here we will focus on Poisson variability.



#### 9.4.1 Poisson variability

Poisson variability (Fig 10.3a) is defined for a spike count, e.g. the number of spikes elicited by a flash of light that is presented for 10 ms. Spike count is a non-negative integer; it can be 0. Suppose a stimulus  $s$  is presented and the mean spike count of a neuron in response to this stimulus is  $\lambda = f(s)$ , which does not need to be an integer.  $\lambda$  is also called the *rate* of the Poisson process. Then the actual spike count will vary from trial to trial, around  $\lambda$ . For every possible count  $r$ , we seek its probability. A *Poisson process* (or in our context, a Poisson spike train) is defined as follows. Imagine a fixed time interval, and divide it into small bins (e.g. 1 millisecond each). We assume that each bin can contain 0 spikes or 1 spike, and that the occurrence of a spike is independent of whether and when spikes occurred earlier (it is sometimes said that a Poisson process “has no memory”). It can be proved in such a case (see Problems) that for a Poisson process with mean  $\lambda$ , the probability of observing a total of  $r$  spikes on a single trial is given by the *Poisson distribution*,

*Poisson process*

*Poisson distribution*

$$p(r_i | \lambda_i) = \frac{1}{r_i!} e^{-\lambda_i} \lambda_i^{r_i}. \quad (10.4)$$

Here,  $r!$  (read “ $r$  factorial”) is defined as  $1 \cdot 2 \cdot 3 \cdot \dots \cdot r$ .

### Numerical example

The rate of a Poisson neuron is  $\lambda_i=3.2$ . What is the probability that this neuron is silent? That it fires 1 spike? That it fires 10 spikes?

Solution: From Eq. (10.4), the probability that the neuron fires 0 spikes is  $1/0! \cdot \exp(-3.2) \cdot 3.2^0 = \exp(-3.2) = 0.04$ , or 4%. The probability that the neuron fires 1 spike is  $1/1! \cdot \exp(-3.2) \cdot 3.2^1 = \exp(-3.2) \cdot 3.2 = 0.04 = 0.13$ , or 13%. The probability that the neuron fires 10 spikes is  $1/10! \cdot \exp(-3.2) \cdot 3.2^{10} = 0.001$ , or 0.1%.

The Poisson distribution is shown for  $\lambda=3.2$  and  $\lambda=9.5$  in Figure 10.3b. Keep in mind that, while  $r$  is an integer,  $\lambda$  can be any positive number. For low  $\lambda$ , the distribution is less symmetrical than for high means. In fact, at high mean firing rates, the distribution looks roughly Gaussian. However, note that the Poisson distribution is discrete, so drawing it as a continuous curve would be a mistake.

An important property of the Poisson distribution is that the variance of a Poisson-distributed variable is equal to its mean: if the mean firing rate of a Poisson neuron is  $\lambda$ , then the variance of this neuron’s spike count is also  $\lambda$ . (Problem 10.2 and Fig. 10.3c). The ratio of the variance to mean of a neuron’s spike count is called the *Fano factor*; for a Poisson process, the Fano factor is 1.

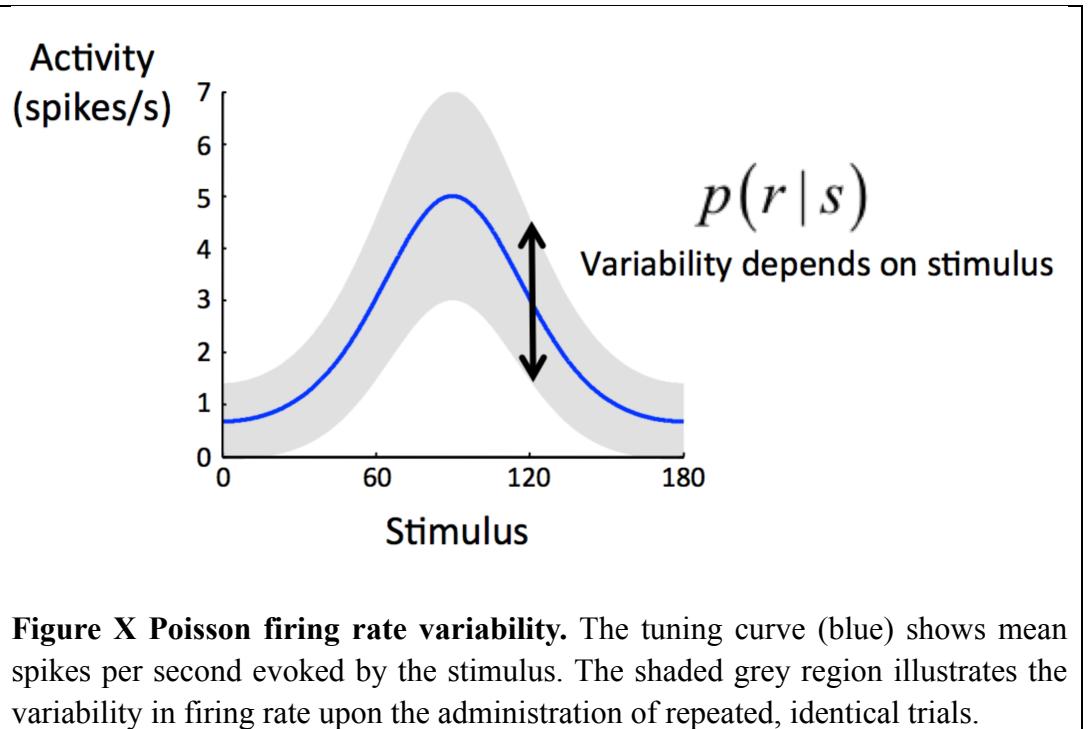
For our generative model of neural firing, we need to specify the probability of a firing rate,  $r$ , as a function of the stimulus,  $s$ . To do this, we note that  $\lambda$  is a function of the stimulus: it is the height of the tuning curve (the neuron’s average firing rate) at stimulus level  $s$ . Therefore, in terms of the stimulus, Eq. (10.4) can be written as

$$p(r_i | s) = \frac{1}{r_i!} e^{-f_i(s)} f_i(s)^{r_i}. \quad (10.5)$$

This is the form we will use frequently. Note that the neuron’s tuning curve,  $f_i(s)$ , plays a role in the neuron’s variability, but that it was not necessary to commit to any specific tuning curve shape to derive the form of the variability. In other words, Poisson variability can go together with any type of tuning curve, whether bell-shaped or monotonic. It is a common mistake to confuse the tuning curve with variability. This is understandable when one compares plots like those in Figures 10.2a and 10.3b, but the meaning of the axes in these plots is completely

*Fano factor*

different. The relationship between the tuning curve and the variability in firing is illustrated in Fig. X:



*refractory period*

It may at first appear that stimuli that evoke higher firing rates will be less informative, because higher firing rates are associated with more variability (e.g., spike rate variance = mean spike rate for a Poisson process). However, higher firing rates in fact convey more information. To see this, consider the mean as the signal and the standard deviation of the variability as the noise. Then the noise (square root of the variance) equals the square root of the mean. The signal-to-noise ratio therefore increases as the square root of the mean. We will later make this statement more precisely for the case of a population of neurons.

#### 9.4.2 More realistic models

Poisson variability is reasonably physiologically realistic, but with a number of caveats. Real Fano factors of cortical neurons are often close to 1, but can take values as low as 0.3 and as high as 1.8. Another unrealistic aspect of Poisson variability is that it assumes that spikes are independent of previous spikes. This is clearly not true: after a neuron fires a spike, it cannot fire again for a short duration, called the *refractory period* (typically several milliseconds). Thus, during that period, the probability of firing a spike is 0, contradicting the way we

defined the Poisson process. There exists a literature that extends the models we discuss here to more realistic models but it is beyond the scope of this book.

### 9.4.3 The origin of variability

The origin of neural variability is unknown and likely a combination of factors. Part of it has an external origin: when the same value of a visual stimulus variable is presented, this might not mean that the retinal image is identical. In many experiments, stimulus reliability is controlled by manipulating the amount of external noise. In those cases, the retinal image will be different even though the stimulus variable of interest is the same. Ideally, variability is measured under repeated presentations of the exact same physical image. This has been done in area MT in macaque, and response variability was still found. This variability can be attributed to internal factors. Internal sources of variability include neurotransmitter release and synaptic failure, both of which are stochastic processes.

#### Key points

A neuron's response to a particular stimulus varies from trial to trial. Such variability or noise can be described by a probability distribution: Poisson or Gaussian are common choices. In real neurons, variance is approximately proportional to the mean. The origin of variability is not yet understood.

## 9.5 Population codes

*population of neurons*

Neural *populations* are groups of neurons that are all selective for a particular stimulus feature. The neurons in a population are often but not necessarily located close to each other in the brain, and they often have similarly shaped tuning curves but with different preferred stimulus values; a population consisting of neurons with the identical tuning curve would not be particularly useful, since all neurons would “cover” the same restricted region of stimulus space.

*population code*

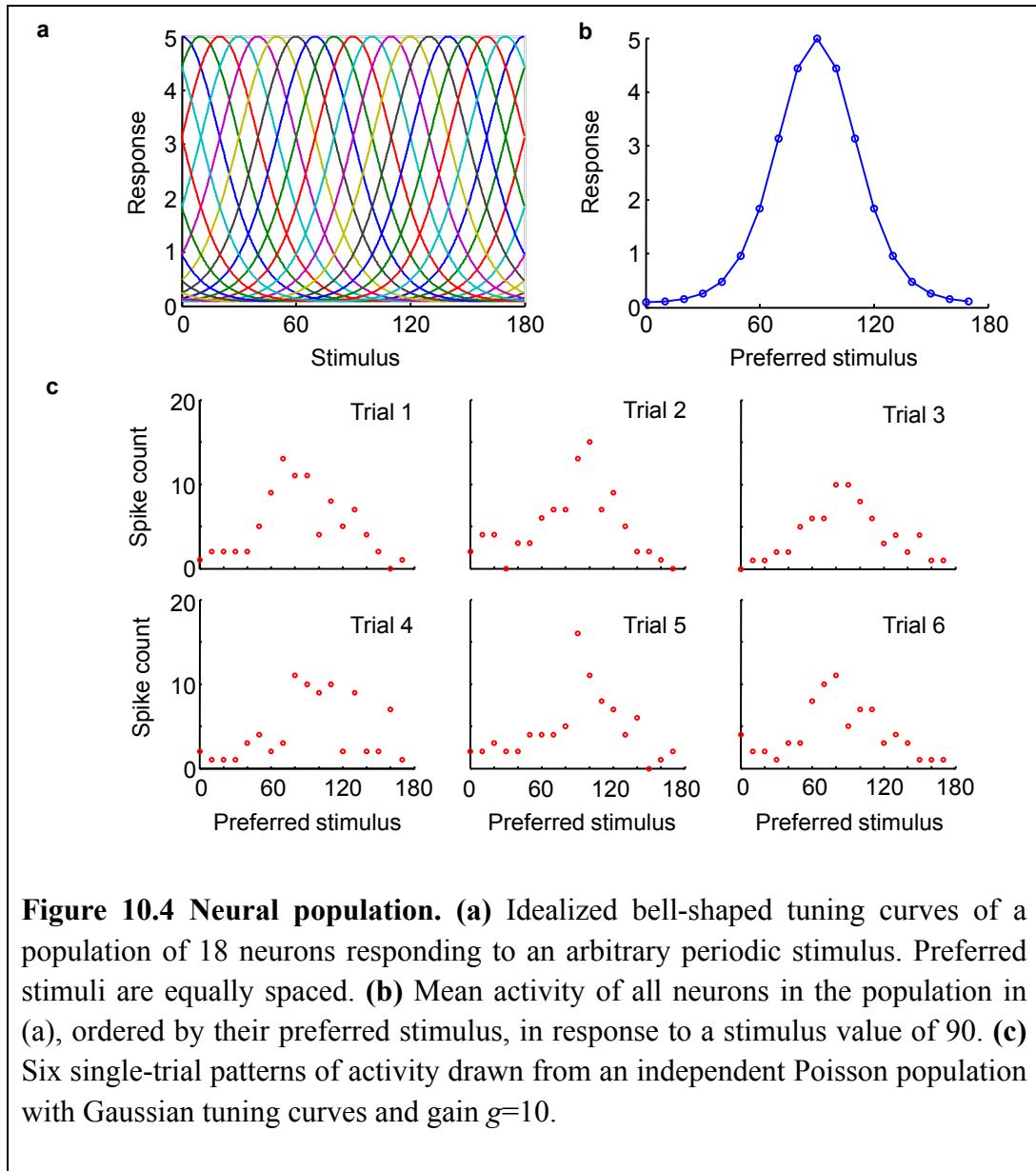
A *population code* refers to the stimulus-evoked firing rates of a population of neurons with different tuning curves. Population codes are believed to be widespread in the nervous system. For instance, in primary visual cortex (V1) and area V4 of the macaque, population codes exist for orientation, color, and spatial frequency. In the hippocampus in rats, a population code exists for the animal's body location. The cercal system of the cricket has a population code for wind direction. Secondary somatosensory area (S2) in the macaque has population codes for surface roughness, speed, and force. The post-subiculum in rat contains a population code for head direction. Primary motor cortex (M1) in macaque uses

populations coding for direction of reach. Even abstract concepts such as number appears to be encoded by population codes in the prefrontal cortex.

The firing rates of the set of neurons depicted in Fig 10.1a forms a population code; in fact, the cricket cercal system population consists of exactly the four neurons shown there. An idealized example with Gaussian tuning curves is drawn in Figure 10.4a. In this example, we show 10 neurons with preferred stimuli equally spaced on an interval. Such equal spacing is an idealization just as an exact Gaussian shape is. Yet, this is usually how a population code is simulated. We call the preferred stimuli  $s_1, \dots, s_N$ , where  $N$  is the number of neurons in the population. The tuning curves only differ in their preferred stimuli, so Eq. **Error! Reference source not found.** is valid;  $f_i(s)$  is the tuning curve of the neuron with preferred stimulus  $s_i$ . The set of tuning curves of all neurons in the population,  $\{f_1(s), \dots, f_N(s)\}$ , is denoted as a vector-valued function  $\mathbf{f}(s)$ .

Just as we modeled the variability of a single neuron's spike count, we can model the variability of the entire population. We denote by  $\mathbf{r}$  the vector of spike counts of the neurons in the population:  $\mathbf{r} = (r_1, \dots, r_N)$ . This is also called a *population pattern of activity*. The probability of observing a pattern of activity  $\mathbf{r}$  in response to a stimulus  $s$  is denoted  $p(\mathbf{r}|s)$ . The mean population pattern of activity over many trials is a smooth curve that resembles the tuning curve (Fig 10.4b). While the tuning curve shows the mean activity of one neuron in response to different stimuli, the population pattern shows the mean activity of every neuron in response to a single stimulus.

*population pattern of activity*



**Figure 10.4 Neural population.** (a) Idealized bell-shaped tuning curves of a population of 18 neurons responding to an arbitrary periodic stimulus. Preferred stimuli are equally spaced. (b) Mean activity of all neurons in the population in (a), ordered by their preferred stimulus, in response to a stimulus value of 90. (c) Six single-trial patterns of activity drawn from an independent Poisson population with Gaussian tuning curves and gain  $g=10$ .

The vector notation for population activity is simply for convenience; it does not have a deeper meaning. One could just as well write  $r_1, \dots, r_N$  wherever  $\mathbf{r}$  appears (and similarly for  $\mathbf{f}$ ), but this would make equations unnecessarily cluttered. Within the vector  $\mathbf{r}$  (or  $\mathbf{f}$ ), the ordering of the neurons has no meaning at all. We will typically order them by their preferred stimulus, only to make visualizations of population patterns like the one in Figure 10.4c look sensible.

In analogy to the single-neuron case, we now discuss Poisson and Gaussian variability in the population.

### 9.5.1 Independent Poisson variability in a population

The simplest assumption we can make about the population is that for a given stimulus, the responses of the neurons are drawn independently from each other, and that each response follows a Poisson distribution (but with its own mean). If random variables are independent from each other (in this case for a given stimulus), their joint probability distribution is the product of the individual distributions (again for a given stimulus). This means that we can write the population variability as

$$p(\mathbf{r} | s) = p(r_1 | s) \text{L} \ p(r_N | s) = \prod_{i=1}^N p(r_i | s). \quad (10.6)$$

The last equality is just a notation for the product. Now we substitute Eq. (10.5) for  $p(r_i | s)$ :

$$p(\mathbf{r} | s) = \prod_{i=1}^N \frac{1}{r_i!} e^{-f_i(s)} f_i(s)^{r_i}. \quad (10.7)$$

This is the probability distribution of independent Poisson variability. Figure 10.4c shows patterns of activity drawn from this distribution, if tuning curves are Gaussian. In a Problem, you will simulate such patterns yourself. Spike count is plotted as a function of the preferred stimulus of the neuron. Each dot corresponds to the activity of one neuron. We could have plotted them in any order, but visually it is most insightful to order the neurons by their preferred stimulus. Each pattern in Fig 10.4c is the population analog of one spike count in the histograms of Fig 10.3b. For the population, it is impossible to draw the histogram, since  $\mathbf{r}$  is now an  $N$ -dimensional vector, and we cannot draw histograms in  $N$  dimensions.

Nevertheless, we can still calculate the probability of each pattern of activity like those in Fig 10.4c.

### 9.5.2 Numerical example

We assume a population of 9 independent Poisson neurons with Gaussian tuning curves and preferred orientations from -40 to 40 in steps of 10. The tuning curve parameters have values  $g=10$ ,  $b=0$ , and  $\sigma_{tc}=20$ . A stimulus  $s=0$  is presented to this population. What is the probability of observing a pattern of activity  $\mathbf{r}=(3,1,7,5,8,8,7,0,2)$ ?

Solution: Under our assumptions about the tuning curves, the mean activity of the  $i$ 'th neuron is  $f_i(s=0)=10 \cdot \exp(-s_i^2/800)$ . Across the population, this gives mean activities (1.3, 3.2, 6.1, 8.8, 10, 8.8, 6.1, 3.2, 1.3). Then from Eq. (10.7):

$$p(\mathbf{r}|s=0) = e^{-1.3} \cdot 1.3^3 / 3! \cdot e^{-3.2} \cdot 3.2^1 / 1! \cdot \dots \cdot e^{-1.3} \cdot 1.3^2 / 2! = 2.4 \cdot 10^{-9}.$$

This number is striking because it is so small. How can it be that a pattern of activity that is not so different from the mean activities is so improbable? The reason is that this is one out of a huge number of possible patterns of activity. To get an idea of this number, let's do a rough estimation. Let's suppose that it is nearly impossible that any individual neuron will fire 20 or more spikes, given the mean rates. Then, each neuron's activity can take 20 values (including 0). There are 9 neurons and they are independent of each other, so the total number of patterns is  $9^{20} = 1.2 \cdot 10^{19}$ . If each of these patterns had been equally likely, each would have had a probability of  $1/(1.2 \cdot 10^{19}) = 8.3 \cdot 10^{-20}$ . Compared to this, the probability of the pattern we calculated above is actually very large! We conclude that it is expected that in an independent Poisson population, each pattern has a low probability, and the more neurons, the lower this probability. If the neurons were Poisson but not independent, fewer patterns would be possible and the probability of a given pattern would tend to be higher.

The patterns in Fig 10.3 make clear that an individual pattern of activity is roughly shaped like the Gaussian tuning curve, but with a different x-axis: preferred stimulus as opposed to stimulus. In fact, if one were to average over many patterns of activity elicited by the same stimulus  $s$ , one would get a mean activity described by the set of numbers  $f_i(s)$  for  $i=1,\dots,N$ . Looking back at Eq. **Error! Reference source not found.**, we see we can plot  $f_i(s)$  as a function of the preferred stimulus  $s_i$ , with  $s$  fixed. This is a Gaussian shape, just like  $f_i(s)$  was a Gaussian shape as a function of  $s$ , with  $s_i$  fixed. In other words, the mean population response to one repeated stimulus has the same shape as the mean response of a single neuron as a function of the stimulus. This is true for any tuning curve in which  $s_i$  and  $s$  can be swapped without affecting the functional form, such as Von Mises and cosine tuning curves (Eqs. (10.1) and Eq. **Error! Reference source not found.**).

### Key points

Neurons are often part of populations that respond to the same state-of-the-world variable. Population activity is described through the tuning curves of individual neurons as well as their joint variability.

## 9.6 Further reading

We refer readers interested in detailed neuron models to *Spiking neuron models* by Gerstner and Kistler. Historical videos recording of Hubel and Wiesel presenting stimuli to cat LGN and V1 neurons can be found online.

- Averbeck BB, Latham PE, Pouget A (2006) Neural correlations, population coding, and computation. *Nat Rev Neurosci* 7:358-366.
- Deneve S, Latham P, Pouget A (1999) Reading population codes: a neural implementation of ideal observers. *Nature Neuroscience* 2:740-745.
- Foldiak P (1993) The 'ideal homunculus': statistical inference from neural population responses. In: *Computation and Neural Systems* (Eckman F, Bower J, eds), pp 55-60. Norwell, MA: Kluwer Academic Publishers.
- Georgopoulos A, Kalaska J, Caminiti R, Massey JT (1982) On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *Journal of Neuroscience* 2:1527-1537.
- Gur M, Snodderly DM (2005) High Response Reliability of Neurons in Primary Visual Cortex (V1) of Alert, Trained Monkeys. *Cereb Cortex*.
- Hoyer PO, Hyvarinen A (2003) Interpreting neural response variability as Monte Carlo sampling of the posterior. In: *Neural information processing systems*: MIT Press.
- Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. *Nat Neurosci* 9:1432-1438.
- Pouget A, Dayan P, Zemel RS (2003) Inference and computation with population codes. *Annual Review of Neuroscience* 26:381-410.
- Sanger T (1996) Probability density estimation for the interpretation of neural population codes. *Journal of Neurophysiology* 76:2790-2793.
- Tolhurst D, Movshon J, Dean A (1982) The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Research* 23:775-785.
- Zhang K, Ginzburg I, McNaughton B, Sejnowski T (1998) Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells. *Journal of Neurophysiology* 79:1017-1044.
- Averbeck BB, Latham PE, Pouget A (2006) Neural correlations, population coding, and computation. *Nat Rev Neurosci* 7:358-366.
- Gur M, Snodderly DM (2005) High Response Reliability of Neurons in Primary Visual Cortex (V1) of Alert, Trained Monkeys. *Cereb Cortex*.
- Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. *Nat Neurosci* 9:1432-1438.
- Tolhurst D, Movshon J, Dean A (1982) The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Research* 23:775-785.

## 9.7 Problems

**Problem 10.1.** Why are monotonic tuning curves always over magnitude-type variables such as width, not over circular variables such as orientation?

**Problem 10.2.** Are the following statements true or false? Explain.

- a) The closer a stimulus is to the preferred stimulus of a Poisson neuron, the lower is the response variance of this neuron when the stimulus is presented repeatedly.
- b) When neurons have similar and equally spaced tuning curves, then the mean population pattern of activity in response to a stimulus has the same width as the tuning curve.
- c) When neurons have similar and equally spaced tuning curves, then the neural posterior has the same width as the tuning curve.
- d) The variance of a single neuron responding to a stimulus can be determined from the value of its tuning curve at that stimulus value.
- e) In any population, the variability of population activity is known if one knows the variability of each single neuron.

**Problem 10.3.** Read the image filter approach in Box 10.2. We assume a Gabor stimulus of orientation  $s$ , spatial wavelength  $\lambda$ , and standard deviation of the Gaussian envelope 1. [give equation] We describe the neuron as a Gabor filter with the same wavelength, the same envelope, and orientation  $s_{\text{pref}}$ .

- a) Derive the orientation tuning curve of this neuron.
- b) How does the tuning depend on  $\lambda$ ?

**Problem 10.3.** We assume a population of 9 independent Poisson neurons with Gaussian tuning curves and preferred orientations from -40 to 40 in steps of 10. The tuning curve parameters have values  $g=10$ ,  $b=0$ , and  $\sigma_{\text{tc}}=20$ . A stimulus  $s=0$  is presented to this population. What is the probability that all neurons stay silent?

**Problem 10.4. Properties of the Poisson distribution.**

- a) Prove that Eq. (10.4) implies that the mean value of  $r_i$  is indeed  $f_i(s)$ . Recall

$$\langle r_i \rangle = \sum_{r_i=0}^{\infty} r_i p(r_i | \lambda_i)$$

that the mean of  $r_i$  is defined as  $\sum_{r_i=0}^{\infty} r_i p(r_i | \lambda_i)$ . You will also need to use a variant of Eq. **Error! Reference source not found.**

- b) Prove that Eq. (10.4) implies that the variance of a Poisson process is equal to its mean. Recall that the variance of  $r_i$  can be written as

$$\text{Var}(r_i) = \langle r_i^2 \rangle - \langle r_i \rangle^2 = \sum_{r_i=0}^{\infty} r_i^2 p(r_i | \lambda_i) - \langle r_i \rangle^2$$

**Problem 10.5.** In a population of independent Poisson neurons with Gaussian tuning curves, examine the claim that  $\sum_i f_i(s)$  is more or less independent of  $s$ .

**Problem 10.6.** Prove that independent Poisson variability, Eq. (10.7), is a special case of Poisson-like variability:

- a) Rewrite Eq. (10.7) in the form of Eq. **Error! Reference source not found..**
- b) Verify that Eq. **Error! Reference source not found.** holds for  $\mathbf{h}(s)$  found in part (a).

**Problem 10.7.\*** In a sequence of  $R$  independent events with two possible outcomes, the probability of having one of both outcomes appear  $r$  times is described by the binomial distribution,

$$p(r) = \binom{R}{r} \left(\frac{\lambda}{R}\right)^r \left(1 - \frac{\lambda}{R}\right)^{R-r}.$$

Here,  $\binom{R}{r}$  is the binomial coefficient,  $\frac{R!}{r!(R-r)!}$ . Prove that the Poisson distribution is a good approximation of the binomial distribution if the sequence is long ( $R$  large) and the probability of the outcome of interest ( $\lambda/R$ ) is small.

**Problem 10.8.\*** Prove that for large means, a Poisson distribution resembles a Gaussian distribution with variance equal to the mean.

**Problem 10.9.\*** Show that in the limit of large  $\kappa$ , the Von Mises function (Eq. (10.1)) becomes the Gaussian function (Eq. **Error! Reference source not found.**). Hint: in this limit, the Von Mises function becomes very strongly peaked around  $s_i$ , and we can use the Taylor series expansion  $\cos(x) \approx 1 - x^2/2$ .

**Problem 10.10\*** Prove that correlated Gaussian variability, Eq. **Error! Reference source not found.**, reduces to Eq. **Error! Reference source not found.** when  $\langle r_i r_j \rangle = \langle r_i \rangle \langle r_j \rangle$  for  $i \neq j$  (i.e. neurons are uncorrelated).

**Problem 10.11.** Consider a population of neurons with known tuning curves, subject to independent noise.

- a) If the noise is drawn from a normal distribution with fixed variance, prove that the maximum-likelihood decoder is equivalent to the template-matching decoder.

- b) If the noise follows a Poisson distribution, tuning curves are Gaussian with zero baseline, and  $\sum_i f_i(s)$  is independent of  $s$ , to which decoder is the maximum-likelihood decoder equivalent? Prove your answer.

**Problem 10.11A.** Show that when neural variability is independent and Poisson, tuning curves are von Mises with zero baseline and  $\sum_i f_i(s)$  is independent of  $s$ , the maximum-likelihood decoder is equivalent to the population vector.

**Problem 10.12.** A Bayesian observer decodes a stimulus  $s$  from a neural population under a cost function,  $C(\hat{s}, s)$ .

- a) Prove that if the cost function is the squared error, the Bayesian estimate is the mean of the posterior distribution.
- b) Derive the Bayesian estimate if the cost function is the absolute error,  $C(\hat{s}, s) = |\hat{s} - s|$ .
- c) What is the cost function corresponding to the maximum-a-posteriori decoder?

**Problem 10.13.** In a discrimination task, an observer decides on each trial whether a stimulus has value  $s_1$  or  $s_2$ . The stimulus elicits activity  $\mathbf{r}$  in a neural population with tuning curves  $f_i(s)$ . Assume that  $\mathbf{r}$  is drawn from an independent Poisson distribution and that  $\sum_i f_i(s)$  is independent of  $s$ .

- a) Calculate the log likelihood ratio and prove that the maximum-likelihood decision is based on the sign (positive or negative) of the inner product of  $\mathbf{r}$  with a vector  $\mathbf{w}$ . Find an expression for the  $i^{\text{th}}$  component  $w_i$  in terms of the numbers  $f_i(s_1)$  and  $f_i(s_2)$ .
- b) What does the absolute value of  $\mathbf{w} \cdot \mathbf{r}$  mean to the observer? Explain.
- c) Compute the mean and variance of  $\mathbf{w} \cdot \mathbf{r}$  from part (a) when  $\mathbf{r}$  is generated by  $s_1$ , and when  $\mathbf{r}$  is generated by  $s_2$ . *Sensitivity* or discriminability is defined as the difference between both means divided by the square root of the mean of both variances (see also Chapter 5). Find an expression for discriminability in terms of the sets of numbers  $f_i(s_1)$  and  $f_i(s_2)$  (for all  $i$ ).

## LAB PROBLEMS

**Problem 10.24 Simulating a Poisson process.**

- a) Define 1000 time points. At each time point, determine whether a spike is fired by generating a random number that leads to a “yes” probability of

- 0.0032 (this corresponds to a mean of 3.2 spikes over all 1000 time points). Count the total number of spikes generated. Repeat for 10,000 trials. Plot a histogram of spike count and compare to Figure 10.3a. Compute the Fano factor.
- Repeat for a mean of 9.5 spikes. Compare the resulting histograms to Figure 10.3b.
  - If your simulation software has a built-in command to randomly generate numbers according to a Poisson distribution (e.g. poissrnd in Matlab), repeat (a) and (b) using this command.
  - A property of a Poisson process is that the time between two subsequent spikes (interspike interval, denoted here  $\Delta t$ ) follows an exponential distribution:  $p(\Delta t) = \exp(-\Delta t/\lambda)/\lambda$ , where  $\lambda$  is the mean of the Poisson process. Verify this by plotting the histogram of interspike intervals across all Poisson spike trains you generated in (a), and comparing with the exponential distribution.

### Problem 10.26

datasetReal (downloadable at [xxx](#)) contains recordings from 35 neurons in the primary motor cortex and about 200 trials, roughly half of which were recorded while the monkey was moving left while the other was recorded while the monkey was moving right (courtesy Miller lab). Lets assume a Gaussian distribution of spike counts, given the direction of movement.

- Calculate for each neuron the average firing rates for left and right movement and also the associated standard deviations.
- Do all neurons have similar average firing rates for left and right movements? Which of the neurons exhibit significant difference between left and right movement?
- What would be a good measure for strength of tuning of a neuron. Which neuron has the strongest tuning to direction?
- If you decoded movement direction based on just this neuron, how good would you do on average?
- If you combine data from all neurons, using a naïve Bayes approach how good can you be at the problem?
- Is this a difficult problem? Could it have real-world relevance? Can you think of an application of naïve Bayes decoding that is more exciting?

$$p(\text{right} | \text{spikes}) \propto \frac{1}{Z} p(\text{right}) \prod \frac{1}{\sigma_{\text{right},i}} e^{(\text{spikes}_i - \mu_{\text{right},i})^2 / (2\sigma_{\text{right},i}^2)}$$

This same approach of assuming that all cues are independent, even when they are not, is used in many domains of machine learning. Naïve Bayes is often used to solve real classification problems and is, for certain problems a competitive machine learning technique. It is particularly strong when there is very little available data.

## Table of Contents

10	The neural representation of uncertainty .....	2
10.1	Chapter structure .....	2
10.2	Neural likelihood function for a single neuron .....	3
10.2.1	Case 1: A bell-shaped tuning curve .....	4
10.2.2	Case 2: Monotonic tuning curve .....	7
10.3	The neural likelihood function based on a population of neurons .....	8
10.3.1	Log likelihood functions .....	9
10.3.2	Case 1: Bell-shaped tuning curves .....	10
10.3.3	Properties .....	12
10.3.4	Case 2: Monotonic tuning curves .....	14
10.3.5	Concluding remarks .....	14
10.4	Toy models .....	15
10.4.1	Case 1: Bell-shaped tuning curves .....	15
10.4.2	Case 2: Monotonic tuning curves .....	21
10.5	Relation between behavioral and neural concepts .....	21
10.6	Statistics of likelihood mode and width across many trials .....	24
10.6.1	Toy model .....	25
10.6.2	Distinction between $p(r s)$ and $p(r I)$ .....	27
10.7	Applications .....	27
10.7.1	fMRI .....	27
10.7.2	Brain machine interfaces .....	28
10.8	Problems .....	29

## 10 The neural representation of uncertainty

*How might neurons represent probability to support probabilistic inference?*

The brain performs inference based on neural activity. In the earlier chapters, we abstracted this process by introducing the measurement, denoted  $x$ , defining the likelihood  $L(s;x)$ , and computing the posteriors in a variety of tasks (Ch 2-8) based on such likelihood functions. To understand the neural basis of inference, we introduced in the previous chapter the formalism of populations of neurons responding to sensory stimuli. In the present chapter, we will use this formalism to define the likelihood function over a stimulus represented by the spike counts in a neural population observed on a particular trial. In the next chapter, we will then use this likelihood function as a building block for inference based on spike counts in neural populations.

Any neural representation of variables in the world implies a likelihood function about variables in the world. More specifically, in the previous chapter, we expressed how stimuli,  $s$ , in the world give rise to neural responses,  $r$ , in the form of a conditional probability distribution  $p(r|s)$ . The corresponding likelihood function is the same expression but viewed from the perspective of the observer, who has access to  $r$  but not to  $s$ :  $L(s;r)=p(r|s)$ . We will call this likelihood function the *neural likelihood function* to distinguish it from the likelihood function  $L(s;x)$  that we encountered in earlier chapters. All information that the neural activity  $r$  provides the observer about the stimulus  $s$  is contained in the neural likelihood function over the stimulus.  $L(s;r)$  is the answer to the question “If the stimulus had value  $s$ , what would be the probability that it would have produced the observed activity  $r$ ?”

Like any likelihood function, each likelihood function we will encounter here allows us to define an associated uncertainty. Following Section 3.X, we will define uncertainty as the width (standard deviation) of the likelihood function. Here, this uncertainty will reflect the uncertainty about the stimulus that remains after having observed  $r$ .

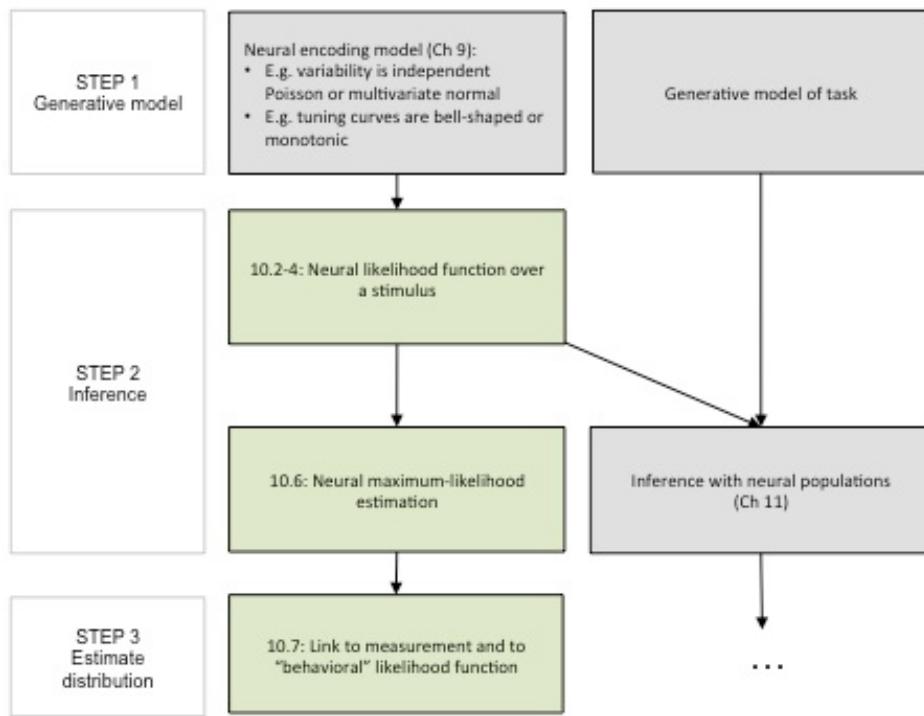
### 10.1 Chapter structure

To understand the neural likelihood function, this chapter will discuss several examples in increasing order of difficulty. In Sections 2 and 3, we will assume that neurons are independent and that the spike count of each neuron follows a Poisson distribution; this was the main framework in the previous chapter. In both sections, we will consider two types of tuning curves: bell-shaped and monotonic. Section 2 discusses the case of a single neuron, section 3 that of a population of neurons. In Section 4, we will compute population likelihood functions in cases where they have a compact functional form. In Section 5, which is advanced, we will discuss a generalization to non-independent (correlated) neurons; this will be facilitated by describing neural activity as a continuous variable (firing rate) that follows a multivariate Gaussian

distribution. In Section 6, we will link the neural likelihood function to the behavioral likelihood function from Chapters 2 to X. Finally, in Section 7, we will describe an application to brain-machine interfaces.

Throughout this chapter, keep in mind that “the stimulus” is a simple stimulus feature that is encoded at an early stage of sensory processing, such as the contrast of a blob of light, the orientation of a line segment, the location of a sound, or the width of a groove. It does *not* include states of the world that are more complicated or task-dependent, such as “the curvature of a contour”, “the presence of a search target”, “the category to which a tone belongs”, or “the amount of scatter in a cloud of dots”. Correspondingly,  $\mathbf{r}$  is the neural activity in that early stage of sensory processing – e.g. retina, LGN, V1, A1, or S1.

### Internal and external connections of Chapter 10



## 10.2 Neural likelihood function for a single neuron

We start with the likelihood function based on the spike count of a single neuron. We described the spike count of a neuron in a given time interval using a Poisson distribution. To recapitulate, if a neuron has a tuning curve  $f(s)$ , then the probability that this neuron will fire  $r$  spikes is

$$p(r | s) = \frac{1}{r!} e^{-f(s)} f(s)^r \quad (10.1)$$

This probability distribution, sometimes called the *distribution of neural variability* or *neural noise distribution*, serves as a generative model: it tells us the statistics of the observation  $r$  given a world state  $s$ .

Suppose now that  $r$  spikes are observed, where  $r$  is a specific number such as 0, 2, or 11. Given  $r$ , the neural likelihood of a hypothesized stimulus value  $s$  is the probability that  $r$  spikes were produced by that value of  $s$ . In other words, we copy Eq. (10.1) but consider it as a function of  $s$  rather than  $r$ .

$$L(s;r) = \frac{1}{r!} e^{-f(s)} f(s)^r. \quad (10.2)$$

We consider two example cases which we will follow throughout this section:

### 10.2.1 Case 1: A bell-shaped tuning curve

Suppose that the tuning curve of the neuron has a Gaussian shape with mean  $s_{\text{pref}} = 0$ , width  $\sigma_{\text{tc}} = 10$ , baseline  $b = 1$  and gain  $g = 5$ :

$$f(s) = ge^{-\frac{(s-s_{\text{pref}})^2}{2\sigma_{\text{tc}}^2}} + b \quad (10.3)$$

In a picture

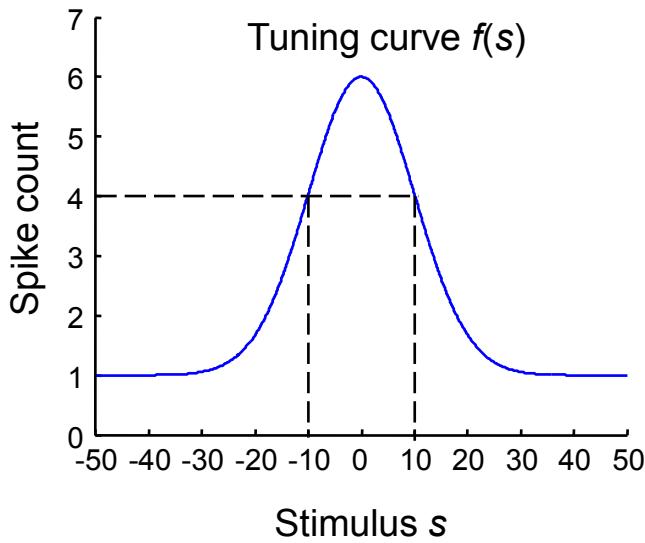


Figure 10.1: Tuning curve of a single neuron with preferred stimulus 0

A point on the blue curve is the mean spike count of the neuron in response to a particular stimulus. Eq. (10.1) tells us the distribution of spike count for any given  $s$ .

## Intuition behind likelihood

Suppose we are faced with the inverse problem: we are told this neuron fired 4 spikes in a given time interval, and asked what we can say about the stimulus. Based on Fig. 10.1, we might say the stimulus was approximately -10 or +10, because then the neuron would produce the expected number of spikes. However, Fig. 10.1 only shows us the average spike count over many trials. The trial-to-trial response is noisy, as expressed by Eq. (10.1). Therefore, a total of 4 spikes could also have been produced by a stimulus value of say 3.7 – it just so happened that on this trial, the neuron fired fewer spikes than average. 4 spikes could even indicate that the stimulus was -21, although it would require that the neuron happened to fire many more spike than its average spike count at this stimulus. Clearly, some stimulus values are more likely than others, and we can define the likelihood of a hypothesized stimulus value as the probability of observing 4 spikes in response to that stimulus value.

### Equation for the likelihood function.

We can formalize this intuition by simply substituting Eq. (10.3) into Eq. (10.2). That gives the following equation for the neural likelihood function:

$$L(s; r) = \frac{1}{r!} e^{-ge} \frac{(s-s_{\text{pref}})^2}{2\sigma_{\text{tc}}^2} - b \left( ge \frac{(s-s_{\text{pref}})^2}{2\sigma_{\text{tc}}^2} + b \right)^r$$

We have plotted this for  $r=4$  in Fig. X

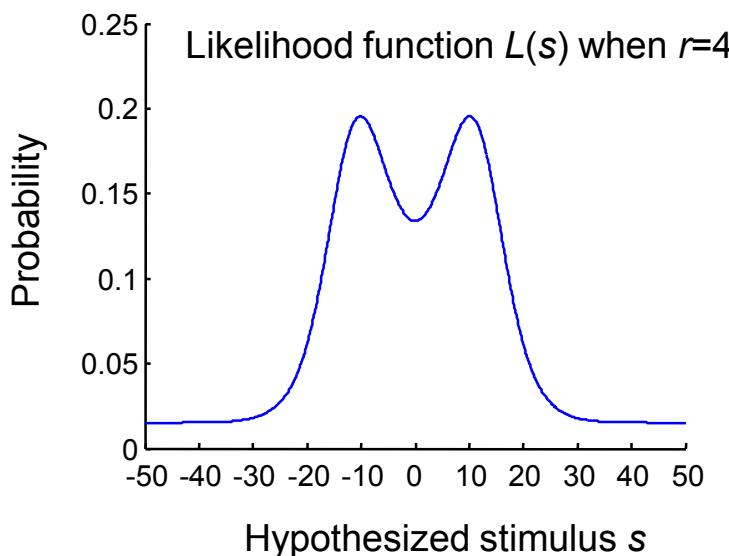


Figure 10.2: Likelihood function over the stimulus when 4 spikes are observed in a one-neuron brain

This odd-looking function tells us how likely each possible stimulus value is based on the observation (the spike count of 4). The shape confirms our intuition: values of (approximately) +10 and -10 are most likely, 0 is still quite likely, but -30 is very unlikely. To compute the likelihood function, we not only used the tuning curve (Eq. 10.2) but also the form of neural variability (Eq. 10.1). This allows us to say more about the stimulus than only that +10 and -10 are most likely.

## Properties

This likelihood function has several interesting properties. First, unlike the likelihood functions we encountered in earlier chapters (Eg. Ch 2), this likelihood is not Gaussian. In fact, there is no tuning curve  $f(s)$  we could have substituted in Eq. (10.2) to get a likelihood function that is Gaussian for arbitrary  $r$ .

A second important point is that the likelihood function in Fig. 10.2 is not normalized. In fact, the area under the curve is infinite! The “tails” extend to arbitrarily large values. This is simply because (see Fig 11.1) the tuning curve has a baseline of 1 spike, so the probability that the observed spike count was 4 is nearly the same value for  $s=30$  as, say, for  $s=1000$ . In the Appendix, Section X, you can find another example of a likelihood function that is not normalized (farmers and states example). In general, likelihood functions are not normalized. It just happened that in our behavioral models, the likelihood function over the stimulus was usually a normalized Gaussian (although we also encountered non-normalized likelihoods, in Chapter 5 when dealing with binary variables). In the present chapter, as we are discussing neural models, the likelihood function over the stimulus will never be normalized.

Fig. 10.2 shows the likelihood function over the stimulus when the observed spike count was 4. We can also calculate the likelihood function over the stimulus for other observed spike counts. This is done in Fig. 10.3.

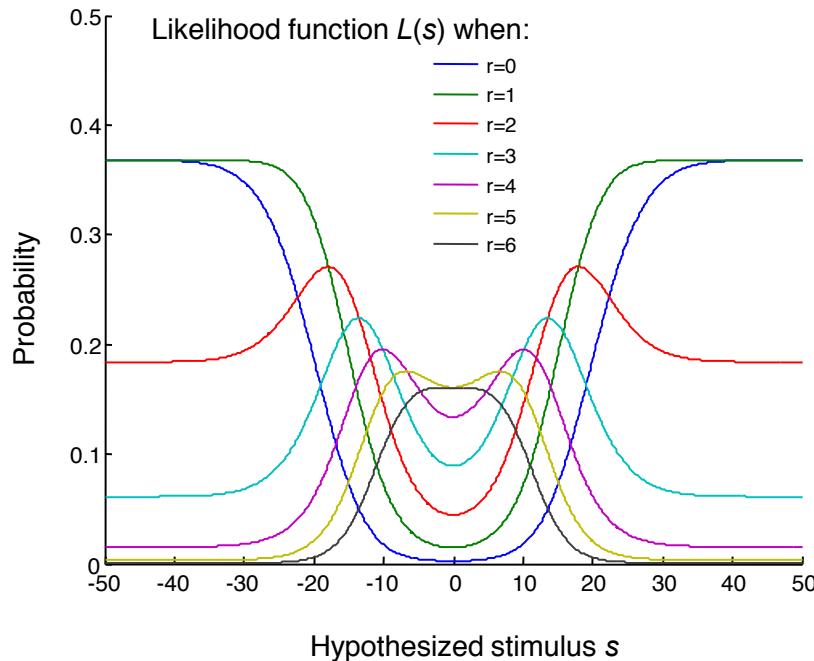


Figure 10.3: Likelihood functions over the stimulus for different observed spike counts in a one-neuron brain

This shows that the likelihood function can have a completely different shape for different observations. For example, when 0 spikes are observed, any of the central values of the stimuli are very unlikely, so the likelihood function has an inverted U-shape.

### 10.2.2 Case 2: Monotonic tuning curve

We have so far considered a real-valued variable with bell-shaped tuning. We will now consider a non-negative variable with monotonic tuning, as we encountered in Chapter 9.X. An example would be a power law tuning curve with baseline:

$$f(s) = as^b + c , \quad (10.4)$$

where we enforce  $a > 0$  and  $c \geq 0$ , so that mean spike counts are guaranteed non-negative regardless of  $s$ . Such a tuning curve would make sense for magnitude variables such as length, weight, contrast, and loudness (see Section 3.X). This tuning curve is shown for  $a=0.5$ ,  $b=1$ ,  $c=1$ , and various values of  $r$  in Fig. X.

[INSERT HERE FIGURE OF TUNING CURVE.]

### Intuition

What do we expect the likelihood to look like? Suppose we observe that this neuron fires 4 spikes. The likelihood reflects how probable this observation is under different hypothesized values of  $s$ . From the tuning curve, we know that this neuron fires 4 spikes on average when the stimulus is  $s=6$ ; therefore, we expect the probability that it will fire exactly 4 spikes to be quite high for a hypothesized stimulus value of 6. On the other hand, if the stimulus were 0, then the neuron's average number of spikes would be 0.5, making it improbable for the neuron to fire 4 spikes. Similarly, if the stimulus were 100, then the neuron's average number of spikes would be 51, which would again make our observation of 4 spikes improbable. Therefore, we would expect the likelihood to be high for  $s=6$  and drop off gradually as  $s$  moves away from 6 in either direction. Thus, we expect that the neural likelihood will be bell-shaped. This intuition is not specific to a spike count of 4. In general, any particular spike count will give the highest likelihood to one stimulus value and lower likelihoods to stimuli on either side.

### Equation

Substituting Eq. (10.4) into Eq. (10.2), the neural likelihood function is

$$p(r|s) = \frac{1}{r!} e^{-as^b-c} (as^b + c)^r$$

This is shown for  $a=0.5$ ,  $b=1$ ,  $c=1$ , and various values of  $r$  in Fig. X. As our intuition predicted, this is a bell-shaped tuning curve!

### 10.3 The neural likelihood function based on a population of neurons

The single neuron likelihood function (Fig. 10.2) is not only very wide, it also has two peaks: +10 and -10 are equally likely. This is not very satisfactory if we are interested in localizing the stimulus. However, when we recall that this likelihood function was based on the firing of just a single neuron, and that this neuron was noisy (Poisson), it is in fact remarkable how much we can already say about the stimulus. Fortunately, most of us have more than one neuron in our brain, and therefore the information that we have about stimuli in the world is based on the simultaneous firing of a population of neurons.

### Independence assumption

We will now consider a population consisting of an arbitrary number of neurons; we call the number of neurons  $n$ . On a given trial, the neurons in this population will produce a set of spike counts,  $(r_1, \dots, r_n)$ , which we will often denote shorthand by a vector  $\mathbf{r}$  and call the “pattern of population activity”. Mathematically,  $\mathbf{r}$  is a high-dimensional vector. If 1000 neurons were selective to  $s$ , then  $\mathbf{r}$  would be a 1000-dimensional vector. We assume that the variability in  $\mathbf{r}$  across trials is independent across neurons conditioned on  $s$ , as described by Eq. X.X.

$$p(\mathbf{r} | s) = p(r_1, \dots, r_n | s) = \prod_{i=1}^n p(r_i | s). \quad (10.5)$$

As a consequence, when we observe a specific pattern of population activity  $\mathbf{r}$ , the neural likelihood function over  $s$  is

$$L(s; \mathbf{r}) = \prod_{i=1}^n p(r_i | s).$$

Each factor in this product can be thought of as the likelihood function based on a single neuron's spike count. Thus, the population likelihood function is the product of single-neuron likelihood functions.

### Poisson assumption

To make further progress, we assume that every neuron's spike count follows a Poisson distribution, but each neuron has its own mean – for the  $i^{\text{th}}$  neuron, the mean is given by the  $i^{\text{th}}$  tuning curve evaluated at  $s$ . These tuning curves are described by functions  $f_i(s)$ , with  $i$  ranging from 1 to  $n$ . Then,

$$p(r_i | s) = \frac{1}{r_i!} e^{-f_i(s)} f_i(s)^{r_i}.$$

As a consequence, when we observe a specific pattern of population activity  $\mathbf{r}$ , the neural likelihood function over  $s$  is

$$L(s; \mathbf{r}) = \prod_{i=1}^n \left( \frac{1}{r_i!} e^{-f_i(s)} f_i(s)^{r_i} \right). \quad (10.6)$$

Like in the previous section, we will evaluate this likelihood for two cases: bell-shaped and monotonic tuning curves.

#### 10.3.1 Log likelihood functions

Before we do this, we will show what happens when one takes the logarithm of the likelihood function – for which we will use the shorthand terminology of “log likelihood function”. With the independence assumption only, the population log likelihood becomes the sum of the log likelihood functions of the individual neurons:

$$\log L(s; \mathbf{r}) = \sum_{i=1}^n \log p(r_i | s)$$

With both the independence assumption and the Poisson assumption, we have:

$$\begin{aligned} \log L(s; \mathbf{r}) &= \sum_{i=1}^n \log \left( \frac{1}{r_i!} e^{-f_i(s)} f_i(s)^{r_i} \right) \\ &= -\sum_{i=1}^n \log r_i! - \sum_{i=1}^n f_i(s) - \sum_{i=1}^n r_i \log f_i(s) \end{aligned}$$

While completely equivalent to Eq. (10.6), the logarithmic form is often considered easier to work with, because it has sums instead of products. At any time, one can recover the likelihood function from the log likelihood function by exponentiating it.

### 10.3.2 Case 1: Bell-shaped tuning curves

We consider the neural population from Section 9.X, in which the neurons have the following tuning curves:

$$f_i(s) = g_i e^{-\frac{(s-s_{\text{pref},i})^2}{2\sigma_{\text{tc},i}^2}} + b_i, \quad (10.7)$$

where  $g_i$ ,  $\sigma_{\text{tc},i}$ , and  $b_i$  are the gain, width, and baseline of the tuning curve of the  $i^{\text{th}}$  neuron. These look like this:

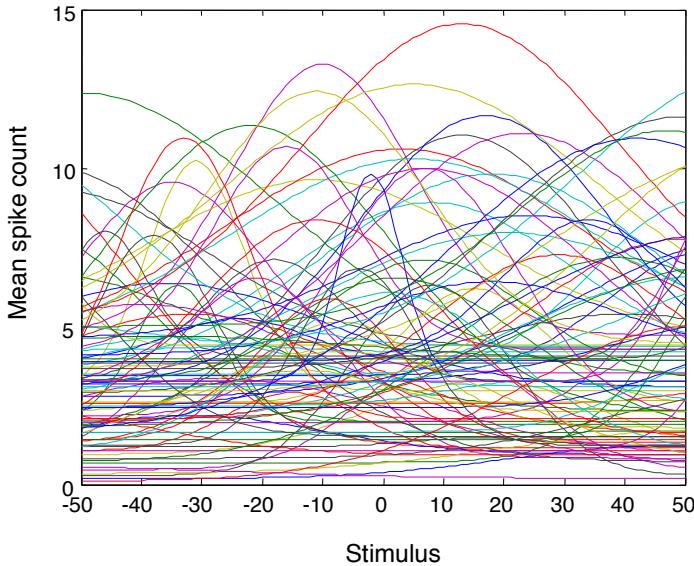


Figure 10.9: Tuning curves in a less regular (heterogeneous) population consisting of 100 neurons.

Amplitude, width, and baseline are highly variable, as is the case in real recordings.

We simulated three patterns of activity in the population of 100 independent Poisson neurons with tuning curves as in Fig. 10.9 (preferred stimuli equally spaced between -60 and 60), elicited by the stimulus  $s=0$ . These patterns would correspond to neural recordings on three trials on which  $s=0$  was presented. Such a pattern could look like Fig. 10.7.

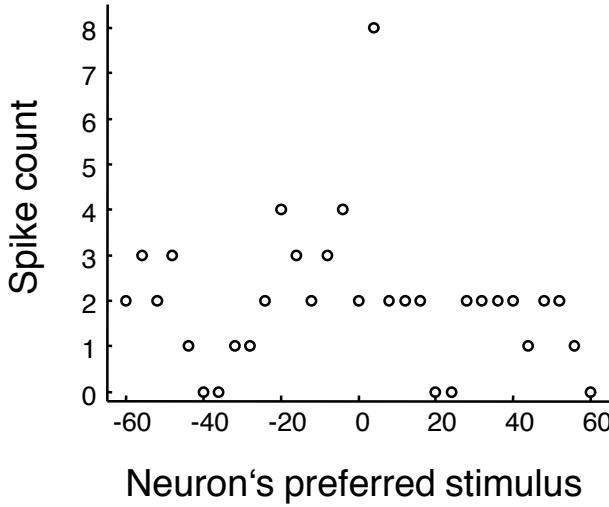


Figure 10.7: Example population pattern of activity.

The likelihood over  $s$  is obtained by substituting these specific values of  $r_1, \dots, r_n$  into  $p(r_1, \dots, r_n | s)$ :

$$\begin{aligned} L(s) &= p(r_1 = 2, r_2 = 3, \dots, r_{31} = 0 | s) \\ &= p(r_1 = 2 | s) p(r_2 = 3 | s) \cdots p(r_{31} = 0 | s). \end{aligned} \tag{10.8}$$

More generally, the likelihood is obtained by substituting Eq. (10.7) into Eq. (10.6):

$$L(s; \mathbf{r}) = \prod_{i=1}^n \left( \frac{1}{r_i!} e^{-g_i e^{-\frac{(s-s_{\text{pref},i})^2}{2\sigma_{\text{tc},i}^2}}} - b_i \left( g_i e^{-\frac{(s-s_{\text{pref},i})^2}{2\sigma_{\text{tc},i}^2}} + b_i \right)^{r_i} \right). \tag{10.9}$$

Three resulting likelihoods are plotted together in Fig. 10.10.

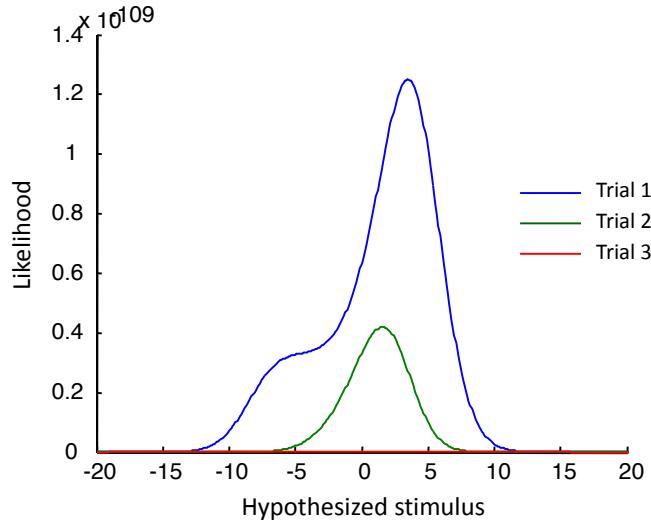


Fig. 10.10: Three likelihood functions from the same population, obtained with  $s=0$

### 10.3.3 Properties

Several properties of the likelihood functions merit discussion. First, the scale on the y-axis is very small ( $10^{-109}$ ). This low magnitude of the likelihood functions results merely from the fact that the probability of multiple events (spikes in multiple neurons) is always less than the probability of any one of those events.

Second, even in a large population, the likelihood can be far from Gaussian (see Trial 1). Some likelihood functions have two maxima, others have a flat top, yet others are skew. At the same time, almost all likelihood functions have a distinct, dominant peak, and most look more or less Gaussian. This is an empirical generality: the more neurons a population encoding a continuous stimulus contains (and the more they fire), the closer to Gaussian the likelihood functions look. The smooth and structured form of the likelihood function stands in contrast to the messy and apparently structureless population pattern of activity in Fig. 10.7.

Third, the likelihood functions vary enormously in peak height. The one from Trial 3 is so low overall that it is not even visible on this scale. Its peak height is  $3.8 \cdot 10^{-113}$ . Since the shape of the likelihood function (the single peak and the narrow width) are its most important features, as these will influence the observer's conclusion as to the value of the stimulus. Here, we have normalized the same three likelihood functions:

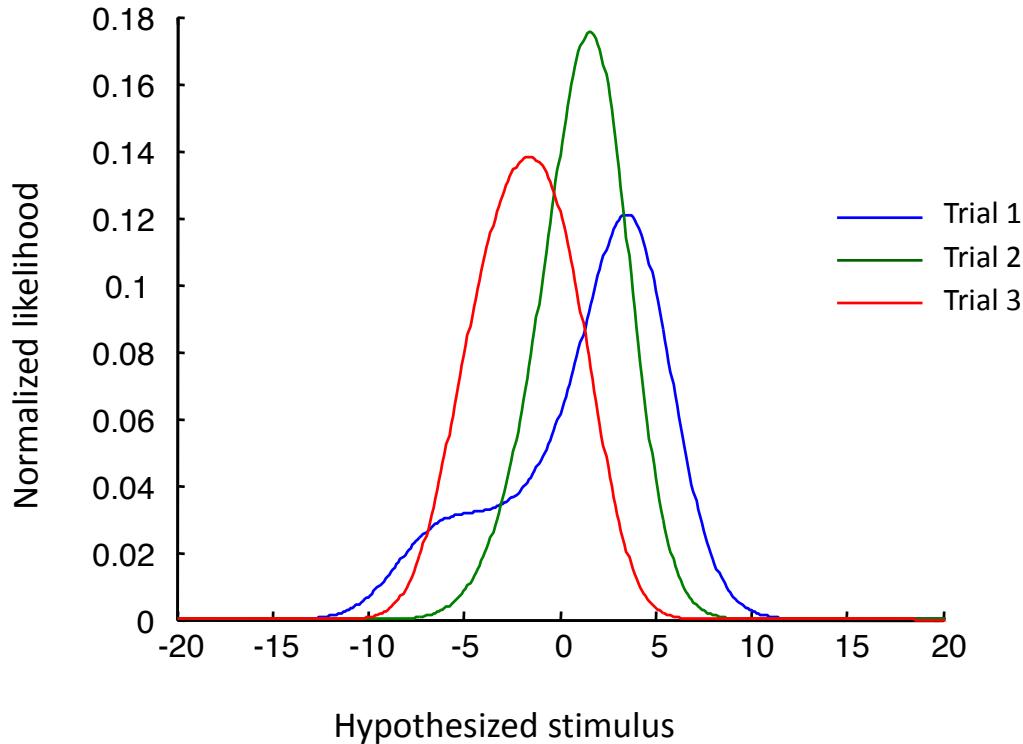


Fig. 10.11: Normalized versions of the likelihoods in Fig. 10.10.

Because of the normalization, the area under each likelihood function is now equal to 1. This makes the Trial 3 likelihood function clearly visible. The normalized likelihood function is a posterior distribution. In particular, it is the posterior distribution when the prior is uniform.

Exercise: Why?

Besides visibility in plots, there is usually no reason to normalize the likelihood function. The shape is the same with or without normalization, and usually the shape is most important. The unnormalized likelihood function is the fundamental entity that is encoded in the neural population pattern  $\mathbf{r}$ .

As a final illustration of the diversity of likelihood functions one obtains even when the stimulus is kept fixed, we now show likelihood functions from 10 trials from the population in Fig. 10.9:

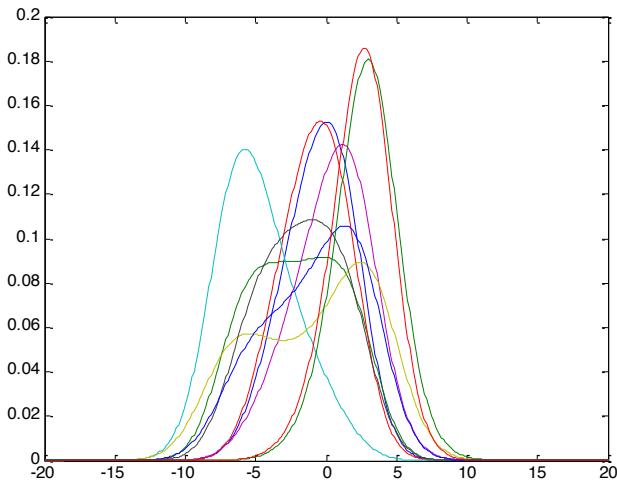


Fig. 10.12: More likelihood function from the same population, still with  $s=0$ .

Exercise: Write your own code to create likelihood functions like those in Fig. 10.12. Run your code multiple times to get an idea of the diversity.

[REDO FIGURES. CONSOLIDATE CODE FOR CH 9 and 10. PARAMETERS MUST MATCH EXACTLY]

#### 10.3.4 Case 2: Monotonic tuning curves

$$f_i(s) = a_i s^{b_i} + c_i$$

where all  $a_i > 0$  and  $c_i \geq 0$ .

[WRITE THIS SECTION IN COMPLETE PARALLEL TO CASE 1]

#### 10.3.5 Concluding remarks

The computation of the likelihood function based on a set of neurons with independent noise is conceptually similar to cue combination as discussed in Chapter 4. A neuron's spike count is analogous to a measurement, and the likelihood (Eq. (10.8)) is obtained by multiplying the likelihoods from the individual neurons, just as the likelihoods from individual cues are multiplied together in cue combination. The difference is that each of the individual neuronal likelihoods is by no means Gaussian (see Fig. 10.2), and yet their product is.

Points to remember:

- A likelihood function over the stimulus can always be obtained from single-trial patterns of neural population activity. This likelihood function can be used by a Bayesian observer in subsequent computation.
- The likelihood function is different on every trial, even if the stimulus is kept the same. This is because the likelihood function is determined by the pattern of neural activity on a trial, and this pattern varies stochastically from trial to trial.
- Likelihood functions come in a variety of shapes and sizes (peak heights), but when the number of neurons is large, often look Gaussian.

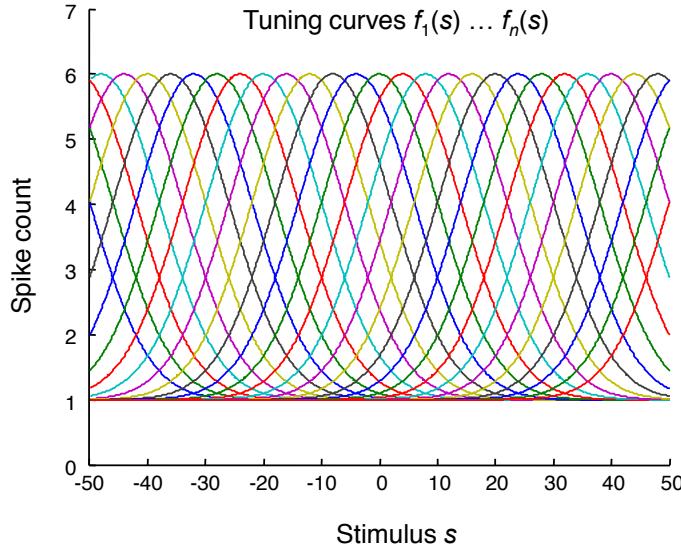
## 10.4 Toy models

So far, our population likelihood functions such as (10.9) were complicated and did not provide any intuition about how its properties depend on neural activity. To achieve such intuition, we will in this section consider specific examples of the bell-shaped and monotonic curve families, which allow for compact expressions. We emphasize that this will force us to consider biologically rather unrealistic settings; however, the intuition we will gain will generalize. Borrowing from the language of physics, the models we describe in this section will be “toy models”: they capture the essence without being realistic.

### 10.4.1 Case 1: Bell-shaped tuning curves

Within this class, we can gain intuition by assume that preferred stimuli of the neurons are equally and densely spaced across the entire real line (there are thus infinitely many neurons), their tuning curves are translated versions of each other, and have baseline 0. In other words, we use Eq. with  $b=0$  (depicted in Fig. 10.X):

$$f_i(s) = g e^{-\frac{(s-s_{\text{pref},i})^2}{2\sigma_{\text{tc}}^2}}. \quad (10.10)$$



[REPLACE TUNING CURVE FIGURE BY BASELINE 0]

We start from the expression for the log likelihood from Eq. :

$$\log L(s; \mathbf{r}) = -\sum_{i=1}^n \log r_i! - \sum_{i=1}^n f_i(s) - \sum_{i=1}^n r_i \log f_i(s)$$

The first two factors are independent of  $s$ , they are just multiplicative constants. The first factor does depend on the spike counts so will change from trial to trial, but on a given trial it is just a constant.

We now assume that the sum of the neural tuning curves over neurons is more or less independent of the stimulus:

$$\sum_{i=1}^n f_i(s) = k \quad (10.11)$$

In general, the left-hand side will depend on  $s$ , but if tuning curves are sufficiently dense and sufficiently similar to each other, Eq. (10.11) is a good approximation. For the population in Fig. 6, that is the case. For example, for the population in Fig. 10.9, it is not and the sum looks like this:

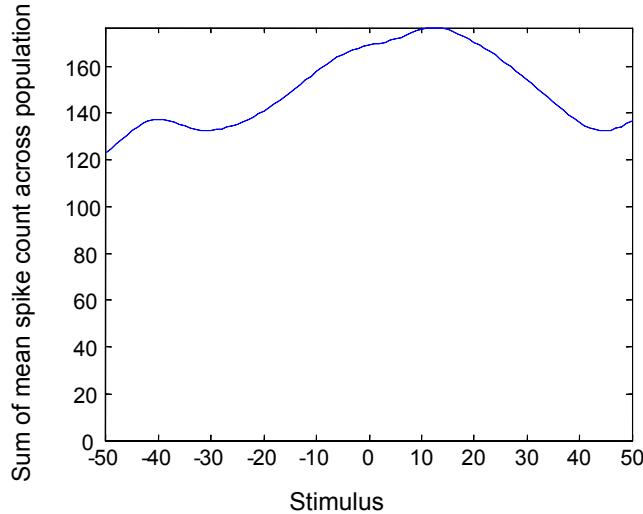


Figure 10.13: Sum of the tuning curves in Fig. 10.9 across all neurons in the population.

(Since we only defined tuning curves in a limited region of space, the approximation will only hold in that region; the sum will drop to zero for values of  $s$  outside this region.) In this case, Eq. (10.11) is only a coarse approximation.

If one is only interested in the *shape* of the likelihood function, as is usually the case (for example to obtain an estimate of the stimulus, or an estimate of uncertainty), then the multiplicative constants don't matter. We can then even write Eq. **Error! Reference source not found.** as

$$L(s) \propto \prod_{i=1}^n f_i(s)^{r_i}, \quad (10.12)$$

where the proportionality sign absorbs all  $s$ -independent factors. This is not the same as normalizing, since we do not compute the normalization constant. To summarize what we have found so far, using the constant-sum approximation we obtained a concise expression for the neural likelihood function, which takes the form of products of tuning curves raised to the powers of the corresponding spike counts. The higher the spike count, the higher the power, and the more influence that neuron's tuning curve has on the likelihood function. Thus, if tuning curves are single-peaked (unimodal), as in Fig. 10.9, then the likelihood tends to peak near the preferred stimuli of the highest-firing neurons.

We substitute Eq. (10.10) into Eq. **Error! Reference source not found.**, to find

$$\begin{aligned}
\log L(s) &= \sum_{i=1}^N r_i \log \left( g e^{-\frac{(s-s_{\text{pref},i})^2}{2\sigma_{\text{tc}}^2}} \right) + \text{constant} \\
&= \sum_{i=1}^N r_i \left( \log g - \frac{(s-s_{\text{pref},i})^2}{2\sigma_{\text{tc}}^2} \right) + \text{constant} \\
&= \sum_{i=1}^N r_i \log g - \frac{1}{2} \sum_{i=1}^N r_i \frac{(s-s_{\text{pref},i})^2}{\sigma_{\text{tc}}^2} + \text{constant} \\
&= -\frac{1}{2\sigma_{\text{tc}}^2} \sum_{i=1}^N r_i (s-s_{\text{pref},i})^2 + \text{constant}
\end{aligned} \tag{10.13}$$

In the last step, we have absorbed the factor  $\sum_{i=1}^N r_i \log g$  into the additive constant; we can do that because it does not depend on  $s$ ; in other words, the shape of the likelihood is the same regardless of this factor. The sum in the last line of Eq. (10.13) consists of a quadratic function of  $s$  in every term. We can regroup these terms into a single second-order ( $s^2$ ) term, a single first-order term ( $s$ ), and a constant. The reader can now see why choosing zero as baseline was important. This simplification would not have been possible without. We now have a Gaussian likelihood function

(10.14)

$$L(s) \propto \exp \left( -\frac{(s-\mu_{\text{likelihood}})^2}{2\sigma_{\text{likelihood}}^2} \right)$$

$$\text{where } \mu_{\text{likelihood}} = \frac{\sum_{i=1}^N r_i s_{\text{pref},i}}{\sum_{i=1}^N r_i} \text{ and } \sigma_{\text{likelihood}}^2 = \frac{\sigma_{\text{tc}}^2}{\sum_{i=1}^N r_i} .$$

Exercise: Show this.

There is a good reason to write the likelihood function like this: we recognize the form of an (unnormalized) Gaussian function! In other words, when a population of independent Poisson neurons have Gaussian tuning curves without baselines, and we make the constant-sum approximation, then the neural likelihood function over the stimulus is Gaussian. This property

makes this special case valuable as a toy model. We plot normalized likelihood functions obtained from this population using the complete equation, Eq. [Error! Reference source not found.](#), overlaid with the approximate expression, Eq. (10.14).

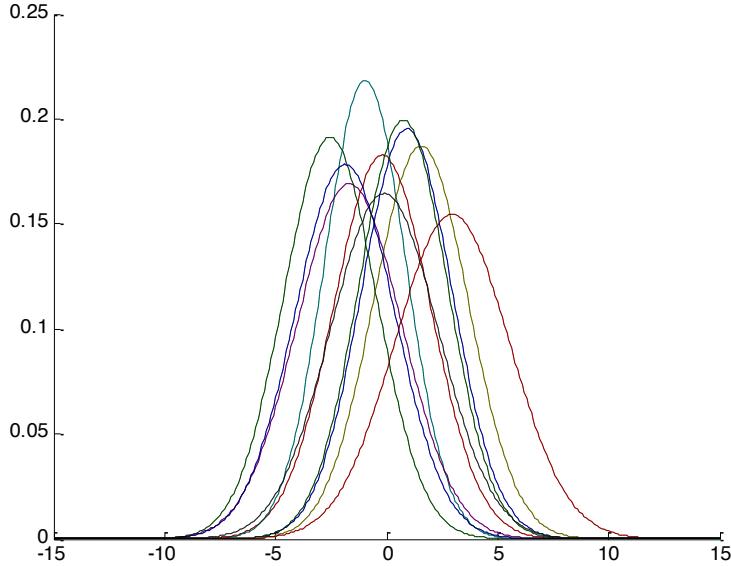


Fig. 10.14: we have overlaid the complete expression with the constant-sum approximation, in black dots. The approximation is indistinguishable from the complete expression. **TO DO:** To make this figure more satisfying, maybe make the black dots larger, but also only plot them at every 10 steps along the x-axis or etc., so that both the dots and the curves show up nicely.

If you only see one line, that is correct: the approximation is essentially indistinguishable from the complete expression. The main things to note in Fig. 10.14:

- The likelihood functions are all proportional to Gaussian distribution, in accordance with Eq. (10.14).
- Both the mode of the likelihood function and its width vary from trial to trial.

Let's examine the mode and width in more detail. We can directly read off expressions for these two quantities from Eq. (10.14), since we know how to recognize mean and standard deviation in the equation of a Gaussian (and for a Gaussian, mode = mean). The mode is called the maximum-likelihood estimate (MLE) of the stimulus. The MLE is

$$\hat{s}_{\text{MLE}} = \frac{\sum_{i=1}^N r_i s_{\text{pref},i}}{\sum_{i=1}^N r_i} \quad (10.15)$$

Thus, under the assumptions made, the maximum-likelihood estimate of the stimulus is a weighted sum of the preferred stimuli of the neurons in the population, with weights given by the neurons' spike counts. This is also called the population vector decoder.

If one were to normalize the likelihood function as we did in Fig. 10.14, one could speak of its variance or standard deviation. For an unnormalized likelihood function, it is more accurate to be more vague and call what would have been the standard deviation the width. It is interpreted as the uncertainty that the observer has about the stimulus, and sometimes also referred to as the *sensory uncertainty*. This width is

$$\sigma_{\text{likelihood}} = \frac{\sigma_{\text{tc}}}{\sqrt{\sum_{i=1}^N r_i}}. \quad (10.16)$$

This expression, plotted in Fig. 10.15, makes intuitive sense: the higher the total spike count in the population, the narrower the likelihood function and the lower the sensory uncertainty. Also, the narrower the tuning curve (smaller  $\sigma_{\text{tc}}$ ), the narrower the likelihood function. We can think of Eq. (10.16) as stating that sensory uncertainty is *encoded in the trial-to-trial neural activity* in the population. Sensory uncertainty information is present in a *distributed* manner, since all neurons contribute to the sum in Eq. (10.16). Even though the equation was derived under specific assumptions, the general concept that sensory uncertainty is encoded in the trial-to-trial population activity is completely general.

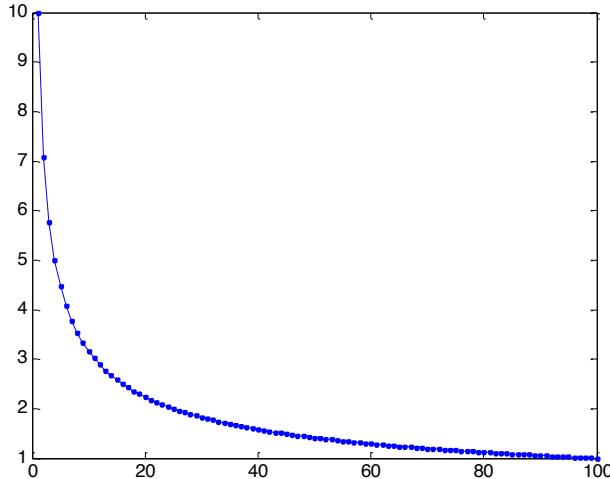


Fig. 10.15: Sensory uncertainty from a neural population: Width of the neural likelihood function as a function of the total spike count in the population, for an independent Poisson population with tuning curves as in Fig. 10.6 (tuning curves have width 10).

Exercises: 1) When only one neuron fires one spike, Eq. (10.16) states that the width of the likelihood is equal to the width of the tuning curve. Explain why this is correct and why this scenario is different from the one-neuron brain discussed in Section 10.2?  
2) What is the likelihood function when the entire population is silent?

Something that you might not have anticipated is that sensory uncertainty itself will vary from trial to trial as the total spike count varies; in other words, not all likelihoods are equally wide. This stands in contrast to the first half of the book, in which the likelihood always had the same width for the same stimulus condition. However, the neural formulation is more accurate than the behavioral formulation; in real neural systems, the width of the likelihood function will vary from trial to trial.

The assumptions we made (independent Poisson; Gaussian tuning curves with zero baseline; constant-sum approximation) allowed us to write down explicit equations for the maximum-likelihood estimate and the width of the likelihood function. That we were able to do this depended strongly on these assumptions: if we had removed any one of them, no such equations could have been formulated. More generally, it is rare that closed-form expressions for the maximum-likelihood estimate and the width of the likelihood function can be written down; therefore, one typically needs to compute likelihood functions via numeric simulation.

To summarize, the case of independent Poisson neurons with Gaussian tuning curves and the constant-sum approximation is a very useful toy model, because the likelihood function is exactly Gaussian and we can find analytical, intuitive expressions for both the maximum-likelihood estimate (population vector or weighted average) and the likelihood width (inversely related to the total spike count).

#### 10.4.2 Case 2: Monotonic tuning curves

A compact equation for the neural likelihood function for a population in which the tuning curves are monotonic can be obtained when  $b_i=1$  and  $c_i=0$ :

$$f_i(s) = a_i s$$

Then the population likelihood function is

$$\begin{aligned} L(s; \mathbf{r}) &= \prod_{i=1}^n \left( \frac{1}{r_i!} e^{-a_i s} (a_i s)^{r_i} \right) \\ &\propto e^{-s \sum_{i=1}^n a_i} s^{\sum_{i=1}^n r_i} \end{aligned}$$

This is proportional to a gamma distribution with..

#### 10.5 Relation between behavioral and neural concepts

Let us take stock of what we have learned in the previous two sections. We introduced a *generative model for neural activity*, namely independent Poisson variability combined with certain assumptions about tuning curves. Using this generative model, we computed the *neural likelihood function* based on a population pattern of activity  $\mathbf{r}$ . We found specific expressions for

the maximum-likelihood estimate and the width of the likelihood function. At this point, it is useful to compare and contrast these results with the generative model we introduced in Chapter 2, which was the basis for all behavioral (non-neural) models, up to and including Chapter 8. We have put corresponding quantities in the table below:

Quantity	Behavioral model	Neural model
Observation	Scalar measurement $x$ Possible values: same as of stimulus $s$	Vector of spike counts $\mathbf{r} = (r_1, \dots, r_n)$ Possible values: positive integers
Noise distribution	$p(x s)$ , typically Gaussian with mean $s$ and standard deviation $\sigma$	$p(\mathbf{r} s)$ , for example independent Poisson with Gaussian tuning curves and a constant-sum approximation
Likelihood over $s$	$L(s) = p(x s)$ , Gaussian if noise distribution is Gaussian	$L(s) = p(\mathbf{r} s)$ , Gaussian in the example but not in general
Maximum-likelihood estimate of $s$	$x$	In the example, $\frac{\sum_{i=1}^N r_i s_{\text{pref},i}}{\sum_{i=1}^N r_i}$
Width of likelihood function	$\sigma$	In the example, $\frac{\sigma_{\text{tc}}}{\sqrt{\sum_{i=1}^N r_i}}$

We can see from this table that the behavioral model was simplified in several ways: first, the likelihood function was always Gaussian, while in the neural model, it is not. Second, the maximum-likelihood estimate is identical to the measurement, which is made possible by the fact that the measurement lives in the same space as (has the same domain as) the stimulus; in the neural model, the observation lives in a completely different space (the space of  $n$ -dimensional vectors of positive integers) than the stimulus, and therefore also than the maximum-likelihood estimate of the stimulus. Third, in the behavioral model, the likelihood width was the same from trial to trial; in the neural model, since it depends on the observation, it varies from trial to trial.

Looking back at the behavioral models, we can appreciate now that the concept of a measurement was an abstraction. The brain itself does not have scalar measurements with which to do inference; it only has neural action potentials. In fact, we could now *define* the concept of a measurement in terms of the neural model: the measurement is the maximum-likelihood estimate of the stimulus based on the neural observation, namely the population pattern of activity  $\mathbf{r}$ . We can think of the measurement  $x$  as a “processed form” of the neural activity  $\mathbf{r}$ . This is illustrated in Fig. 10.18.

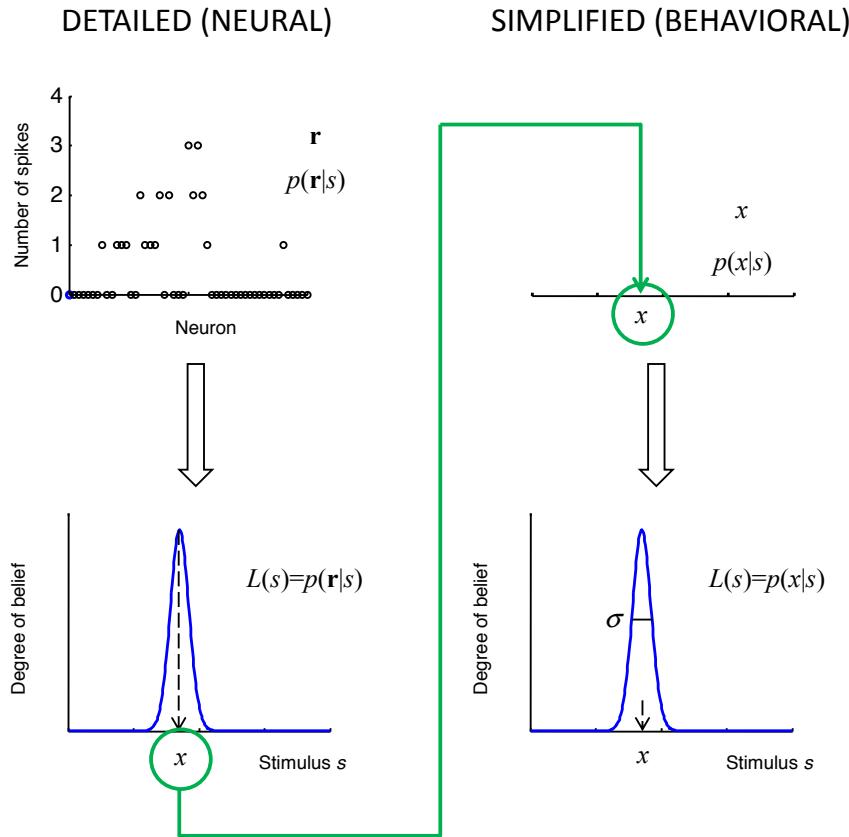


Fig. 11.18

**Definition:** The measurement  $x$  of a stimulus  $s$  is the value of  $s$  under which the observed neural activity  $\mathbf{r}$  is most probable.

Under this definition, the Gaussian distribution to describe  $p(x|s)$  is an approximation to the neural distribution of the maximum-likelihood estimate, which we will examine in Section 10.6

The internal representation may directly relate to the readings of one of our primary sensors. For example, the pattern of activation of photoreceptors in the retina or of the mechanoreceptors embedded in our skin may be considered an internal representation. Alternatively, the internal representation might be associated with activity in a downstream brain area. For example, the activity in primary visual cortex is also an internal representation.

The neural activity elicited by a stimulus will vary randomly from trial to trial, even when the physical stimulus itself is identical each time. Thus, we say that the internal representation of a stimulus is noisy. As we discussed in Chapter 2, noise originates from many sources, including thermal noise, photon shot noise, neurotransmitter release, and ion channel opening and closing. We define a *measurement* as the best possible guess about the stimulus based on the internal representation alone. “Best possible” here means that this value has the highest probability of generating the observed internal representation. If the stimulus is an orientation of a line and the

internal representation is the pattern of activation of retinal photoreceptors, then the measurement would be the best guess of orientation obtained from this pattern. If the stimulus is the size of an object and the internal representation is the pattern of activation of skin mechanoreceptors when holding the object, then the measurement would be the best guess of size based on this pattern. In the auditory localization example, the measurement could be the best guess of location based on the activation of the hair cells in the inner ears.

Since the internal representation is noisy, the measurement will be noisy as well. The full internal representation typically occupies a high-dimensional space that is very different from the stimulus space; it could for example be a space of neural activity, such as the firing rates of a population of sensory neurons in the cortex. One could think of the mapping from internal representation to measurement as “pre-processing”, since the computation we will focus on takes the measurement(s) as input.

## 10.6 Statistics of likelihood mode and width across many trials

Step 3 of the Bayesian modeling framework outlined in Chapter 2 and followed throughout the behavioral chapters was to predict behavior across many trials, or, to be more precise, to calculate the distribution of the MAP estimate conditioned on the stimulus. In the neural context, this distribution almost always has to be computed numerically.

There is a particular inference task that was trivial in the behavioral context but is not in the neural one, which is simply to estimate the stimulus on a continuum under a uniform prior. In that case, the observer’s MAP estimate is the maximum-likelihood estimate and the width of the posterior is equal to the width of the likelihood function. In the behavioral model, the maximum-likelihood estimate was  $x$ , so the distribution of the maximum-likelihood estimate given the stimulus was the same as the noise distribution, namely  $p(x|s)$ . To be mathematically precise,

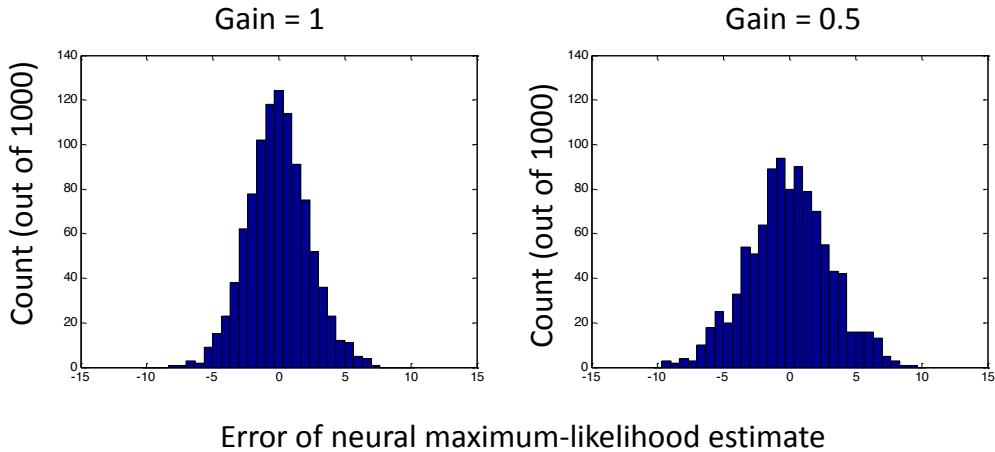
$$p_{\hat{s}|s}(\hat{s}|s) = p_{x|s}(\hat{s}|s) .$$

The likelihood function on every trial, the posterior distribution on every trial, and the distribution of the maximum-likelihood estimate all had standard deviation  $\sigma$  in this case..

### 10.6.1 Toy model

In the neural toy model of section 5, the maximum-likelihood estimate is  $\frac{\sum_{i=1}^N r_i s_{\text{pref},i}}{\sum_{i=1}^N r_i}$  and the width of the likelihood function is  $\frac{\sigma_{\text{tc}}}{\sqrt{\sum_{i=1}^N r_i}}$ . Both quantities are random variables, with distributions that they inherit from the distribution of the spike counts  $\{r_i\}$ . We will now examine the distributions of the mode and the width of the likelihood function over many trials.

The distribution of the quantity  $\frac{\sum_{i=1}^N r_i s_{\text{pref},i}}{\sum_{i=1}^N r_i}$  as  $r_i$  is drawn from its Poisson distribution, cannot be calculated analytically. However, we can simulate it, which we have done in Fig. 11.16 for  $s=0$  and two levels of contrast (which correspond to two levels of gain,  $g=1$  and  $g=0.5$ ).

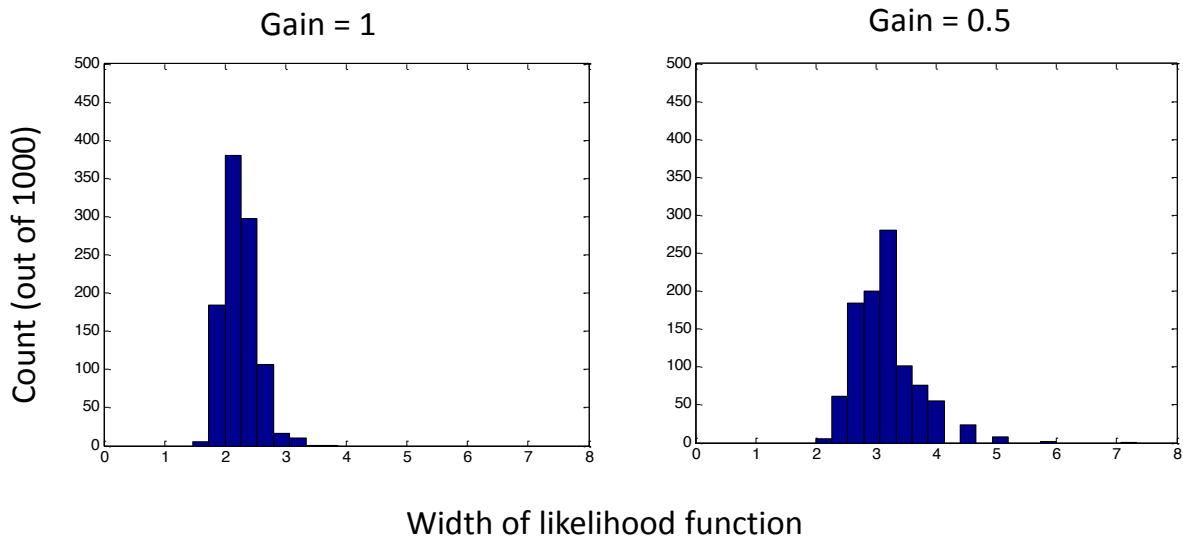


These distributions resemble Gaussians, but that is no proof that they are theoretically Gaussians. In fact, they are not theoretically Gaussians. Only in the limit that gain is very high (high spike counts on average) can one state that statistically, the distribution of the maximum-likelihood estimate becomes indistinguishable from a Gaussian. This is a property known as asymptotic normality; it is essentially a limit when the amount of information in the population is large, because a sufficient number of neurons has responded with a sufficient number of spikes. If this condition is not satisfied, it is not guaranteed that the Gaussian distribution is a good approximation to the true distribution of the maximum-likelihood estimate. Under similar conditions, it can be shown as well that the mean value of the maximum-likelihood estimate over

many trials is equal to the stimulus itself, i.e.  $\langle \hat{s}_{\text{ML}} \rangle = s$ . The technical statement is “the maximum-likelihood estimate is asymptotically unbiased”. These properties provide some degree of justification for our assumptions in the behavioral models that the measurement has a Gaussian distribution and is on average equal to  $s$ . However, it is important to keep in mind that both are approximations to the underlying neural model.

The width of the likelihood function,  $\frac{\sigma_{\text{tc}}}{\sqrt{\sum_{i=1}^N r_i}}$ , also varies from trial to trial, and its

distribution cannot be calculated analytically either. The result of simulation is shown in Fig. 11.17. As one would expect, the width of the likelihood function is on average lower when gain is higher. Moreover, what would have been harder to anticipate, the variation in the width is also lower when gain is higher. These distributions are skewed, with a sharp rise and a longer tail, but otherwise difficult to characterize. Intuitively, the fact that these distributions are not very narrow means that the observer’s sensory uncertainty will vary substantially even as the physical stimulus is the same.



We can now add two rows to Table 11.X that pertain to statistics across trials:

Quantity	Behavioral model	Neural model
Distribution of MLE	Gaussian with mean $s$ and standard deviation $\sigma$ (same as noise distribution)	No analytical form; only Gaussian in the limit of many spikes
Distribution of likelihood width	Delta function at $\sigma$ (always the same)	Wide distribution (no analytical form)

Besides looking at the distributions of the mode and the width of the likelihood function separately, we can also look at the trial-to-trial correlations between both quantities. Intuitively

such a correlation would mean that on trials when you are less certain, you also perform worse. We will examine this in a Problem.

### 10.6.2 Distinction between $p(r|s)$ and $p(r|I)$

Having discussed the neural likelihood function provides us with a basis to formulate a neurally based framework for perception (Fig. 3.1a). The general structure of a perception model we described in Chapter 2, consisted of a stimulus  $s$ , a measurement  $x$ , and a stimulus estimate. This was a simplified conceptualization. In describing the auditory task in Chapter 2, we somewhat loosely referred to sound location as the “stimulus”, ignoring its neural component.

The likelihood over  $s$  based on  $r$ , denoted  $L_r(s)$ , was the focus of this chapter so far. We could also have computed a likelihood function over  $s$  from the sensory input  $I$  rather than the neural activity  $r$ :

$$L_I(s) = p(I|s)$$

This function represents all information that can be obtained about  $s$  from the sensory input  $I$ . It is not necessarily the same as  $L_r(s)$ . In general, information will be lost between the sensory input  $I$  and  $r$ , as mentioned above. As a consequence,  $L_r(s)$  is wider than  $L_I(s)$ , reflecting a greater amount of uncertainty. We are faced with the interesting problem: what can we, statistically, say about the world given our observed neural representation  $r$ .

### 10.7 Generalization: multivariate normal variability (\*)

$$p(r|s) = \frac{1}{\sqrt{\det 2\pi\Sigma}} e^{-\frac{1}{2}r^T\Sigma^{-1}r}$$

[...text]

### 10.8 Applications

Although the focus of this book is how the brain performs inference based on noisy sensory information, there is a parallel literature on how an experimenter can decode brain state on a trial-to-trial basis. Traditionally, this literature has focused on point estimates; however, in recent years, more attention has been given to decoding entire likelihood functions – thus linking to the rest of this chapter.

#### 10.8.1 fMRI

Functional magnetic resonance imaging (fMRI) is a method that uses big magnets and microwaves to measures the three dimensional oxygenation of blood in the head due to neural activity. The brain is divided into voxels (like pixels, but volume elements – small cubes), and

one records “percent signal change” in each voxel in response to presented stimuli. Of natural interest is an application of “mind-reading” of decoding what the brain thinks or sees based on fMRI data.

We consider an example in which the observer views oriented stimuli, and the orientation is the only feature that changes. The logic here is a bit different from the rest of the book, since our model is not a model of the observer: the human subject’s observations are not voxel activities, but receptor neuron activities. Instead, we as experimenters use the generative model to decode the stimulus from fMRI activity on every trial. Thus, the observer is the experimenter, not the human subject. Just like we saw in chapter [LEARNING] that the observer must first learn the parameters of the generative model of its observations to perform inference, we as experimenters have to learn the parameters of the generative model of voxel activity. For that purpose, we use training data, in which the orientation is considered known on every trial.

### Generative model

Voxel activity is noisy. We denote by  $s_j$  is the orientation in the  $j^{\text{th}}$  trial). We denote by  $\mathbf{B} = (B_1, B_2, \dots, B_{N_{\text{voxel}}})^T$  is an  $N_{\text{voxel}} \times 1$  vector of voxel activity. We assume that each voxel contains orientation “channels”, and the proportions of different channels vary across voxels. We denote by  $\mathbf{C}(s) = \{C_k(s)\}$  an  $N_{\text{channel}} \times 1$  vector of channel tuning curves, and by  $\mathbf{W} = \{W_{ik}\}$  an  $N_{\text{voxel}} \times N_{\text{channel}}$  matrix of channel weights for all voxels. We assume that the activities of the voxels in  $\mathbf{B}$  are independent given the channel activity  $\mathbf{C}(s)$ . The  $i^{\text{th}}$  voxel follows a Gaussian distribution with mean and  $\sum_k W_{ik} C_k(s)$  standard deviation  $\sigma$ . Then the distribution of the entire

activity vector  $\mathbf{B}$  is a product:

$$p(\mathbf{B} | s; \mathbf{W}, \sigma) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(B_i - \sum_k W_{ik} C_k(s))^2}{2\sigma^2}}$$

after the parameters of this model,  $\mathbf{W}$  and  $\sigma$ , have been learned using a set of training trials, we can compute, on each testing trial, a voxel-based likelihood function over  $s$ :

$$L(s) = p(\mathbf{B} | s; \hat{\mathbf{W}}, \hat{\sigma}) = \prod_i \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} e^{-\frac{(B_i - \sum_k \hat{W}_{ik} C_k(s))^2}{2\hat{\sigma}^2}}$$

### 10.8.2 Brain machine interfaces

The calculation of posterior distributions based on neural activities for perception has a close analog in the technical problem of decoding in the context of brain machine interfaces.



**Figure 10.8.** A monkey feeding itself through a BMI system. Photo courtesy Schwartz lab.

Following a spinal cord injury that results in paralysis, a person might be outfitted with prosthetic arms. But how can these prosthetic limbs be controlled? One possibility is to use voice commands, but this procedure is cumbersome. Another possibility is to use eye movements. A third possibility, that at first seems worthy of a science-fiction story, is to control the prosthetic device with ones thoughts. This can be accomplished through the use of a Brain-Machine interface that reads the user's intent from her neural activity. Recorded neural signals from the motor cortex, properly interpreted via Bayesian inference, indicate her intended movements. In such BMI scenarios, we are interested in calculating the posterior distribution over the movement intent, given the recorded neural activity:  $p(C|r)$ .

## 10.9 Problems

**10.1** In Section **Error! Reference source not found.** (toy model), we assumed zero-baseline Gaussian tuning curves.

- What changes if the baseline is not zero?
- For each of the baseline values 0, 0.25, 0.5, and 1, numerically compute and plot 10 likelihood functions (assume the true stimulus is zero), and describe what you see.

**10.2** In the toy model, we assumed the same tuning width  $\sigma_{tc}$  for all neurons. Derive the equivalent of Equations (10.15) and (10.16) for the mode and the width of the likelihood function if every neuron has its own tuning width, say  $\sigma_{tc,i}$  for the  $i^{\text{th}}$  neuron.

**10.3** In Section 10.6, we mentioned that the width of the likelihood function might be correlated with the error of the maximum-likelihood estimate.

- Simulate the toy model of this Chapter for 10,000 trials (all at  $s=0$ , and gain 1) and create a scatter plot of the squared error of the maximum-likelihood estimate versus the squared width (variance) of the likelihood function. What is the correlation coefficient?
- Choose several different values of gain. Does the strength of the correlation depend on gain?

c) Now divide the trials into four quartiles for the variance of the likelihood function. Compute the variance of the maximum-likelihood estimate for each of these four groups of trials. Is there a correlation?

**10.4** We introduced the Poisson-like family of distributions.

- a) Relate the assumptions we made in the toy model to properties of  $\varphi$  and  $\mathbf{h}$  in the Poisson-like equation.
- b) If  $\mathbf{h}(s)$  is quadratic in  $s$ , say  $\mathbf{h}(s) = \mathbf{a}s^2 + \mathbf{b}s$ , derive expressions for the mode and width of the neural likelihood function,

**10.5** We determined earlier that the mode and width of the likelihood function in the toy model are independent of contrast. However, when contrast is unknown, the Bayesian observer would marginalize out contrast:  $L(s) = p(\mathbf{r} | s) = \int p(\mathbf{r} | s, c) p(c) dc$ . Show that the likelihood mode and width only depend on  $\mathbf{r}$ , not on the prior over contrast,

- a) in the toy model
- b) for Poisson-like variability.

**10.6** Consider the case that the tuning curve width depends on contrast in the toy model. What can then be said about the likelihood function

- a) when contrast is known to the observer;
- b) when contrast is unknown and marginalized over.

**10.8.** In Problems 2.20 and 3.6, we introduced inference on stimulus variables that take values on the circle, such as motion direction, which takes values between (for instance)  $-\pi$  and  $\pi$ . In a laboratory experiment, motion direction is drawn from a Von Mises distribution with circular mean  $\mu_s$  and concentration parameter  $\kappa_s$ :

$$p(s) \propto e^{\kappa_s \cos(s - \mu_s)}$$

- a) Assume that motion direction, denoted  $s$ , is encoded in a population of  $n$  independent Poisson neurons. The tuning curve of the  $i^{\text{th}}$  neuron has a Von Mises shape with gain  $g$ , preferred direction  $s_{\text{pref},i}$ , and concentration parameter  $\kappa_{\text{tc}}$ :

$$f_i(s) = g e^{\kappa_{\text{tc}} \cos(s - s_{\text{pref},i})}.$$

Show that the likelihood function over the stimulus based on a population pattern of activity in this population,  $\mathbf{r} = (r_1, \dots, r_n)$ , is proportional to a Von Mises distribution,

$$L(s) \propto e^{\kappa_L \cos(s - \mu_L)},$$

and find expressions for  $\cos(\mu_L)$ ,  $\sin(\mu_L)$ , and  $\kappa_L$ , all in terms of the set  $\{r_i\}$ .

- b) To which well-known population decoder is the maximum-likelihood estimator  $\mu_L$  obtained in part (a) equal?
- c) In Matlab, draw 2000 motion directions  $s$  from the stimulus distribution described in part (c) with  $\mu_s=0$  and  $\kappa_s=4$ . To do this, use the function `circ_vmrnd.m`, which is downloadable from [http://www.mathworks.com/matlabcentral/fileexchange/10676-circular-statistics-toolbox-directional-statistics/content/circ\\_vmrnd.m](http://www.mathworks.com/matlabcentral/fileexchange/10676-circular-statistics-toolbox-directional-statistics/content/circ_vmrnd.m). Each drawn stimulus represents one experimental trial. For each trial, draw a pattern of activity of the population described in part (a) in response to the motion direction on that trial; assume  $g = 0.2$ , preferred directions at every multiple of  $10^\circ$ , and  $\kappa_{tc}=1$ . Hint: the command “`meshgrid`” may be useful when computing the neurons’ mean activities.
- d) Then, again for each trial, compute  $\mu_L$  and the mean of the posterior. You may use the expressions obtained in part (a) and Problem 2.20. If you did not solve those parts, you can do the computations numerically. Hint: the command “`atan2`” is very convenient.
- e) Create a figure consisting of 2 by 2 subplots. The top left subplot should show a scatterplot (command: “`scatter`”; use marker “.” instead of “o”) of  $\mu_L$  against  $s$ . The top right subplot should show a scatterplot of the mean of the posterior against  $s$ . Both subplots should have both an x-range and a y-range from  $-\pi$  to  $\pi$ ; also draw the diagonal as a dashed black line, for reference. The bottom left subplot should show a histogram of the value of the utility function in part (e) when  $\hat{s} = \mu_L$  (maximum-likelihood estimation). Use 20 bins. The bottom right subplot should show the same histogram but with  $\hat{s}$  the mean of the posterior.
- f) Does the maximum-likelihood estimate or the posterior mean have a higher expected utility? Which correlates better with the true stimulus (use “`corr`”)? Are these properties expected?

## Contents

11	Neural computation with probabilities.....	1
11.1	Doing inference with the neural likelihood function .....	2
11.2	Likelihood with prior.....	2
11.3	The duality of probabilities and neural representations .....	3
11.4	The neural posterior for various tasks .....	4
11.5	Implementing the neural posterior with downstream neurons.....	6
11.5.1	Relating back to behavior .....	7
11.6	The neural implementation of Bayesian inference.....	8
11.7	Alternative neural encodings of uncertainty .....	9
11.7.1	Probabilistic population codes.....	10
11.7.2	Sampling codes .....	10
11.7.3	Explicit probability codes.....	10
11.7.4	Convolution codes .....	11
11.7.5	Encoding moments of a distribution .....	12
11.8	Problems .....	13

## 11 Neural computation with probabilities

*How could neurons represent and compute with probabilities?*

If the likelihood functions described in the previous chapter are relevant to neural computation, then the nervous system needs to compute with it. Here we will lay out the basis for how such computations are possible. We will talk about a broad range of inferences where the neural likelihood function matters, but focus on prior-likelihood combination and cue combination.

Following the discussion of how the nervous system could compute with an uncertainty representation that is based on the generative model of chapter 10, we will discuss alternatives. There are many ways of representing uncertainty that have been suggested by the computational neuroscience community. Each of these proposals can easily explain certain biological findings and has trouble explaining others. Also, for each of these proposals some computations are easy to implement while others are very difficult. We will thus discuss a broad set of such proposals.

**Outline of the Chapter:** We start out by discussing how the neural likelihood function from the last chapter can be combined with priors. Then we will move on to the idea of probabilistic population codes and how cue combination can be implemented simply by adding activities in neural populations. We will then continue to discuss other proposed representations of probability distributions, discussing their biological relevance and computational strengths and weaknesses.

## 11.1 Doing inference with the neural likelihood function

In chapter 9, we have defined the neural generative model, which is Step 1 of the Bayesian modeling recipe we outlined in Chapter 2, and in chapter 10 we have derived the neural likelihood function over the stimulus for one particular encoding model, which is part of inference (Step 2). However, the likelihood over the stimulus is typically not all of inference. We already know this from the behavioral models in Chapter 2 through 7: sometimes the world state variable of interest is not any stimulus itself, but a categorical variable (Chapters 4 to 6), or there are multiple stimulus likelihoods that need to be combined (Chapters 3 and 7), or the likelihood must be combined with a prior (Chapter 2). In all these cases, each likelihood function over a stimulus is an elementary building block that is used to build the posterior distribution over the world state of interest. The situation is exactly the same for the neural likelihood. Everywhere where in Chapters 2 to 7 we used a stimulus likelihood function  $L(s) = p(x|s)$ , we can now replace it by a neural likelihood function,  $L(s)=p(\mathbf{r}|s)$ , and everything else would go through as before.

## 11.2 Likelihood with prior

Here we want to go through the example of estimating the posterior over a stimulus variable if we have a prior and the neural likelihood function from the previous chapter. Let us say that  $s$  is drawn from a stimulus distribution  $p(s)$ , the posterior over the stimulus is (from Chapter 2),

$$p(s|\mathbf{r}) \propto p(\mathbf{r}|s)p(s) .$$

Under the toy model of Section **Error! Reference source not found.** (but not in general), the likelihood is proportional to a Gaussian with mean given by Eq. **Error! Reference source not found.** and standard deviation by Eq. **Error! Reference source not found.** Suppose now, as we did throughout Chapter 2, that the stimulus distribution  $p(s)$  is Gaussian with mean  $\mu$  and standard deviation  $\sigma_s$ . Then we are exactly back to the case of Chapter 2, with the only difference the substitutions from Table 11.X,

$$x \rightarrow \frac{\sum_{i=1}^N r_i s_{\text{pref},i}}{\sum_{i=1}^N r_i}$$

$$\sigma \rightarrow \frac{\sigma_{\text{tc}}}{\sqrt{\sum_{i=1}^N r_i}}$$

Thus, we can immediately import the equation for the MAP estimate from Chapter 2, Eq. 2.X, and make these substitutions:

$$\hat{s}_{\text{MAP}} = \frac{\frac{x}{\sigma^2} + \frac{\mu}{\sigma_s^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}} \rightarrow \frac{\frac{1}{\sigma_{\text{tc}}^2} \sum_{i=1}^N r_i s_{\text{pref},i} + \frac{\mu}{\sigma_s^2}}{\frac{1}{\sigma_{\text{tc}}^2} \sum_{i=1}^N r_i + \frac{1}{\sigma_s^2}}.$$

This equation provides the answer to the question: if the sensory input activity is  $\mathbf{r}$  and the stimulus is drawn from  $p(s)$ , how does the brain make the best possible estimate of the stimulus? In other words, it is a neural Bayesian stimulus-response mapping.

The inference problems from Chapters 3 through 8 can be treated through the same substitution of  $L(s) = p(x|s)$  by a neural likelihood function,  $L(s)=p(\mathbf{r}|s)$ . This shows that the same behavioral formalism that we developed for behavior can be translated to neural terms. However, unless we make severe simplifying assumptions for  $p(\mathbf{r}|s)$  like in our toy model, we will in the neural context not be able to find closed-form expressions for the MAP estimate, or for the decision rules in binary tasks. Moreover, there is more to be said about how neural circuits actually implement the operations needed for MAP estimation, which we will address in the next chapter.

### 11.3 The duality of probabilities and neural representations

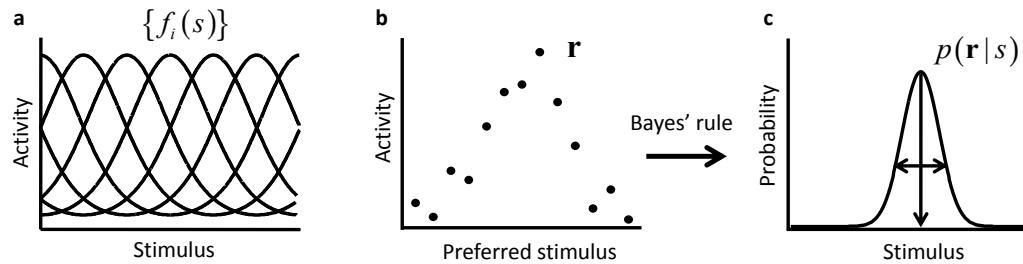
For any given neural representation  $\mathbf{r}$ , if one knows the neural code, one can always calculate the probability distribution  $p(s|\mathbf{r})$  of the associated stimulus variable  $s$ . As the dimensionality of the neural representation is generally far higher than the dimensionality of the variable of interest, there may be many neural representations for the same stimulus probability distribution.

This duality of course exists for any scheme by which the nervous system may encode uncertainty. How to decode or read out a distribution from this kind of data is an entirely different question. Importantly, it is a question that is potentially unnecessary to answer. The nervous system does not generally need to read out all the posteriors over variables. In fact, all that it needs to do is be able to calculate for which action, when integrating over possible values.

In the text below we thus want to discuss two different problems:

- (A) *Decoding*. Given activity vector  $\mathbf{r}$  (or vectors in the case of sampling) how could we and how could the nervous system decode the actual distribution of beliefs over  $s$ .
- (B) *Cue combination*. Given activity vectors  $\mathbf{r}_A$  and  $\mathbf{r}_B$  that both characterize two input cues (generated according to  $p(\mathbf{r}_A|s)$ ,  $p(\mathbf{r}_B|s)$ ), and are otherwise assumed conditional independent on  $s$ : how can we calculate a third activity vector (ideally in a neutrally plausible way) a new activity vector  $\mathbf{r}_C$  so that  $p(\mathbf{r}_C|s) \approx p(\mathbf{r}_A|s)p(\mathbf{r}_B|s)$

In the text below we will focus on two extreme models for these two problems:  
Distinct encoding of mean and variance and probabilistic population codes.



**Figure 13.1: Schematic illustration of probabilistic population coding.** a. Bell-shaped tuning curves of 6 neurons in a hypothetical population. (In real data, these do not look nearly as smooth and identical.) Notation:  $s$  is the stimulus value,  $i$  labels the neuron, and  $f_i(s)$  is the average activity of the  $i$ 'th neuron in response to  $s$ . b. Population pattern of activity elicited by a stimulus (e.g. the orientation of a line segment) on a single trial. Neurons are ordered by their preferred stimuli. c. This pattern of activity encodes a likelihood function over the stimulus (right), providing not only the most likely value of the stimulus (indicated by the arrow), but also its uncertainty (indicated by the double arrow).

## 11.4 The neural posterior for various tasks

- Likelihood + prior (Ch2)
- Cue combination (Ch 4)
- Binary decisions (Ch 5)
- Marginalization (Ch 6)

Also linear neuron

We will start by asking how neurons could implement efficient cue combination (Chapter 4). Let's say we want to localize a stimulus  $s$ , and estimate the direction in which it occurred. For example, a person could be snapping their finger at position  $s$ , which generates two cues about direction, an auditory (A) and a visual (V) cue, each of which is indicative about the direction of the stimulus. The cue in each modality is represented by a neural population; we will denote their patterns

of activity by  $\mathbf{r}_A$  and  $\mathbf{r}_V$ . To implement efficient cue combination a brain area would need to compute the posterior based on these two activity vectors.

How could a downstream area combine the neural activities to allow cue combination? We will assume that each modality-specific area has the same number of neurons, that neurons in each area have the same tuning curves  $f_i(s)$ , and that given the tuning neural responses follow a Poisson distribution. Let the only difference between the two areas be the gains: the mean activities of the auditory neurons are equal to  $g_A f_i(s)$ , whereas the mean activities of the visual neurons are  $g_V f_i(s)$ . If one of these gains are larger, then that area will produce more spikes, and intuitively we should expect it to be more important for the calculation. Characterizing the effect of these gains on the relevant computations promises to be a first step towards understanding how neurons can compute posteriors.

We are now interested in the posterior distribution over  $s$  that is encoded by  $\mathbf{r}_A$  and  $\mathbf{r}_V$  together, we are interested in optimal cue combination. Using Bayes' rule, and assuming that given the stimulus there is no correlation across the two sensory areas, the posterior can be written as:

$$p(s | \mathbf{r}_A, \mathbf{r}_V) \propto p(\mathbf{r}_A, \mathbf{r}_V | s) = p(\mathbf{r}_A | s)p(\mathbf{r}_V | s) \quad (1)$$

As we assume that given a stimulus neural responses follow a Poisson distribution we can put in the equation for Poisson distributions (see Chapter 10) and as we are only interested in  $s$  we can put all factors that do not depend on  $s$  into proportionality factors to obtain:

$$p(s | \mathbf{r}_A, \mathbf{r}_V) \propto \left( \prod_{i=1}^N e^{-g_A f_i(s)} f_i(s)^{r_{Ai}} \right) \left( \prod_{i=1}^N e^{-g_V f_i(s)} f_i(s)^{r_{Vi}} \right). \quad (2)$$

This can be rewritten as:

$$\begin{aligned} p(s | \mathbf{r}_A, \mathbf{r}_V) &\propto \exp \sum_{i=1}^N \left( -(g_A + g_V) f_i(s) + (r_{Ai} + r_{Vi}) \log f_i(s) \right) \\ &= \prod_{i=1}^N e^{-(g_A + g_V) f_i(s)} f_i(s)^{r_{Ai} + r_{Vi}} \end{aligned} \quad (3)$$

We have thus calculated the posterior distribution over  $s$ , given the neural responses in the two sensory areas,  $\mathbf{r}_A$  and  $\mathbf{r}_V$ . We have calculated the posterior

distribution, in probability space, that goes along with the joint information from both neural areas.

## 11.5 Implementing the neural posterior with downstream neurons

However, for neurons to implement cue combination, the result, the posterior, needs to be encoded as neural responses, as a neural representation  $\mathbf{r}_{AV}$  that represents the resulting posterior (Eq. (3)). So which neural activity distribution would be such that  $p(s| \mathbf{r}_{AV}) = p(s| \mathbf{r}_A, \mathbf{r}_V)$ ? Looking carefully at Eq. (3) suggests the answer: addition. We construct a new population pattern of activity,  $\mathbf{r}_{AV}$ , by summing the activities of corresponding pairs of neurons in the auditory and visual populations:

$$\mathbf{r}_{AV} = \mathbf{r}_A + \mathbf{r}_V. \quad (4)$$

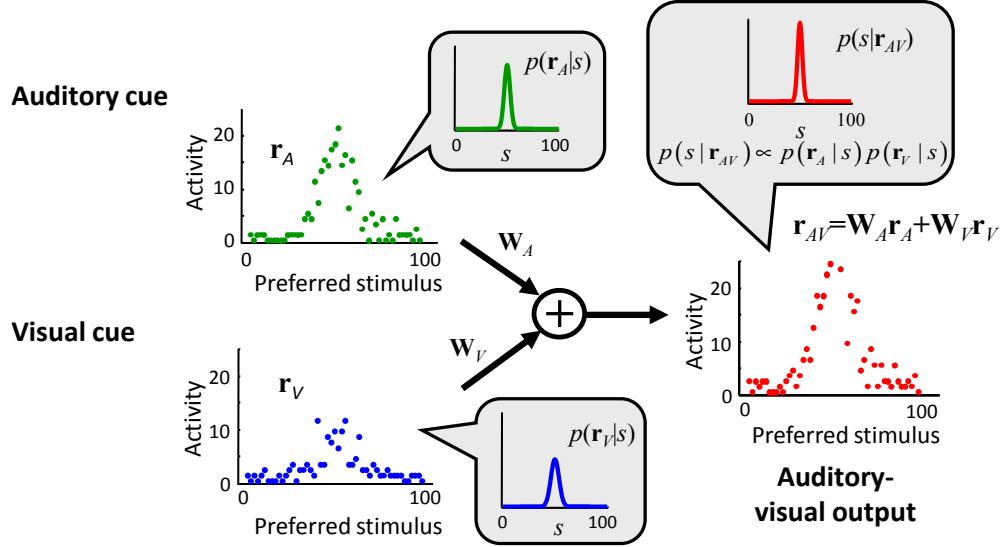
The output population pattern,  $\mathbf{r}_{AV}$ , will still obey independent Poisson variability across many trials, since the sum of two Poisson processes is again Poisson.

This distribution is identical to the one in Eq. (3). We conclude that adding independent Poisson population patterns of activity implements a multiplication of the probability distributions over the stimulus that are encoded in those patterns. In probability space  $\mathbf{r}_A$  and  $\mathbf{r}_V$  correspond to exactly the same posterior distribution as  $\mathbf{r}_{AV}$ .

We have used Auditory and Visual cue combination as an example but other kinds of cue combination could be modeled using the same formalism.

### GENERALIZATION PARAGRAPH TO BE ADDED:

If the brain is Poisson on every layer and you do cue combination then stuff should be linear which can be experimentally tested.



**Figure 13.2: Optimal cue integration with probabilistic population codes.** The cues elicit activity in input populations  $\mathbf{r}_A$  and  $\mathbf{r}_V$ , indicated by green and blue dots. The dialogue boxes show the probability distributions over the stimulus encoded in each population on a single trial. A simple linear combination of the population patterns of activity,  $\mathbf{r}_{AV} = \mathbf{W}_A \mathbf{r}_A + \mathbf{W}_V \mathbf{r}_V$ , guarantees optimal cue integration, if neural variability is Poisson-like. Optimal cue integration means that the probability distribution over the stimulus encoded in the multisensory population is proportional to a product of the likelihoods encoded in the unisensory populations. The synaptic weight matrices  $\mathbf{W}_A$  and  $\mathbf{W}_V$  depend on the tuning curves and covariance matrices of the input populations, but do not have to be adjusted over trials.

### 11.5.1 Relating back to behavior

We will now examine how the neural operation  $\mathbf{r}_{AV} = \mathbf{r}_A + \mathbf{r}_V$  relates to the behavioral equations for multisensory mean and variance that we encountered in earlier chapters, Eq. ... In behavioral modeling of cue integration discussed in this book, it is assumed that the posterior distribution  $p(s|\mathbf{r})$  is Gaussian:

$$p(s|\mathbf{r}) \propto e^{-\frac{1}{2}s^2 a(\mathbf{r}) + sb(\mathbf{r})}, \quad (5)$$

where  $a(\mathbf{r})$  and  $b(\mathbf{r})$  are functions of  $\mathbf{r}$ . Comparing with Eq.

**Error! Reference source not found.**, we see that these functions must be of the form  $a(\mathbf{r}) = \mathbf{a} \cdot \mathbf{r}$  and  $b(\mathbf{r}) = \mathbf{b} \cdot \mathbf{r}$ , where now  $\mathbf{a}$  and  $\mathbf{b}$  are constant vectors. From Eq. (5), we can find the mean  $\mu$  and variance  $\sigma^2$  of the Gaussian, since the exponent of a

Gaussian is of the form  $-\frac{(s-\mu)^2}{2\sigma^2} = -\frac{s^2}{2\sigma^2} + \frac{s\mu}{\sigma^2} + \text{constant}$ . They are given by

$$\frac{1}{\sigma^2} = a(\mathbf{r}) = \mathbf{a} \cdot \mathbf{r} \quad (6)$$

and

$$\frac{\mu}{\sigma^2} = b(\mathbf{r}) = \mathbf{b} \cdot \mathbf{r}. \quad (7)$$

Since  $\mathbf{r}_{AV} = \mathbf{r}_A + \mathbf{r}_V$  for optimal cue integration, applying the inner product with  $\mathbf{a}$  gives (from Eq. (6)):

$$\frac{1}{\sigma_{AV}^2} = \frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2}, \quad (8)$$

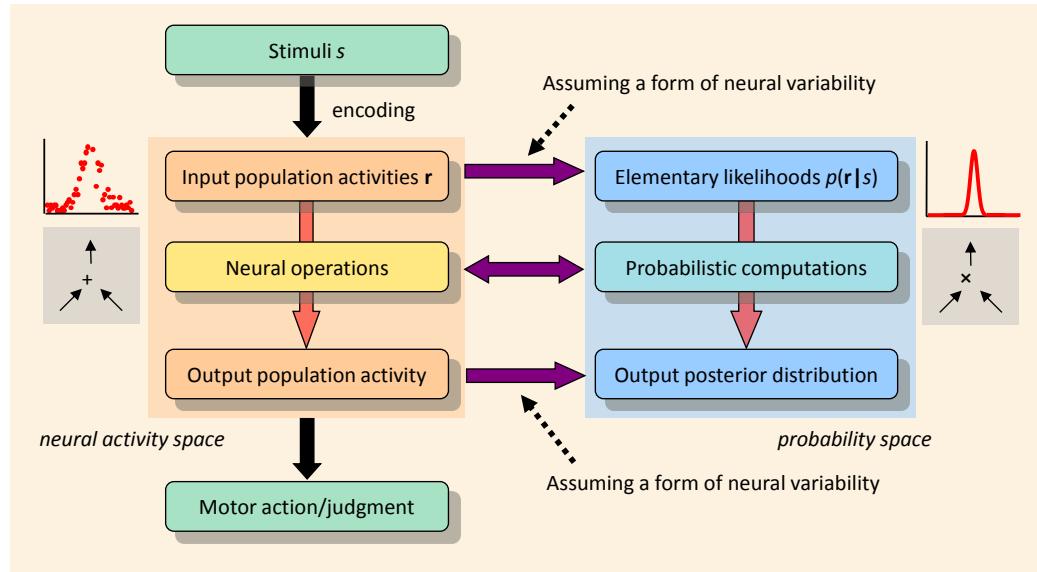
which is the *single-trial* version of our well-known equation for optimal combination of variances. For the mean, Eq. (7) gives

$$\frac{\mu_{AV}}{\sigma_{AV}^2} = \frac{\mu_A}{\sigma_A^2} + \frac{\mu_V}{\sigma_V^2} \quad (9)$$

(the extra assumption here is that the trial-to-trial fluctuations in the inverse variance are small). This is the *single-trial* version of the optimal combination of means we encountered in the cue combination chapter.

## 11.6 The neural implementation of Bayesian inference

We have laid out a theoretical framework for how optimal cue integration could be implemented by neural populations. The main significance of this framework does not merely lie in understanding multisensory perception in a principled manner, but in the fact that it provides a blueprint for finding neural implementations of other forms of Bayes-optimal computation, as discussed in Chapters 2 to 8. Probabilistic population coding provides a roadmap for identifying a neural implementation of each of these computations: first work out the Bayesian model at the behavioral level, then assume that probability distributions in this model are encoded in neural populations with Poisson-like variability, and finally identify the neural operations that map onto the desired operations on probability distributions. This general scheme for neural computation is illustrated in Figure 13.3. It is useful because it produces a natural bridge between the computational layer in which Bayesian models are usually formulated and the mechanistic layer in which models of brain function are often formulated.



**Figure 13.3:** Schematic of perceptual computation using probabilistic population codes. One or multiple stimuli elicit population patterns of activity. Each pattern encodes a likelihood function over a stimulus. In perceptual tasks, these elementary likelihood functions have to be manipulated in specific ways to achieve optimality (e.g. multiplication in cue integration). The key problem is to establish a “dictionary” between such probabilistic computations (e.g. multiplication) and neural operations on population patterns of activity (e.g. addition), assuming a form of neural variability (e.g. Poisson-like). Using those neural operations, the brain will retain full probabilistic information about the variable(s) of interest at all stages of computation. Eventually, a motor action is generated or a high-level judgment is made. From Ma, Beck, and Pouget (2008).

## 11.7 Alternative neural encodings of uncertainty

In the text so far we have seen one hypothesis about the way neurons may represent uncertainty. If their variability is only Poisson and otherwise the tuning curves relate to the stimuli, then we can (under certain assumptions) calculate the neural likelihood function in an analytical way. If we assume probabilistic population codes (PPCs), which to us means that the form of noise is the same in the pre- and the postsynaptic neural population then we can directly solve problems like cue combination in a neural way. However, there are many other ways of encoding uncertainty and for different ways of encoding uncertainty different neural likelihood functions emerge. Importantly, for each proposed way how neurons might encode uncertainty, there are certain computations that are very natural and certain neural phenomena that seem supportive, and other computations and phenomena that are harder to explain.

Importantly, different theories for the encoding of uncertainty are generally not mutually exclusive. It is possible that some brain areas represent probability distributions in one way and that other brain areas represent it differently. It is also possible that distinct neurons use different codes. Lastly, it is possible that neurons generally use mixtures of all these ideas.

Here we will give an overview of some popular coding schemes.

### 11.7.1 Probabilistic population codes

The probabilistic population codes are particularly well adapted at modeling cue combination, which just amounts to a single addition. Over time a broad range of problems have been formulated in this framework including xxx, xxx and xxx. They are compatible with certain biological findings. For example, lowering the contrast of visual stimuli lowers neural activities and increases behavioral uncertainty, as predicted. At the same time, there are other findings that are harder to explain. For example, PPCs predict that higher uncertainty generally leads to lower firing rate but many papers in the fMRI literature show higher firing rates. There are also other computations that do not map easily onto PPCs, including cases where uncertainty is supplied by other cues, or cases where the system needs to marginalize over many irrelevant variables.

### 11.7.2 Sampling codes

Another popular coding scheme, sampling proposes that the brain represents a probability distribution over the stimulus by drawing samples from this distribution over time. At any one point the brain would represent one possible world state  $s$  but would visit probable world states particularly often. More precisely, the probability that a neural representation that is associated with  $s$ ,  $r_s$  is encountered is equal to the probability that the stimulus is actually  $s$ .

$$p(\hat{s}(\mathbf{r}_s) = s) = p(s)$$

If our belief distribution over  $s$  is narrow, then the different samples of  $r$  will be similar to one another and if it is more broad then the samples will be more dissimilar. Just as in the previous proposals it is not entirely clear how the system would actually sample from the posterior.

A broad range of problems have been formulated in this framework including dreaming, high level cognitive estimations and visual perception. They are compatible with certain biological findings. For example, neural activities in the visual system of dreaming animals matches those of animals that watch movies of natural scenes, as predicted. At the same time, there are other findings that are harder to explain. For example, sampling codes predict that average activity should not be affected by uncertainty, which conflicts both with contrast data in the visual system and fMRI data in higher level cortex. There are also other computations that do not map easily onto sampling codes, including cases where speed is crucial.

### 11.7.3 Explicit probability codes

If we have variables that are binary, or we have a binary representation of a continuous distribution, then we could directly represent the probability distribution. Horace Barlow, in a seminal article (cite) has argued that in such a scheme it would make a lot of sense for neurons to exhibit a firing rate that is proportional to the probability of a feature being there:

$$p(s) = \exp(\kappa r)$$

here,  $K$  is a scaling factor determining average firing rates.

Alternatively, one might want to have neurons directly represent probabilities

$$p(s) = \kappa r$$

where again  $\kappa$  is a scaling factor determining average firing rates. Again, this coding scheme somehow assumes that the nervous system can calculate the posterior probability distribution and it is unclear how it will do so.

Appealing about these views is that indeed, in multiple domains including eye-movements, hand movements and others, the activity of neurons representing an action appear to change parametrically with the probability of that action.

A broad range of problems have been formulated in this framework including XXX. They are compatible with certain biological findings. For example, XXX, as predicted. At the same time, there are other findings that are harder to explain. For example, explicit codes predict that XXX but experiments show XXX. There are also other computations that do not map easily onto explicit codes, such as the representation of continuous probability distributions.

#### 11.7.4 Convolution codes

Convolution codes consider the activity of a neuron as a vote for a particular function associated with that neuron. Different probabilistic codes are compared in the table.

Code	Encoding	Decoding
Explicit probability code	$\langle r_i \rangle \propto p(s=s_i) + \text{constant}$	Unclear; requires prior over distributions $p(s)$
Log probability code	$r_i = [a \log p(s=s_i) + b]_+$	Winner-take-all
Log likelihood ratio code	$r_i = \left[ a \log \frac{p(s=s_1)}{p(s=s_2)} + b \right]_+$	Winner-take-all; limited to binary variables $s$
Convolution code	$r_i \propto \int \varphi_i(s) p(s) ds$	$\hat{p}(s) = \frac{\sum_i r_i \psi_i(s)}{\sum_i r_i}$
Probabilistic population code	observed variability, $p(\mathbf{r} s)$	$p(s \mathbf{r}) \propto p(\mathbf{r} s) p(s)$

A probabilistic population code with Poisson-like variability is related to other probabilistic codes. Applying Bayes' rule and assuming a flat prior, we find  $\log p(s|\mathbf{r}) = \mathbf{h}(s) \cdot \mathbf{r} + \tilde{\Phi}(\mathbf{r})$ , where  $\tilde{\Phi}(\mathbf{r})$  is such that  $p(s|\mathbf{r})$  integrates to 1. This probabilistic population code is therefore similar to a convolution code, except for the logarithm. The kernel  $\mathbf{h}(s)$  is specified by the statistics of the population and can be estimated through logistic regression. The log probability code is a special case, for

which the components of the kernel are delta functions,  $h_i(s) = \delta(s - s_i)$ . This kernel would, however, not be compatible with tuning curves and covariance structures found in real neurons.

### 11.7.5 Encoding moments of a distribution

Finally, one could imagine a code in which mean, variance, and perhaps if needed higher moments such as skewness and kurtosis of a distribution are each encoded by a neuron or a group of neurons. For example, distinct neurons could encode the mean and the variance of a distribution. One attractive way of representing probability distribution over stimuli is to simply have one representation that encodes uncertainty and another representation that encodes the best estimate of the mean. The mean of the activities of one group of neurons could encode the precision  $1/\sigma_s^2$  of the posterior  $p(s|r)$ . Another group of neurons could encode the estimate of the mean  $\mu_s$  of the posterior. For example we could have

$$r_i = \kappa_i / \sigma_s^2$$

with a neuron-specific gain factor  $\kappa_i$  for the variance encoding neurons and

$$r_j = \kappa_j \mu_s$$

with a neuron-specific gain factor  $\kappa_j$  for the mean encoding neurons. As opposed to PPCs and sampling codes, in this coding scheme the nervous system actually calculates the mean and the variance of the resulting posterior distribution.

This scheme has clearly a number of attractive properties. Mean and standard deviation are separated. As such, brain processes required to calculate learning rates or weights, calculations that only require uncertainty but not means would only need inputs from the uncertainty neurons. Readouts of means would only require inputs from the mean neurons.

There is some evidence for neurons that separately encode uncertainty. A wide range of electrophysiological studies have found that some neurons in the basal ganglia, the basal forebrain, and the amygdala appear to encode probability of events that are associated with rewards. Other neurons in sensory cortices appear to be well suited for a readout of means.

They are compatible with certain biological findings. For example, dopaminergic neurons in the ventral tegmental areas appear differentially affected by uncertainty, as predicted. At the same time, there are other findings that are harder to explain. For example, moment encoding codes predict that extra neurons that represent uncertainty are needed and, certainly in low level cortices, few neurons have been characterized with these properties. There are also other computations that do not map easily onto PPCs, including the problem that uncertainty is, in principle, associated with any existing variable which would necessitate a very large number of moment encoding neurons.

**Nature or nurture: where does Bayesian behavior come from?** The coding schemes discussed in this chapter ask how neurons may represent and calculate with uncertainty. These models generally assume that there is a specific *code* for uncertainty. However, there is a whole continuum of models that do not quite have that flavor.

On one end of the continuum do we have models where every spike is a sample from a probability distribution and where the brain is wired so that it has no choice but being Bayesian. On the other end of this continuum do we have models where the brain is good at learning, but there is nothing Bayesian about it – it's just that learning eventually leads to Bayesian behavior. Along this continuum there are many intermediate models. For example, we may have a model that has connections that are such that learning to be Bayesian is easy.

Experiments so far have generally analyzed highly overtrained behaviors like hand movements and depth estimation. As such, behavior is generally only analyzed in cases where the brain had ample experience for it to learn to approximate Bayesian behavior. So if the brain was not wired to be Bayesian but trained to be Bayesian we would not know.

This distinction between prewired codes versus learned codes is highly important for the resulting neurophysiological predictions. If we have a system (e.g. using deep learning) just learns to be Bayesian then the representation of uncertainty could be completely heterogeneous, with say some neurons coding for variance, and many others being modulated in distinct ways by uncertainty. If on the other hand there is a simple pre-wired code for uncertainty, then we have the chance of discovering it and considerably simplifying neuroscience going forward.

## 11.8 Problems

10.1 Give one example for each of the coding schemes where it seems ill placed.

**10.7** In Section 11.1, we sketched how to use the neural likelihood function to do subsequent inference. Apply this process to a discrimination task from Chapter 4, left/right classification of orientation, in the context of the toy model. Express the MAP decision rule in terms of  $r$ .

# Table of Contents

12	Outlook and limitations .....	2
12.1	Misconceptions about the state of the art .....	2
12.1.1	It has been shown that many behaviors are optimal .....	2
12.1.2	The brain could know nothing about probabilities .....	3
12.1.3	Its been shown that the brain's circuitry works following Bayesian mechanisms .....	3
12.1.4	The brain is messy, so its impossible that it follows Bayesian mechanisms .....	3
12.1.5	Bayesian models can be successfully applied to any problem .....	3
12.1.6	Bayesian models can only be applied to trivial problems .....	4
12.2	Misconceptions about the role of Bayesian models .....	4
12.2.1	Bayesian models compete with mechanistic models .....	4
12.2.2	Being optimal requires computing with probability distributions .....	4
12.2.3	Bayesian models can not inform mechanistic models .....	5
12.2.4	Bayesian models can explain anything and can hence not be falsified .....	5
12.3	Limitations .....	6
12.4	Limitations: There are suboptimal behaviors .....	6
12.4.1	Suboptimal inference: wrong generative model .....	6
12.4.2	Suboptimal probability estimates: base rate neglect .....	7
12.4.3	Suboptimal probability estimates: biased probability estimates .....	7
12.4.4	Suboptimal cue combination: conjunction fallacy .....	7
12.4.5	Suboptimal choices: matching .....	8
12.4.6	Suboptimal choices: framing effects .....	8
12.4.7	Suboptimal choices: loss aversion .....	8
12.4.8	Suboptimal choices: anchoring .....	9
12.4.9	Suboptimal behavior when the task is really difficult .....	9
12.4.10	High inter-subject variability in cognition .....	10
12.4.11	Bad economic behavior vs efficient perceptual and sensorimotor behavior .....	10
12.4.12	Bayesian models of "suboptimal" behavior .....	10
12.5	Limitations: There are mathematical challenges .....	11
12.5.1	Modeling higher complexity situations .....	11
12.5.2	Hierarchical structured models .....	11
12.5.3	Generalizing statistics from past events .....	12
12.5.4	Better approximate inference methods .....	12
12.6	Limitations: Combination with viable mechanistic models .....	12
12.6.1	The impossibility of many mechanistic models .....	12
12.6.2	Hybrid models – introducing Bayesian ideas into mechanistic models .....	13
12.6.3	Hybrid models – introducing mechanistic data into Bayesian models .....	13
12.7	Limitations: There are experimental challenges .....	13
12.7.1	More complex behaviors .....	13
12.7.2	Interaction with cognitive factors .....	13
12.7.3	Uncertainty in neural recordings .....	13
12.8	Outlook .....	14

## 12 Outlook and limitations

In this book we have introduced the tools, concepts, and questions associated with the Bayesian modeling of perception and action. To complete the exposition there are three remaining issues that need attention. There are a range of *misconceptions* held by many scientists building Bayesian models. We want to address the most common ones. There are *limitations*. Like most mathematical modeling, Bayesian modeling is useful for understanding some aspects of the world, and less useful for understanding other aspects. We will discuss roughly the scope of where Bayesian models are particularly useful. Lastly, we want to discuss our *outlook* on the future of the field. There are big questions to be asked and difficult problems to be solved. There are three natural classes of misconceptions about Bayesian approaches. Here, we will discuss (a) Misconceptions about the state of the field, about the things that have or have not been shown. A broad reading of the literature is necessary to deal with these misconceptions. (b) Misconceptions about the meaning of the results of Bayesian models. Dealing with these issues requires extensive discussion and to some level soul searching. Some issues boil down to the philosophical stance taken by the reader and thus remain open to disagreement.

### 12.1 Misconceptions about the state of the art

Another class of misconceptions are about the state of the art of the Bayesian modeling of perception field. There are both sentiments that the field is less and that the field is more advanced than it really is. A balanced view about the state of the art is important, both in evaluating past and future contributions.

#### 12.1.1 It has been shown that many behaviors are optimal

A good numbers of papers have reported that they found human behavior to be optimal. In the best cases, this comes from a statistical statement: the null-hypothesis that the data is indistinguishable from the model predictions cannot be rejected based on the data at hand. However, statements of optimality almost universally overstate their case. (A) Many of the papers do not even test if the null-hypothesis of being identical to the optimal model can be rejected. In fact, most papers that we are aware of do not even test this. Instead most models compare a Bayesian model with a non-Bayesian model and find that the Bayesian model does better. This in no way establishes that behavior is optimal. (B) In some cases the null-hypothesis can, indeed, not be rejected. However, even in this case, the argument for human optimality is unconvincing. Virtually all studies are underpowered to see even relatively strong deviations from optimality. This problem is compounded by the fact that the models themselves have a good number of free parameters.

So, what should we believe about the state of uncertainty about the brain? As discussed at several places in the book, we should probably believe that humans are pretty good at the things they do frequently (and have been doing on evolutionary timescales). Given the many noise-sources that appear to exist in the nervous system we might not expect that the brain is actually optimal.

### **12.1.2 The brain could know nothing about probabilities**

The opposite argument is also frequently formulated. Many scientists that do not use Bayesian models believe that the brain does not care about probabilities at all. However, there are many experiments that use the following logic. Design two experimental situations that differ in the level of associated prior or likelihood uncertainty. Keep everything else as constant as possible between the two situations. These papers have near universally found that behavior is different between the two situations. In other words, it seems outlandish at the moment to argue that the brain does not know anything about probabilities.

### **12.1.3 Its been shown that the brain's circuitry works following Bayesian mechanisms**

Many theoretical papers during the last decade or so have asked how the nervous system might represent uncertainty. These approaches typically propose that Bayesian statistics is basically built into the microscopic code of the nervous system. In other words, we are born to be Bayesian. What is appealing about this view is that it naturally would explain why we are able to deal with uncertainty. Some of the schemes (see chapter 11) are explored better than others, but all of them have alternatives. Moreover, while each of these encoding schemes have physiological data that supports the notion, each also has plenty of physiological data that it does not explain. As such, it is premature to assume that the brain is Bayesian at a microscopic level. One radical alternative to there is just strong learning. We know that the brain is good at learning. We know that being Bayesian is important in certain domains of life (e.g. cue combination). As such, we should expect the brain to be able to act as if it was Bayesian. Only real data can help ask to which level statistics is built into the makeup of the brain.

### **12.1.4 The brain is messy, so its impossible that it follows Bayesian mechanisms**

The opposite argument is also often made. Experimentalists state that they have extensively looked at the neural code, have not seen anything that resembles Bayesian statistics and hence the idea that the brain is Bayesian at the microscopic level must be wrong. They describe the huge complexity of the brain, different cells, different firing patterns, etc. Because the brain is thus “messy” it cannot cleanly represent a Bayesian machine. This view is highly problematic for two reasons. (A) There is actually little electrophysiological data where uncertainty was varied while other parameters were held constant. As such, even if the code for uncertainty would be simple, chances are that no one would have seen it. (B) Many if not most of the proposed coding schemes for uncertainty actually predict that it is hard to see the reflection of the encoding of uncertainty. It thus seems that the data just is not there to make this kind of an argument.

Also, in reality this whole question is about a continuum. On one end we have a brain that basically has Bayes rule built into its synapses. On the other end we have a brain that has nothing Bayesian about it but just learns to behave like a Bayesian system. However, anything on this continuum is possible. Some layouts of the brain may make it easier to learn to be Bayesian. Some might make it easier to represent priors (which requires memory). Where we are on this continuum requires careful experiments. The a-priori rejections that some scientists give to either view is unwarranted.

### **12.1.5 Bayesian models can be successfully applied to any problem**

In theory, for any problem that we can formulate we can write down the Bayesian solution. However, in practice this is not true. For example, for the problem of vision we can relatively easily formulate a good generative model. Only that we are entirely unable to actually do

inference on such real world problems. There is some promising research (e.g. in the work of Yuille, Sudderth, replace with citations?) but all Bayesian models are still a long distance away from actually being applicable to the real world problems that we would like them to be applicable to.

### 12.1.6 Bayesian models can only be applied to trivial problems

Again, the opposite assumption is just as widely held and equally wrong. Bayesian models are routinely used quite successfully in many technical domains, including speech recognition, data mining, and collaborative filtering. Bringing Bayesian models to progressively more interesting problems is still one of the exciting directions in today's research.

## 12.2 Misconceptions about the role of Bayesian models

Yet another set of misconceptions that are widely held are of a philosophical nature. What are Bayesian models about? What does it mean for them to be successful? What can we learn from them? And what can we not learn from them? In this section we will deal with these kinds of questions.

### 12.2.1 Bayesian models compete with mechanistic models

Many scientists see Bayesian models as competing with mechanistic models. If there is a mechanistic interpretation of the data, do we need Bayesian models? However, both approaches aim at answering entirely different questions. The Bayesian approach asks why behaviors are the way they are, in the sense that it asks why this is the right way of dealing with the available information. The mechanistic approach asks why behaviors are the way they are, in the sense that it asks which mechanistic factors give rise to the observed behavior. These questions are fundamentally different. For the same behavior we can have a Bayesian explanation (e.g. this happens because it minimizes variance) and a mechanistic explanation (e.g. this happens by synapses that add this and that firing rate). Both are answers to different questions and do not compete at all.

### 12.2.2 Being optimal requires computing with probability distributions

It is a common misconception that being optimal means computing with probability distributions (cite? Ma, WJ (2010). *Signal detection theory, uncertainty, and Poisson-like population codes*. Vision Research 50: 2308-2319., or somewhere else? Had to move it for balance.). In fact, not all optimal computation is probabilistic, and not all probabilistic computation is optimal. Being optimal with respect to the (0,1) cost function (see Section **Error! Reference source not found.**) simply means using the MAP estimate. The MAP estimate might or might not require knowledge of the input likelihood function(s). For example, we discussed in Section **Error! Reference source not found.** that when a Gaussian stimulus distribution is very wide, the MAP estimate,

$$\hat{s}_{\text{MAP}} = \frac{\frac{x}{\sigma^2} + \frac{\mu}{\sigma_s^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}}, \text{ reduces to } \hat{s}_{\text{MAP}} = x. \text{ Then, the width of the input likelihood function, } \sigma, \text{ does}$$

not appear in the expression for the MAP estimate. Estimation that requires knowledge of the width of likelihood function(s) is called *probabilistic*. A probabilistic estimate does not even

have to be a MAP estimate. For example,  $\hat{s} = \frac{\frac{x}{\sigma} + \frac{\mu}{\sigma_p}}{\frac{1}{\sigma} + \frac{1}{\sigma_p}}$  would be a probabilistic, but suboptimal estimator. Therefore, probabilistic and optimal computations are distinct concepts.

Even probabilistic estimators do not always require that the brain encode the full input likelihood function(s) on a trial-by-trial basis. It can for instance happen that because the likelihood width,  $\sigma$ , is the same over many trials, observers learn its value through feedback. To avoid this, one could induce different values of  $\sigma$  by varying the reliability of the stimulus, and/or by withholding feedback. This is relevant if the goal of the experiment is to provide evidence for the notion that neurons encode and compute with probabilities on a trial-by-trial basis.

### 12.2.3 Bayesian models can not inform mechanistic models

A good number of scientists state that they are exclusively interested in a mechanistic understanding of the nervous system and, therefore, unless the mechanisms themselves are Bayesian that Bayesian models are not interesting to them. However, in our views this ignores a good number of the contributions that Bayesian models can. To start with, Bayesian models may help us understand which variables are important for a given estimation problem. There are countless variables in the world, and some are more important than others. Bayesian models allow us to understand which variables are important. This understanding can help guide mechanistic experiments. For example, the role of uncertainty has long been ignored in electrophysiological experiments, a situation that is only slowly improving. Moreover, Bayesian models allow us to come up with a characterization of what is necessary for good solutions. Basically a mechanistic model can not be correct if it would lead to worse performance than actual humans. Lastly, Bayesian models can generate predictions of microscopic codes. These predictions may be right or wrong, but they can inspire experiments to test these hypotheses.

### 12.2.4 Bayesian models can explain anything and can hence not be falsified

It is true, that if priors and likelihoods can be freely chosen, then any behavior can be fit by a Bayesian model. However, the same is obviously true for any modeling framework. For example, any behavior could be explained by a mechanistic model as well. In general, any modeling framework can explain any kind of data. Falsifying a modeling framework is thus impossible.

However, any one given model can of course be falsified. Once we specify the generative model and the cost functions there is only one mapping from data to estimates and behaviors. As such, each individual model can be falsified. No one would ask if mechanistic models can be falsified, which does not make sense. The same is true for Bayesian models.

It deserves to be said, however, that within the Bayesian framework, not every set of assumptions about likelihood, prior and cost function is meaningful. The first two are assumed to

approximate reality, which can be tested with experiments that quantify properties of the world (cite?). Most cost functions would not make any sense, after all the cost function is assumed to approximate the true cost of the subject.

So far we have discussed the methods used in the Bayesian modeling of perception, exposed the data that has been gained, discussed the process of comparing data and models, discussed neural implementations and the misconceptions that scientists often hold. However, our book would not be complete with at least some discussion of the opportunities and challenges that lie ahead.

### 12.3 Limitations

After first discussing misperceptions we want to discuss limitations of the framework. There are clearly a wide range of tasks for which typical Bayesian models fall short, where they can not explain the main findings.

### 12.4 Limitations: There are suboptimal behaviors

We are all aware of decisions that did not seem all that good in hindsight, and in fact we frequently find ourselves making the same, wrong decisions. In fact this phenomenon is what Amos Tversky so nicely called “predictably irrational” in his delightful little book of that title (REFERENCE). In this section, we review how human subjects often differ from optimal and discuss both challenges and opportunities that arise from this.

There are several ways in which humans fail to behave optimally. A first group of biased behavior concerns the estimation of probability. Under certain circumstances, human subjects systematically tend to misestimate probability. Even when people have sufficient time to learn about a probability situation, they often make systematically suboptimal choices. There are a large number of cases where subjects show systematic mistakes in economic situations where they tend to utilize probabilities and utilities incorrectly. In addition, the brain has fundamental limitations that prevent it from optimally solving all decision tasks. Here we consider each of these suboptimal behaviors in turn.

#### 12.4.1 Suboptimal inference: wrong generative model

A very important and common way of performing suboptimal inference is to assume a wrong generative model. For example, an observer assumes that two cues are conditionally independent while in fact they are correlated. Or an observer believe the prior is a Gaussian distribution while in fact it has a different shape. Or an observer believes that the target is present with probability 0.5, while in fact it is with probability 0.43. In each of these cases, the false belief about the generative model will lead to suboptimal performance, even if the observer performs MAP estimation under the generative model they themselves assume. We saw an example of this in the cue combination chapter, where we studied an estimator that applied the wrong weights to the measurements. Many suboptimal estimators and decision rules can in fact be reformulated as

optimal under a wrong generative model, although in practice it is difficult to determine whether the observer assumed a particular wrong generative model. Changes in the environment often induce discrepancies between the true and assumed generative models. For example, if I come from a country where cars drive on the right of the road, but move to a country where they drive on the left, I might in my first days, when crossing the street, still first look to my left to see if a car is approaching. Wrong beliefs about generative models can be modified through learning, when feedback is given.

#### **12.4.2 Suboptimal probability estimates: base rate neglect**

Suppose a particular disease affects 1% of the population. Your doctor decides to run a test on you. This test will be positive 100% of the time if you have the disease and positive 5% of the time if you do not have the disease. The test comes up positive. What is the probability that you have the disease? Many people will estimate this probability to be about 95% (we hope that you won't, given that you have followed the book). The correct answer (found using Bayes' rule) is 17%, but subjects systematically overestimate this number. They appear to ignore the fact that the base rate of the disease is only 1%.

Many studies have found base rate neglect for medical decision making and social decision making. Base rate neglect is even a widespread problem in the medical profession, and it is estimated that many medical errors result from this problem. It turns out that asking the question in a different way helps people to avoid base rate neglect. We will often obtain reasonably accurate answers if we ask our doctor: "Out of the number of people who test positive overall, how many end up actually having the disease?"

#### **12.4.3 Suboptimal probability estimates: biased probability estimates**

Kahneman and Tversky have run a good number of experiments to ask if people are unbiased in their estimates of probability. For example, let's say we ask human subjects about the probability of dying from a wide range of diseases. Say Ebola (a very rare disease), stroke (a frequent cause of death), and the common cold. When asking many people about judgments of these probabilities, we would find systematic biases. Very rare events are estimated to be more likely than they are. Very frequent events are estimated to be less likely than they are. Roughly it seems that the estimated probability  $\pi$  relates to the real probability  $p$  as follows

$$\pi = p^\alpha / (p^\alpha + (1-p)^\alpha)^{1/\alpha}$$

It thus appears that people are systematically wrong at calculating probabilities. Hence, it would seem quite unlikely for them to be able to choose the right actions in the general case.

#### **12.4.4 Suboptimal cue combination: conjunction fallacy**

In a now famous experiment, Tversky and Kahneman (cite) asked subjects questions about scenarios of the following kind: "Linda is 31 years old, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice and also participated in anti-nuclear demonstrations." The research subjects were asked questions that assessed their estimates of both regular and joint probabilities. For example, they were asked "how likely is it that Linda is a bank teller?" and "how likely is it that Linda is a bank teller and active in the feminist movement?" When it comes to the probability calculus it is clear

that the second probably cannot be greater than the first: ( $p(A, B) = p(A)p(B | A) \leq p(A)$ ). Yet, the subjects on average responded to the second question with a greater probability estimate. This experiment, and many similar ones since, reveal systematic biases in subjects' probability judgments. What is interesting about these sets of experiments is the stark contrast to cue combination when it is given in a sensory context. Such sensory cue combination generally appears to be quite close to optimal.

#### 12.4.5 Suboptimal choices: matching

Lets say we are confronted with a two-armed bandit slot machine. We receive a reward randomly 70% percent of the time when we pull the left lever and 30% of the time when we pull the right one. As we saw in the previous chapter, it is clear that the optimal strategy (to maximize expected reward) would be to choose the left lever 100% of the time. Indeed, this is exactly what many animals such as rats do. Interestingly, human subjects and some other animals, e.g. monkeys and, surprisingly, pigeons, do not use that strategy. Instead they will pull the left lever 70% of the time and the right lever 30% of the time. Instead of choosing the best option, they match their response distribution to the reward distribution.

Interestingly, this choice of behavior, which persists even when there are real rewards involved, significantly lowers the overall rewards that subjects obtain. Despite being able to report that there are more rewards on one side, human subjects seem unable to act on their realization.

#### 12.4.6 Suboptimal choices: framing effects

In expected utility theory, subjects care about the outcomes, their probabilities and utilities. It does not matter how one reaches an outcome. It only matters what the outcomes are and how likely it is that we will find ourselves in each of them. However, research in behavioral economics has described a large number of cases in which people's behavior differs from that predicted by expected utility theory.

One famous deviation from optimal utility calculations is the *endowment effect*. A group of economics students were asked to bid for mugs and it was recorded how much they generally would be willing to pay for those mugs. In another experiments a group of similar economics students were given the mugs and it was assessed for how much money they would be willing to sell the mug. Interestingly, the price for selling the mugs was far higher than the price for buying them. This implies that by virtue of owning an object, our estimate of its value increases. Monkeys show a similar effect. They tend to prefer juice over peanut butter. But when given the peanut butter, very few of them are willing to exchange it for juice. These examples show that utility estimates can be affected by history, not just by outcome.

#### 12.4.7 Suboptimal choices: loss aversion

Another interesting effect is *loss aversion*. It is known that utility functions have decreasing marginal gains. However, if someone is relatively rich (say they own \$1,000,000) then more money will still make them better off and less money less well off. However, utility for such small changes should be roughly linear with the amount of money gained or lost. And yet, the loss of a given amount of money is felt much more strongly (in the negative) than the gain of the same amount is felt (in the positive). Consequently, subjects are far more willing to work to avoid losing a given amount of money than they are willing to work to make the same amount.

This loss aversion is of the order of a factor 2: a lost dollar is as bad as two gained dollars are good.

The *loss aversion* effect can only really be understood if we assume *framing*. Instead of comparing the total utility between one situation and another, we evaluate a single transaction: we frame it. Hence, it seems as if the utility function evaluated the transaction, not our general situation in life. This is a general violation of the assumptions of expected utility maximization.

#### 12.4.8 Suboptimal choices: anchoring

Yet another class of effects relate to *anchoring*. If we do not know the value of a potential option, we use seemingly incorrect ways of estimating this value. For example, it is hard for us to know the value of a specific bottle of wine, say a Côtes du Rhône 1998. Ariely did an interesting experiment in which he asked students to write the last two digits of their social security number (SSN) (which is essentially random) onto a piece of paper. Next he asked them to bid on the bottle of wine. Interestingly, there was a very strong correlation between the students SSN and the price they were willing to pay for the wine. Somehow, the SSN worked as an anchor; the students compared the price of the wine to their SSN. When we are uncertain about the value of something, our way of evaluating it may be arbitrary.

There are many real world implications of anchoring effects. We all know of car salesmen who show us an overpriced piece of junk before introducing us to the car they really want to sell us. In comparison with that anchor car, the final car looks like a great deal. One well-documented case involves bread machines. In the early 1990s, the Williams Sonoma company came up with a bread-making machine for \$275. These machines were delivered to many stores, and were quite unpopular. Subsequently, the same company came up with another bread making machine that was essentially identical, just somewhat larger and 50% more expensive. All of the sudden the first machine sold exceptionally well. The larger machine had set a price point or anchor that buyers could compare to, and bread-making machines became quite popular.

#### 12.4.9 Suboptimal behavior when the task is really difficult

There are some cases where optimal behavior would actually be very difficult. For example, lets take the problem of chess. There exists a solution. If black and white play optimally, then either white would win every time, black would win every time (less likely) or there would be a draw (possibly most likely). An optimal chess player would thus never loose. Computers cannot actually solve this problem at the moment and neither can humans. However, in principle there is a solution. This example highlights that ultimately, optimal behavior must break down. We are not optimal when the task is too difficult.

Many problems that appear in Bayesian statistics and decision making are known to be computationally hard. For example, the general problem of optimal control is known to be NP hard. Inference in arbitrary nonlinear systems is NP hard. It would thus seem, that it is quite likely to be impossible for the brain to optimally solve such problems.

In this book we have only dealt with those problems that for one reason or another can be solved efficiently. In doing so, we have given the reader an overview of many fundamental methods but we have not touched much on approximate methods. However, when problems are sufficiently hard, the brain has to resort to approximate methods.

There are many approximate methods to solving Bayesian problems and our little book is to small to even list a representative number of them. Optimal control can be approximated by the more tractable hierarchical optimal control. Arbitrary probability distributions can be approximated by a large number of particles. Or by a parametric distribution. There are sampling methods that allow approximating arbitrary probability distributions relatively compactly. Importantly, many methods exist for dealing with approximate inference and we can recommend the interested reader to peruse the relevant literature (cite Bishop, MacKay etc)

### **NP hard problems**

Computer science generally focuses on the scaling of problems. For example, if I can solve a puzzle with 10 pieces in 10 seconds, how long would it take me to solve the same puzzle with 20 pieces. There are some problems that are considered easy. For example, sorting. If I need to sort twice the number of items it only takes me a bit more than twice the time to sort them. More specifically it takes me of the order of  $N \log N$  to sort them. There are other problems, such as optimal control, for which we take an exponential time, i.e.  $e^N$ . These problems are called NP hard. It is known that von Neumann machines can not solve these problems efficiently.

#### **12.4.10 High inter-subject variability in cognition**

Typical studies in vision research run 3 subjects, observe that the three of them are almost indistinguishable and publish papers based on  $N=3$ . It is expected that all subjects behave in the same way. Typical studies in sensorimotor integration run 7 subjects, find that they are quite similar to one another. It is expected that the subjects behave in a way that is quite similar. Typical studies in economics or cognitive decision making run between 50 and 100 subjects, find that they are generally all quite different from one another but that across such a large population effects are solid. Somehow there is something about cognition that makes it incredibly variable, and quite a bit unpredictable. Understanding this variability and its structure is one of the central questions in cognitive science.

#### **12.4.11 Bad economic behavior vs efficient perceptual and sensorimotor behavior**

In the early chapters of the book we have seen a broad range of settings for which behaviors seemed to be quite good, often close to optimal. Now, given the violations of optimal decision-making we have listed so far, it seems that in many ways our decision making in economic situations is quite bad. And yet, the book thus far has focused on how the brain solves problems in a near-optimal way. Why is there a disparity between near-optimal perceptual decision making and badly suboptimal decision making in the economics realm? A recent study from the Maloney lab has started to ask this question. They created a sensorimotor task that was identical in structure to an economics task. They found that in the sensorimotor task subjects were near-optimal, whereas in the economics task they were not.

#### **12.4.12 Bayesian models of “suboptimal” behavior**

There are two interpretations of the deviations of optimality that we have discussed above. Either *behavior is actually suboptimal*, possibly due to limitations of the available processing power

(e.g. in the example of chess), or due to the algorithm implemented in the brain. Alternatively *behavior may be optimal, albeit for a different task*. For example, it has been argued that probability matching may be an optimal strategy. If we are in a completion situation with others then matching makes us less predictable, which in turn may optimize our ultimate yield. While there is not actually an adversary in the two-armed bandit problem we may still be using mechanisms that were evolved to deal with an adversarial situation.

One potentially important approach in this context are meta-bayesian methods, methods in which the nervous system could learn to be Bayesian (as in the work of Tenenbaum, Griffiths, etc, cite?). Such systems basically estimate the model about the world based on data, not just the parameters of an existing model. Such approaches often promise to explain deviations from optimal behavior.

## 12.5 Limitations: There are mathematical challenges

Bayesian models are clearly attractive to model both behavior and solve many technical problems. However, it is currently hard to formulate and solve them for situations that are complex and thus closer to those of everyday life. In this section we will discuss exciting developments in Bayesian statistics that promise to improve the way we deal with these models.

### 12.5.1 Modeling higher complexity situations

The models exposed in this book, and in fact virtually all Bayesian models that have been used to describe human perception assume pretty simple generative models. In fact, most experiments that have been performed estimate one dimensional variables and assume Gaussian distributions. For example, most cue combination studies involve a single, one-dimensional stimulus variable and two measurements. However, the problems that we solve in the real world are far more complex. We do not combine cues, we make sense of a cluttered room full of things. We do not use a prior over a one-dimensional variable, we use a complicated high-dimensional model of what our spouse may think. The world is not structured into a small number of perfectly repeatable events over which we can calculate statistics, in fact we never experience the same situation twice. However, we are currently missing the mathematical tools to deal with such complicated situations.

### 12.5.2 Hierarchical structured models

One highly exciting approach is the construction of complex hierarchical generative models for data where inference is still possible. For example, in work of sunderth et al, he defines a generative model for visual scenes. Each scene consists of a number of objects (drawn from a scene specific distribution). Each object consists of subparts (drawn from an object specific distribution), with some spatial relations to one another. And each subpart is somehow visually observed in images (through the use of so-called SIFT filters). Interestingly, when formulated carefully, inference can be quite efficient in this model. Similar models by Yuille allow the combination of text, faces, and textures. In a completely orthogonal approach Geoffrey Hinton is modeling hierarchical generative models of visual objects using new neural network training procedures (cite). The development of hierarchical models for complicated domains where efficient inference is possible is one of the exciting current developments in Bayesian statistics.

### 12.5.3 Generalizing statistics from past events

In statistics papers, we generally deal with very simple generative models. These models allow drawing arbitrary amounts of statistical data. In the real world, however, we never encounter the same situation twice. This means that we can not ever calculate statistics in the sense of how often an event has happened; after all each event only happens once. Instead we realistically need to estimate the probability of events based on other, similar events, that we have observed. This thus requires generalization of statistical information from past situations to the current situation – somewhat akin to the problem generally encountered in machine learning.

### 12.5.4 Better approximate inference methods

Ultimately, as you will have seen over the course of this book, doing Bayesian statistics is simple, apart from the fact that there are these ugly integrals that we need to solve all the time. Recent research has given rise to a large number of solution strategies that allow solving Bayesian problems of progressively larger size. Several approaches are notable in that regard. Markov Chain Monte Carlo (MCMC) methods allow dealing with a large class of generative models and are quite efficient in practice (provided that the resulting high probability regions of parameter space are reasonably connected). Variational Bayes methods are quite good at describing probability distributions that have certain parametric form. Empirical Bayes methods are quite efficient for many datasets. There is a broad range of exciting developments happening at the moment and we can only refer our readers to the relevant literature.

## 12.6 Limitations: Combination with viable mechanistic models

Many scientists, both those working on Bayesian approaches and those not using Bayesian are interested in mechanistic models of the brain. If we knew how neurons interacted with one another, we should be able to calculate the effect of potential perturbations to the nervous system. Such a mechanistic understanding should allow a deep understanding of how the brain works at a microscopic level and also give rise to cures to many of the complex diseases affecting the nervous system. However, at the moment, there appears to be an unfortunate antagonism where scientists who work towards a mechanistic understanding reject Bayesian approaches and vice versa.

### 12.6.1 The impossibility of many mechanistic models

Many scientists following the Bayesian approach believe that it is too early to build mechanistic models. If we are honest, the amount of information that we have about the nervous system is quite limited. We know tuning curves in a good number of brain areas, but only for a relatively small number of neurons which are, to make matters worse, generally sampled in a biased way from the population. We know synaptic properties of some neurons, we know anatomical aspects about the brain and results from imaging. However, with all this, we can not even get close to being able to construct a model that contains all the synaptic connections, the right nonlinear processing, and the right biophysics. We can not even do such simulations for the tiniest organisms (e.g. *c. elegans*) and much less so for any larger animal. It seems that to simulate even a cubic millimeter of brain tissue would require millions if not billions of unknown parameters.

Of course, we might be lucky. Maybe these parameters do not matter. Maybe these parameters are very well conserved across groups of neurons and hence we do not need to measure them *in vivo*. However, evidence for these views is essentially nonexistent and given the importance of learning it seems very unlikely that the information stored in the brain is low-dimensional.

These, and other considerations consider many scientists working on Bayesian approaches that the time is not ripe for mechanistic models. However, this view seems to ignore the fact that the two approaches can be synergistic as we will discuss below.

### **12.6.2 Hybrid models – introducing Bayesian ideas into mechanistic models**

Bayesian models could be used to simplify the building of mechanistic models. For example, lets say I build a mechanistic model. There are some parameters that I can measure. There are other parameters that I can not measure. I could then use Bayesian models to fill in the parameters that I can not measure. Basically use those parameters that would make behavior closest to the Bayesian optimum.

### **12.6.3 Hybrid models – introducing mechanistic data into Bayesian models**

Chapter 12 of this book extensively deals with neural decoding. When building Bayesian models, one possibility would be to replace the theoretical encoding of probabilities into the  $\mathbf{r}$  vector with actual neural activities, maybe those recorded simultaneously in an experiment. This approach promises to make Bayesian models closer to the physiological reality implemented in the nervous system.

## **12.7 Limitations: There are experimental challenges**

Possibly the most important direction of development are new experiments. If we want to understand how the brain deals with uncertainty in progressively more complicated situations we need to find good, well controlled experiments to measure this dependency. If we want to understand how the nervous system encodes uncertainty we need better physiological experiments that characterize the dependency of neural activities on uncertainty.

### **12.7.1 More complex behaviors**

Most of the experiments that Bayesian models are currently being fitted to are of a very simple nature. A prior and a cue. Or two cues. Or one variable that evolves over time. We therefore have a relatively poor understanding how good Bayesian models are at explaining behaviors over the broader set of realistic behaviors. One approach that probably is promising in that direction is to start with simple problems and progressively make them more difficult. This would allow the development of more complicated Bayesian models and the measurement of human data from progressively more complicated behaviors to go in lockstep with one another.

### **12.7.2 Interaction with cognitive factors**

There are clear domains where aspects of cognitive models that are not caught by Bayesian models are important. Attention, memory, emotions all are bound to play important roles for actual human behavior. It is an exciting possibility to look for the interactions of such systems. How will Bayes-like behavior be affected by cognitive or attentive loads? How will it interact with emotional processes? To which level could those cognitive systems be included in a Bayesian treatment and to which level are they fundamentally different.

### **12.7.3 Uncertainty in neural recordings**

One last important area are experiments that ask how uncertainty is encoded in the nervous system. We have no dearth of models of how the brain may represent probability distributions and how it may compute with them. However, experimental neuroscience (maybe apart from the fMRI field and the reward processing field) has not to any degree of extent asked how uncertainty could be represented. The right way of doing such experiments would probably be to keep everything else constant while varying aspects of uncertainty, maybe likelihood or prior

uncertainty, or information about causality. One could then measure how neurons are affected by such changes of uncertainty. Such experiments promise to tell us how neurons represent uncertainty and to which level the same coding scheme is used throughout the nervous system.

## 12.8 Outlook

There are success stories of Bayesian statistics as applied to the study of perception and action that have made the framework a staple of behavioral and neural modeling. There are also numerous limitations that limit the scope where these models are useful. Particularly important are deviations from optimality, as well as mathematical and experimental challenges. Progress in the field has been rapid at dealing with all these issues. Bayesian statistics is really just calculus of probabilities and as such, it seems unthinkable that it is not going to have its place in future models and experiments. However, there are many challenges and opportunities that lie ahead.

## Contents

Appendix: Basics of probability theory .....	2
1 Objective and subjective probability .....	4
2 The intuitive notion of probability .....	5
3 Complementary event .....	5
4 Venn diagram representation .....	6
5 Random variables and their distributions .....	7
5.1 Discrete versus continuous random variables .....	7
5.2 Total probability = 1 .....	8
5.3 Discrete probability distributions .....	8
5.4 Continuous probability distributions .....	9
5.5 Formal definition of the probability density function (Advanced) .....	11
5.6 Normalization .....	12
5.7 A note on notation .....	12
6 Mean, variance, and expected value .....	13
7 The normal distribution .....	14
7.1 Definition .....	14
7.2 Central limit theorem .....	14
7.3 Multiplying two normal distributions .....	15
7.4 Multiplying N normal distributions .....	16
7.5 The error function .....	16
7.6 The Von Mises distribution .....	17
8 The delta distribution .....	18
9 The Poisson distribution .....	19
10 Distributions involving multiple variables .....	19
10.1 Joint probability .....	20
10.2 Marginalization .....	21
10.3 Conditional probability .....	22
10.4 Independence .....	25
10.5 Bayes' rule .....	25
11 Functions of random variables (Advanced) .....	27

11.1	Functions of one variable: changing variables .....	27
11.2	Marginalization formulation.....	31
11.3	Functions of multiple variables .....	33
12	Drawing from a probability distribution.....	35
12.1	Drawing from a uniform distribution .....	35
12.2	Drawing from a normal distribution.....	35
12.3	Drawing from a Poisson distribution.....	36
12.4	Drawing from other built-in distributions .....	36
12.5	Drawing from an arbitrary one-dimensional distribution.....	36
12.6	Drawing from a mixture of distributions.....	37
12.7	Proof of the Cramer-Rao bound (Advanced) .....	37
13	Problems .....	39

## Appendix: Basics of probability theory

This Appendix provides some basics of probability theory. It is by no means an exhaustive introduction as one would find in a textbook on probability theory. Instead, it is a tutorial only on the specific concepts and calculations we use in the book.

### Probability cheat sheet

#### General

Random variable  $X$

Can be discrete, e.g. die roll, coin toss, number of children, spike count, or continuous, e.g. person's height, elapsed time, membrane potential

We denote a value that  $X$  can take by  $x$ . This can be confusing, but think of  $X$  as a label and  $x$  as a number. It is slightly sloppy but quite common to use the same notation for both.

For a discrete variable, we use  $p_X(x)$  or, if there is no misunderstanding what  $X$  is,  $p(x)$ , to denote the probability that  $X$  takes the value  $x$ . This is called a probability mass function. For a continuous variable,  $p_X(x)$  is the probability that  $X$  takes a value in a small bin around  $x$ , divided by the size of the bin. This is called a probability density function.

Discrete variable  $\rightarrow$  pmf; continuous variable  $\rightarrow$  pdf. Pmfs can never take values  $> 1$ , pdfs can.

Normalization: total probability = 1

Discrete variable	Continuous variable
-------------------	---------------------

$\sum_x p(x) = 1$	$\int p(x) dx = 1$
-------------------	--------------------

Mean (also expected value) of  $X$ :

Discrete variable $\langle X \rangle = \sum_x x p(x)$	Continuous variable $\langle X \rangle = \int x p(x) dx$
--	---

Expected value of a function of  $X$ :

Discrete variable $\langle f(X) \rangle = \sum_x f(x) p(x)$	Continuous variable $\langle f(X) \rangle = \int f(x) p(x) dx$
--	---

Variance and standard deviation of  $X$  (both for discrete and continuous), and derived quantities:

$\text{Var } X = \langle (X - \langle X \rangle)^2 \rangle = \langle X^2 \rangle - \langle X \rangle^2$ $\text{Std } X = \sqrt{\text{Var } X}$ $\text{Coefficient of variation of } X = \frac{\text{Std } X}{\langle X \rangle}$ $\text{Fano factor of } X = \frac{\text{Var } X}{\langle X \rangle}$
--

### Binomial distribution

Choose a time window of interest. Divide it into  $n$  bins. The probability of getting a spike in any one bin (independently of the other bins) is  $\lambda$ . This is called a Bernoulli process. Then the probability of having a total of  $r$  spikes is

$$p(r) = \binom{n}{r} \lambda^r (1-\lambda)^{n-r}$$

Mean (in sloppy notation):  $\langle r \rangle = \lambda n$ . Variance:  $\text{Var } r = n\lambda(1-\lambda)$ . Fano factor =  $1-\lambda$ .

### Poisson distribution

Take a Bernoulli process. Choose the number of bins  $n$  very large and the probability of spiking per bin,  $\lambda$ , very small, while keeping  $\lambda n$  fixed. Then this is a Poisson process, and  $\lambda n$ , which we will denote by  $\lambda$ , is called the *rate parameter*. The probability of finding  $r$  spikes is

$$p(r) = \frac{1}{r!} \lambda^r e^{-\lambda}.$$

Mean:  $\langle r \rangle = \lambda$ . Variance:  $\text{Var } r = \lambda$ . Fano factor = 1.

### Exponential distribution

Take a Poisson process with rate parameter  $\lambda$  in a time interval  $[0, T]$ . Define the firing rate as

the expected number of spikes per unit of time:  $f = \frac{\lambda}{T}$ . Then the probability of finding an interspike interval  $t$  is

$$p(t_{\text{isi}}) = f e^{-ft_{\text{isi}}}$$

Mean:  $\langle t_{\text{isi}} \rangle = \frac{1}{f}$ . Variance:  $\text{Var } t_{\text{isi}} = \frac{1}{f^2}$ . Coefficient of variation = 1.

## 1 Objective and subjective probability

Probability is degree of possibility. In its most restrictive sense, probability can be defined as the expected outcome frequency of a repeatable event, such as the probability that a coin will come up heads or the probability that someone rolls a 5 on a die. These events can be repeated an arbitrarily large number of times, and the long-run outcome frequencies tallied. If the proportion of tosses on which a coin lands heads converges to 0.5 as the number of tosses approaches infinity, we can state that the coin has a 0.5 probability of landing heads. This type of probability is sometimes called *objective probability*, and it is the only valid type of probability according to a strict frequentist view of probability.

A much broader – and, we believe, much more useful – conceptualization of probability is as degree of belief in a possibility. This is called *subjective probability*. The everyday terms *confidence* and *uncertainty* refer to subjective probability. If I know that a die has 1/6 chance of landing 5, then my confidence in the proposition that it will land 5 is 1/6. This particular example is trivial, because it involves simply converting an objective probability (a long-run outcome frequency) into a belief statement. However, the vastly wider applicability of the subjective conceptualization becomes clear when we consider degrees of belief in outcomes that cannot be repeated, for example the probability that candidate A will beat candidate B in the next election. This is not a probability that can be obtained by repeating the same event many times, but we may nevertheless have a strong prediction regarding the outcome. Indeed, examples of subjective probabilities that cannot be phrased as long-run outcome frequencies abound in daily life: What is the probability that it will rain today? What is the probability that I will enjoy the course taught

by professor X? Many scientific questions also can be phrased only in terms of subjective and not in terms of objective probabilities: What is the probability that Saturn's mass lies between  $10^{25}$  and  $10^{26}$  kg? What is the probability that disease X is caused by a virus?

The concepts of objective and subjective probabilities are not always clearly distinguishable. For example, to determine the probability that it will rain today, a forecaster might run a large number of simulations starting from the current state of the atmosphere, each with a different instantiation of the stochastic factors in the model, and record the frequency of rain among these runs. While the resulting probability is subjective, it has been obtained in an "objective" way, namely by counting. Similarly, if I observe dark clouds in the sky and express my opinion that there is a high probability of rain, I am expressing a subjective probability judgment, but I am basing this judgment on a large number of previously observed, similar (though not identical) skies.

A great advantage of Bayesian inference is that it treats both subjective and objective probabilities in the same way. Bayesian inference is therefore extremely widely applicable. The same mathematical relationships (Bayes' rule, marginalization, etc.) apply identically to both types of probability. Bayesian models of perception, however, are grounded fundamentally in subjective probability: there is only one true world state, but from the point of view of an organism trying to infer it, there are many possibilities, and degrees of belief can be assigned to these possibilities.

## 2 The intuitive notion of probability

We call the set of all possibilities under consideration the *sample space*. An *event* or *hypothesis* is a subset of the sample space. The term "event" is commonly used when discussing objective probability, and the term "hypothesis" when discussing subjective probability. The sample space could be "all possible numbers I can roll on a die" or "all possible weather patterns that can occur today". Given the former sample space, an event could be "I will roll an even number". Given the second sample space, a hypothesis could be "It will rain today". The probability of an event or hypothesis is a real number between 0 and 1, indicating the degree of possibility of the event or hypothesis. An event that is certain has a probability of 1, and an impossible event has a probability of 0. For events, one can think of probability as the frequency that the event happens among a very large number of random samples from the sample space. For example, the probability that I will roll an even number on a die is  $3/6=0.5$ . As explained above, this can also be conceptualized as a degree of belief. For a hypothesis, the frequency concept does not generally apply but probability still represents a degree of belief; for example, the degree of belief that it will rain today could be 0.35. The probability of an event or hypothesis  $X$  is denoted  $\text{Pr}(X)$ . For example,  $\text{Pr}(\text{it will rain today})=0.35$ , or  $\text{Pr}(\text{coin comes up heads})=0.50$ .

## 3 Complementary event

Given an event or hypothesis, its complementary event or hypothesis is that the first event or hypothesis does not occur or is false. For example, the complementary event to "rolling a 1 on a

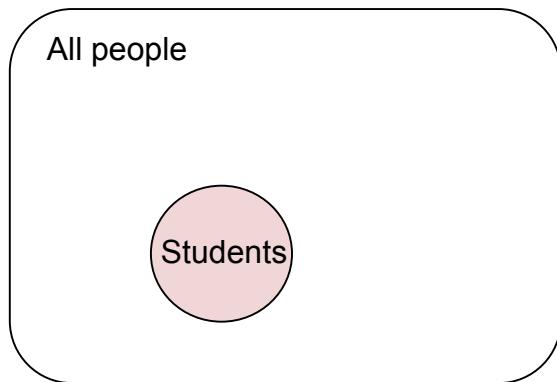
die" is "rolling a 2, 3, 4, 5, or 6 on a die". If the event or hypothesis is denoted  $X$ , then its complementary event or hypothesis, or complement, is denoted  $\neg X$  (read: "not  $X$ "). Here,  $\neg$  is the symbol for a logical negation. The probability of the complementary event or hypothesis is 1 minus the probability of the event or hypothesis:

$$\Pr(\neg X) = 1 - \Pr(X).$$

In some problems, it is easier to calculate the probability of the complement of an event or hypothesis than of an event or hypothesis itself. For example, if you are asked to calculate the probability that the sum of the eyes on two dice is at least 3, it is easiest to first calculate the probability that the sum is lower than 3, and subtract that from 1 (the answer is 35/36).

#### 4 Venn diagram representation

Events and hypotheses can be represented graphically through *Venn diagrams* or *Euler diagrams*. First draw a large rectangle whose interior represents all possible outcomes, i.e. the sample space. We set the area to 1, representing a total probability of 1. Then draw inside this rectangle a circle that represents all outcomes consistent with a particular event or hypothesis.



**Figure A1.** Example of a Venn diagram used to represent probabilities.

For example, the rectangle could represent all people in a group, and the circle all students among them. The area enclosed by the circle is a fraction of the area enclosed by the rectangle; this fraction represents the probability of the event or hypothesis – in our example, the probability that a randomly selected person is a student. The complement of the event or hypothesis is represented by the points that are inside the rectangle, but outside the circle. Its area divided by the area of the rectangle represents the probability that a randomly selected person in the group is not a student. We will make use of the Venn diagram representation in later sections.

## 5 Random variables and their distributions

A random variable is a variable whose values cannot be known with certainty. Examples include the number rolled on a die, the date of birth of a person, the shoe size of a random person on your street, the time it takes to travel from home to work, the number of voters who will participate in an upcoming election, or the price of a stock tomorrow. The opposite of a random variable is a variable whose value is known with certainty. Examples of non-random variables are the number of planets between us and the Sun (2), the number of days in a week (7), the ratio of the circumference to the diameter of a circle, the age of the oldest person on Earth, and the distance between two adjacent cm marks on a ruler (1 cm).

This is not an airtight distinction. Variables that appear non-random might be subject to measurement or production noise, which makes them random. For example, the distance between two adjacent cm marks on a ruler might vary, since the machine that produced the rulers was probably programmed by a computer that set the centimeter marks. However, computer-generated numbers have only a finite number of decimals, perhaps 10. As a consequence, the centimeter marks will never reach femtometer precision. In addition, the paint used for the marks will not attach itself to the surface in an identical way every time a ruler is produced. Therefore, one can think of the distance between two adjacent marks as a random variable. Similarly, we might think we know the age of the oldest person on Earth because we rely on the media or on the Guinness Book of Records; however, their sources of information might be flawed or incomplete. Perhaps in a distant part of the Amazon, a person has grown to be 130 years old without ever being documented as such. Therefore, certainty about this variable does not exist. For reasons like these, it might be useful to think of all variables as random, just with some having very low uncertainty.

Randomness, also called variability, noise, or stochasticity, is often a consequence of a lack of knowledge. If I rolled a die and you could record exactly the position, direction, and speed from which the die left my hand, and you were able to simulate exactly the interactions the die had with air and table, then you would be able to predict with certainty the outcome of every roll. Since nobody knows the values of all these variables, or would care to model them, the die roll is random or variable. Whether true randomness exists is a philosophical question that is beyond the scope of this book.

### 5.1 Discrete versus continuous random variables

Random variables can be distinguished based on the values they can take. The most important distinction is between discrete and continuous random variables.

A *discrete random variable* takes on a set of values that can be counted, even though there might be infinitely many. Examples are the number of children in a household, the number of dots we draw on a piece of paper, the number of action potentials fired by a neuron, in fact any “number of...” variable, the number of moves in a chess game, the age of a person when counted in whole years, the gender of a person, the price of a movie ticket, the set of ingredients

that go into a recipe, or the identity of a spoken word. A discrete random variable that takes on only two possible values is called *binary*.

A *continuous random variable* takes on values on a continuum. By definition, there are infinitely many values on a continuum. Examples are the length of a line segment, the color of a surface, the shape of a polygon, the direction one can walk on an open field, the amount of an ingredient in a recipe, the waiting time in front of a red light, and the frequency of a musical note. One can think of a continuous variable as discrete but with values that come in very small increments. For example, distance is a continuous variable, but when it is measured in whole millimeters, it is a discrete variable. In a computer program, truly continuous variables do not exist; their domain must always be discretized.

## 5.2 Total probability = 1

The total probability of all possible values of a random variable equals 1. This total is a reflection of the fact that the possibilities are mutually exclusive. If one were to increase the probability of one value, the probability of at least one other value has to decrease.

## 5.3 Discrete probability distributions

Discrete probability distributions are functions that assign a probability to each possible value of a discrete random variable. The probability distribution over a discrete random variable is also called a *probability mass function*. A discrete random variable  $X$  taking a particular value  $x$  is an event, denoted  $\Pr(X=x)$ . As we now vary  $x$  over all its possible values, we obtain a function of  $x$ . This is the probability mass function, denoted  $p_X(x)$ .

$$p_X(x) = \Pr(X=x)$$

(Throughout this Appendix, we will denote random variables by capitals and their values by lowercase letters. We do not sustain this convention throughout the book, though.) This means that the probability mass function evaluated at  $x$  is equal to the probability that the random variable takes this value. We use the subscript  $X$  (uppercase) to refer to a random variable, and the argument  $x$  (lowercase) to refer to a specific value of this random variable. The term “mass” is borrowed from physics. Roughly speaking, it is based on using matter as a metaphor for a possibility (a point in the sample space). The larger the probability of an event or hypothesis, the larger the mass of the piece of matter in the metaphor.

Example: The random variable  $X$  is the number rolled on a die. Its possible values  $x$  are 1 to 6. If the die is fair, the probability of each of these values is  $1/6$ , i.e.  $\Pr(X=x)=1/6$  for all  $x$ . This is an example of a discrete uniform distribution. If the die is not fair, the probability of at least two of the values differs from  $1/6$ , and the distribution will no longer be uniform.

For discrete random variables, total probability is computed by summing over all possible values; it should return 1. This is denoted as follows:

$$\sum_x p_X(x) = 1.$$

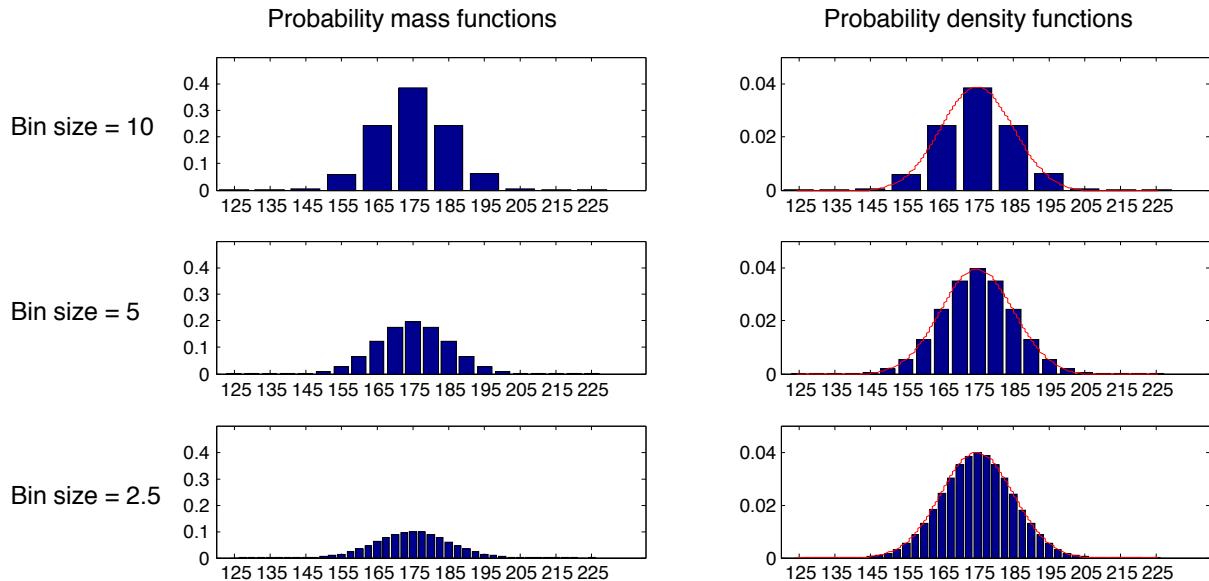
In a specific case where the possible values of  $X$  are given, we can put those above and below the  $\Sigma$  sign. For example, the total probability of a die roll would be

$$\sum_{x=1}^6 p_X(x) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 1$$

Binary random variables are a special case of discrete random variables. Suppose a binary random variable  $X$  can take values  $x_1$  and  $x_2$ . We know that total probability equals 1. Therefore, the probability of  $x_2$  is 1 minus the probability of  $x_1$ , i.e.,  $p_X(x_2) = 1 - p_X(x_1)$ .

#### 5.4 Continuous probability distributions

What is the probability that someone is exactly 160 cm tall? It is zero, since “exactly” means that the length is accurate to an infinite number of decimal places. This problem is characteristic of continuous random variables, and illustrates that probability mass functions, which worked well for discrete distributions, have to be replaced by a different concept in order to accommodate continuous variables.



**Figure A2.** When a random variable can take on a continuum of values (x-axis), a probability mass function is only defined when outcomes are binned. The values of the function will decrease with decreasing bin size (left column). Probability density functions are obtained by dividing the probability mass values by bin size. This yields values that are independent of the bin size. The process of making the bin size smaller can be continued until the bins are infinitesimally small. This produces a continuous probability density function, overlaid in red.

Suppose we are interested in the probability distribution of the height of an adult (Fig. A2). As a first approximation, we could consider possible heights in bins of 10 cm increments: between 120 and 130 cm, between 130 and 140 cm, etc. Each bin has an associated probability, and in this way we can build up a probability mass function. However, we might want to describe height more finely, say in bins of 5 cm: between 122.5 and 127.5 cm, etc. Each original bin is thus replaced by 2 new bins, each of which has on average one half the probability mass of the original bin. Thus, the new probability mass function is scaled to about half the height of the original one (Fig. A2). As we keep decreasing bin width in order to increase precision, the probability mass per bin keeps decreasing as well – it can become arbitrarily small. This is not very satisfactory. Is there a way to prevent the probability mass function from “disappearing”? Yes, this can be accomplished by dividing the probability mass in a bin by the width of the bin. By doing so, the function does not change much as we decrease bin width – it only becomes more precise. The *probability density function* is the result of this process as the bin width approaches zero (red curve). Again, there is an analogy with physics: if the probability in a bin is regarded as mass, then dividing this probability by bin width is analogous to computing a linear density: mass per unit length of the x-axis.

The similarity in notation between the probability mass function for discrete variables and the probability density function for continuous variables, both denoted  $p_X(x)$ , is misleading as there are some important conceptual differences between the two. For a discrete distribution, the probability mass of a single point never exceeds 1, since the probability mass values have to sum to 1. For a continuous distribution, the probability density at a single point is meaningless and can take arbitrarily large values. Consider, for example, a uniform distribution on the interval  $[0,0.01]$ . It will have a probability density of 100 at every point. Only the integral over an interval will always be less than or equal to 1. Stated differently, for a discrete distribution, the probability  $\Pr(X=x)$  is a meaningful number that can take any value between 0 and 1, and is in fact identical to  $p(x)$ . For a continuous distribution,  $\Pr(X=x)$  is always 0, and only probabilities of the form  $\Pr(a < X \leq b)$ , with  $a$  and  $b$  arbitrary numbers, are meaningful.

We use the terms *probability distribution function (pdf)*, *probability distribution*, or simply *distribution* to refer to the probability mass function of a discrete random variable or the probability density function of a continuous random variable.

Just as for discrete variables, the total probability of all values of a continuous variable equals 1. Total probability for a continuous variable is computed not as a sum, but as an *integral*. The integral of a continuous probability density function, as defined above, is the width of a bin multiplied by the function value in that bin, summed over all bins, in the limit that bin width approaches zero. Calculus provides recipes to compute integrals of certain functions. In this chapter, we familiarize ourselves with various integrals over probability density functions, especially because they directly parallel expressions with sums over probability mass functions;

however, we will not evaluate these integrals, so no calculus is needed. The rule of total probability for a continuous variable  $X$  is written as

$$\int p_X(x)dx = 1.$$

The “ $dx$ ” is in essence the width of a very small bin, and the integral sign  $\int$  is a deformed “S” for sum.

The most important continuous distribution is the normal distribution, which we discuss in detail below (Section 7). Another important one is the uniform distribution. The uniform distribution on an interval  $[a,b]$  has a constant pdf

$$p(x) = \frac{1}{b-a} \quad (\text{A.1})$$

The following continuous distributions are also common in applications of probability theory. The exponential distribution is given by  $p(x) = \lambda e^{-\lambda x}$ , with  $\lambda$  a constant and  $x$  defined on the positive real line. The power law distribution is given by  $p(x) \propto x^{-a}$ , with  $a$  a constant and  $x$  again defined on the positive real line. In this book, we occasionally use these distributions.

### 5.5 Formal definition of the probability density function (Advanced)

Consider a continuous random variable  $X$ , such as the waiting time in a queue. The probability that the value of  $X$  is less than or equal to  $x$  is denoted by  $\Pr(X \leq x)$ . This is the *cumulative distribution function* (cdf) of  $X$  at  $x$ , denoted  $P_X(x)$ :

$$P_X(x) = \Pr(X \leq x).$$

By definition, this is a monotonically increasing function that takes values between 0 (at  $x = -\infty$ ) and 1 (at  $x = \infty$ ). The *probability density function* (pdf) of  $X$  is now the derivative of this function:

$$p_X(x) = \frac{dP_X}{dx} \quad (\text{A.2})$$

We use uppercase letters to denote cumulative distribution functions, and lowercase letters to denote probability density and mass functions. Figure 3 shows an example of a cumulative distribution function and a probability density function. For discrete random variables, the cdf can be defined in the same way, but it is not a necessary step in defining the probability mass function.

To go back from the pdf to the cdf, one integrates:

$$P_X(x) = \Pr(X \leq x) = \int_{-\infty}^x p_X(t) dt.$$

The physics equivalent of this statement is that the integral over a density is a mass. It immediately follows that the probability that  $X$  takes values in an interval  $(x_1, x_2]$  can be obtained by integrating  $p_X(x)$  between  $x_1$  and  $x_2$ :

$$\Pr(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p_X(t) dt.$$

It also follows from the definition that  $\int_{-\infty}^{\infty} p_X(t) dt = 1$ .

### 5.6 Normalization

A function can be made into a probability distribution by dividing each value by the total value on the entire domain, provided that this total value is finite. As a result, the probability distribution will integrate (or sum) to 1. This process is called *normalization*. If the total value on the entire domain is infinite, normalization is not possible.

Exercise: Prove that the exponential distribution is normalized.

Exercise: Normalize the power law distribution and find a condition on  $a$  for which normalization is possible.

### 5.7 A note on notation

For discrete and continuous distributions alike, we use  $p$  to denote a probability distribution. However, in a given problem, we might consider multiple random variables, each with its own distribution. Therefore, the proper notation is to label  $p$  by the random variable considered. For example, for a continuous random variable  $X$ ,  $p_X(x)$  would be the probability density belonging to  $X$  evaluated at the value  $x$ . However, since this notation is somewhat cluttered, we often leave out the subscript when no confusion is possible. We consider that to be the case only when the argument (here  $x$ ) corresponds directly to the random variable (here  $X$ ). Whenever another numerical or symbolic value gets substituted, we make sure to attach the subscript. Thus, we will write  $p(x)$  instead of  $p_X(x)$ , but will not simplify  $p_X(3)$  or  $p_X(y^2)$ . Incidentally, we will also denote these as  $p(X=3)$  or  $p(X=y^2)$ . If you are not sure, include the subscript. When the subscript is left out, it is important to keep in mind that the same  $p$  can stand for different functions depending on the argument.

Special notation exists for events or true/false statements. An event is a binary random variable that can take values “occurs” (1) and “does not occur” (0). We often abbreviate the probability that an event  $X$  occurs,  $\Pr(X=1)$  or  $p_X(1)$ , as the probability of the event,  $p(X)$ . The

same holds for true/false statements, which are binary random variables that can take values “true” and “false”. For a true/false statement  $X$ , the notation  $p(X)$  indicates the probability that the statement is true, and  $p(\neg X)$  that it is false.

## 6 Mean, variance, and expected value

For a discrete random variable  $X$ , the mean or expected value of  $X$  is

$$\bar{X} = \sum_x x p(x)$$

Another common notation for the mean is  $\langle X \rangle$ . The notation  $\langle X \rangle_{p(x)}$  is used when there can be confusion about which distribution is used to compute the expected value. The variance, which is a measure of the spread around the mean, is defined as

$$\text{Var } X = \sum_x (x - \bar{x})^2 p(x)$$

The variance can alternatively be written as

$$\text{Var } X = \left( \sum_x x^2 p(x) \right) - \bar{x}^2$$

Exercise: Show that these two expressions for variance are identical.

The standard deviation is the square root of the variance. Mean and variance are special cases of the *expected value* of any function of a random variable. If we denote the function by  $f$ , then the expected value, denoted  $E_X(f(X))$ , is

$$E_X(f(X)) = \sum_x f(x) p(x)$$

The notation  $\langle f(X) \rangle_{p(x)}$  is also used. Thus, the mean is the expected value of  $X$ ,  $\bar{x} = E_X(X)$ , and the variance is  $\text{Var } X = E_X((X - E_X(X))^2) = E_X(X^2) - E_X(X)^2$ . The subscript  $X$  can be omitted when no misunderstanding is possible regarding the variable over which the summation is performed.

For a continuous random variable  $X$  with probability density  $p(x)$ , the analogous expressions are obtained by replacing sums with integrals:

$$\bar{x} = \int_{-\infty}^{\infty} xp(x) dx$$

$$\text{Var } X = \int_{-\infty}^{\infty} (x - \bar{x})^2 p(x) dx = \left( \int_{-\infty}^{\infty} x^2 p(x) dx \right) - \bar{x}^2$$

$$E_X(f(X)) = \int_{-\infty}^{\infty} f(x) p(x) dx$$

## 7 The normal distribution

### 7.1 Definition

The most important continuous distribution in nearly all applications of probability theory is the normal or Gaussian distribution. Its probability density function is

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

We will sometimes use  $p(x) = N(x; \mu, \sigma^2)$  as short-hand notation. Examples of normal distributions can be found in every chapter of this book. The parameters  $\mu$  and  $\sigma^2$  do not have an a-priori meaning (they are just denoted suggestively), but they, of course turn out to be equal to the mean and variance of the distribution, respectively. The factor  $\frac{1}{\sqrt{2\pi\sigma^2}}$  is needed for normalization.

Exercise: Show that this is the correct normalization factor.

### 7.2 Central limit theorem

The importance of the normal distribution derives mainly from the central limit theorem. Roughly, the central limit theorem states that the mean of a large number of independent random variables with identical probability distributions will follow an approximately normal distribution, regardless of the distribution of the original variables. This theorem is most powerful because of its last part: the distribution of the original variables is irrelevant. The theorem can be relaxed to allow for independent, but not identically distributed variables.

In mathematical models of perception, the central limit theorem always plays a role in the background: whenever we assume that the noise corrupting a stimulus is normally distributed, we are essentially motivating this using the central limit theorem. The random variable describing the noise corrupting a stimulus might be the sum of a large number of independent noise processes.

### 7.3 Multiplying two normal distributions

Let's consider the product of two Gaussian probability distributions over the same random variable  $X$ . One has mean  $\mu_1$  and variance  $\sigma_1^2$ , and the other mean  $\mu_2$  and variance  $\sigma_2^2$ . We will multiply these two distributions just as we would multiply regular functions, and then we will normalize the result (since the product is not automatically normalized). What is the resulting probability distribution?

We first write down the expressions for the two probability density functions:

$$p_1(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}, \quad p_2(x) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}. \quad (\text{A.3})$$

Multiplying these two functions comes down to summing the exponents. We will do that first:

$$\text{sum of exponents} = -\frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{(x-\mu_2)^2}{2\sigma_2^2} = -\frac{1}{2} \left( \frac{x^2 - 2\mu_1 x + \mu_1^2}{\sigma_1^2} + \frac{x^2 - 2\mu_2 x + \mu_2^2}{\sigma_2^2} \right).$$

We reorganize by collecting all terms containing  $x^2$ , and all containing  $x$ :

$$\text{sum of exponents} = -\frac{1}{2} \left( \left( \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right) x^2 - 2x \left( \frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2} \right) + \dots \right)$$

Here and in the following, the dots represent all terms that do not depend on  $x$ . When exponentiated, these terms become a multiplicative constant that is independent of  $x$ . Since the resulting product of distributions must be normalized at the end of the calculation anyhow, the multiplicative constant is irrelevant. The sum of exponents can be written as

$$\text{sum of exponents} = -\frac{1}{2 \left( \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right)^{-1}} \left( x^2 - 2x \frac{\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} + \dots \right) = -\frac{1}{2 \left( \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right)^{-1}} \left( x - \frac{\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} \right)^2 + \dots$$

Thus, the product of the distributions in Eq. (A.3) is

$$p_1(x) p_2(x) \propto \exp \left[ -\frac{1}{2 \left( \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right)^{-1}} \left( x - \frac{\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} \right)^2 \right], \quad (\text{A.4})$$

where the proportionality sign absorbs all factors that are independent of  $x$ . We recognize this as

another normal distribution, now with mean  $\frac{\mu_1 + \mu_2}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}$  and variance  $\frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}$ .

Exercise: What is the correct normalization constant in Eq. (A.4)?

## 7.4 Multiplying N normal distributions

We now generalize the previous section to  $N$  normal distributions. This is used in Section 6.X (Sameness judgment). Consider a set of  $N$  normal distributions over the same variable  $x$ . The  $i^{\text{th}}$  distribution has mean  $\mu_i$  and variance  $\sigma_i^2$ . The (unnormalized) product of these distributions is equal to

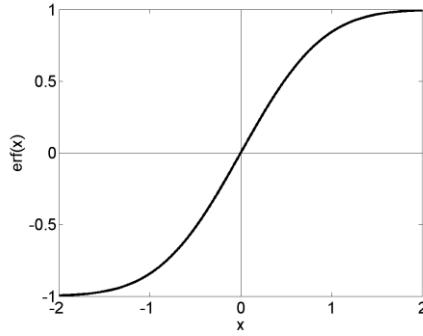
$$\prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma_i^2}\right) = \frac{1}{(2\pi)^{\frac{N}{2}} \prod_i \sigma_i} \exp\left(-\frac{1}{2} \left( \sum_i \frac{1}{\sigma_i^2} \right) \left( \mu - \frac{\sum_i \frac{x_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}} \right)^2\right) \exp\left(-\frac{1}{2} \left( \sum_i \frac{x_i^2}{\sigma_i^2} - \frac{\left(\sum_i \frac{x_i}{\sigma_i^2}\right)^2}{\sum_i \frac{1}{\sigma_i^2}} \right)\right),$$

## 7.5 The error function

The cumulative distribution function of a Gaussian distribution is not an elementary function (i.e. one built from exponentials, logarithms, and powers using addition, subtraction, multiplication, and division). However, because it occurs very often, it has been named a special function. To be precise, the *error function* is defined as

$$\text{erf } x = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

This function takes values between  $-1$  and  $1$  (negative when  $x$  is negative), as shown in Figure A3.



**Figure A3.** The error function.

Exercise: Show that the error function is an odd function, i.e.  $\text{erf}(-x) = -\text{erf}(x)$ .

Exercise: Show that the integral of a Gaussian distribution and the error function are related in the following way:

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_0^x e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt = \frac{1}{2} \text{erf} \frac{x-\mu}{\sigma\sqrt{2}} + \frac{1}{2} \text{erf} \frac{\mu}{\sigma\sqrt{2}}.$$

### Box: Integrals

We need only two integrals in this book:

$$\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sqrt{2\pi\sigma^2}$$

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{2} \text{erf} \frac{b-\mu}{\sigma\sqrt{2}} - \frac{1}{2} \text{erf} \frac{a-\mu}{\sigma\sqrt{2}}$$

The error function is shown in Fig. A3.

## 7.6 The Von Mises distribution

Some variables we consider in this book, such as orientation and motion direction, are circular (periodic). On a circular space, the normal distribution is not well defined. An analog of the normal distribution is the Von Mises distribution. It is defined as

$$p(x) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(x-\mu)},$$

where  $\mu$  is the (circular) mean of the distribution, and  $\kappa$  is called the concentration parameter.  $I_0$  is the modified Bessel function of the first kind of order zero. This is a so-called *special function*, which is defined in terms of an integral or infinite series. Its precise definition is not relevant here; for us, all that matters is that  $I_0(\kappa)$  normalizes the Von Mises distribution. When  $\kappa=0$ , the Von Mises distribution becomes a uniform distribution. The higher  $\kappa$ , the more similar the Von Mises distribution becomes to a normal distribution. This is illustrated in Fig. XX.

Exercise: How would you define the mean of a circular variable?

Exercise: Show analytically that in the limit of large  $\kappa$ , the Von Mises distribution becomes the normal distribution. (Hint: use the Taylor expansion of the cosine.) How is the variance of that normal distribution related to  $\kappa$ ?

Exercise: Consider two Von Mises distributions over  $X$ , one with mean  $\mu_1$  and concentration parameter  $\kappa_1$ , the other with mean  $\mu_2$  and concentration parameter  $\kappa_2$ . Show that the normalized product of these distributions is again a Von Mises distribution, and compute its mean and concentration parameter.

## 8 The delta distribution

A special type of random variable that we will encounter quite often is one that takes only one possible value. There is a special notation for the probability distribution of such a random variable. If  $X$  is a discrete random variable that always takes the value  $x=a$ , then its distribution is sometimes written  $p(x)=\delta_{xa}$ , where  $\delta$  is called the Kronecker delta. It is defined as

$$\delta_{xa} = \begin{cases} 1 & \text{if } x = a \\ 0 & \text{otherwise} \end{cases}$$

If  $X$  is a continuous random variable that always takes the value  $x=a$ , then we write for its distribution,

$$p(x) = \delta(x-a). \quad (\text{A.5})$$

Here,  $\delta$  is the Dirac delta function. It returns 0 unless the argument equals 0, in which case it returns infinity. Of course, infinity is not a number and therefore the Dirac delta function is strictly speaking not an ordinary function. This problem is not a concern, since the only place in which we will use the delta function is inside an integral. There, the following property holds for any function  $f(x)$ :

$$\int \delta(x-a) f(x) dx = f(a). \quad (\text{A.6})$$

In fact, this is the defining property of the Dirac delta. Again, the delta function has the effect of evaluating the function  $f$  inside the integral at a single point,  $a$ .

We find it more convenient to use the same notation for discrete as for continuous variables, i.e., Eq. (A.5). The discrete analog of Eq. (A.6) is

$$\sum_x \delta(x-a) f(x) = f(a), \quad (\text{A.7})$$

where  $f$  is any function on a discrete domain (or viewed differently, a vector in which  $x$  indexes the entries).

## 9 The Poisson distribution

A discrete probability distribution that we use to describe neural activity (Chapters 11 and 12) is the Poisson distribution. The possible values of a Poisson random variable are  $0, 1, 2, 3, \dots$  (there is no upper limit). The Poisson distribution has a free parameter, which we will call  $\lambda$ . The probability distribution of  $X$  is given by

$$p(x) = \Pr(X=x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad (\text{A.8})$$

where  $x! = 1 \cdot 2 \cdot 3 \cdot \dots \cdot x$  is the factorial operation. This distribution is only defined on the non-negative integers, and is therefore discrete. The factor  $\exp(-\lambda)$  acts as a normalization factor.

Exercise: Prove that  $\exp(-\lambda)$  is the correct normalization factor for the Poisson distribution. Hint: use the Taylor expansion of the exponential function.

The Poisson distribution, Eq. (A.8), is plotted in Figure XX for two different values of  $\lambda$  [suggest that we make a single multi-panel figure (or perhaps 4 figures, one for each distribution) to plot each distribution we describe in the appendix: Gaussian, Von Mises, delta, Poisson. Even if these distributions are plotted elsewhere in the book, a figure here would be nice to show.]. Note that unlike  $x$ ,  $\lambda$  does not have to be an integer. The parameter  $\lambda$  has a special meaning that might be apparent from these figures.

Exercise: Prove that both the mean and the variance of the Poisson distribution are equal to  $\lambda$ .

## 10 Distributions involving multiple variables

Random variables can depend on each other in interesting ways. This is formalized in joint and conditional probability distributions, and in Bayes' rule, which we derive here formally. The concepts discussed in this section apply to both continuous and discrete variables. Thus, probability or  $p$  can refer to either probability mass or probability density. Since we consider

multiple variables at the same time, we will generally use a subscript on  $p$  to denote which random variable(s) the probability distribution belongs to.

### 10.1 Joint probability

The *joint probability distribution* of random variables  $X$  and  $Y$  is denoted  $p_{X,Y}(x,y)$  or, with our notation convention from Section 5.7 in mind,  $p(x,y)$ . It is the probability of the values  $x$  and  $y$  as a pair. Summing over both  $x$  and  $y$  gives 1:

$$\sum_x \sum_y p(x,y) = 1$$

For continuous variables, the double sum is replaced by a double integral:

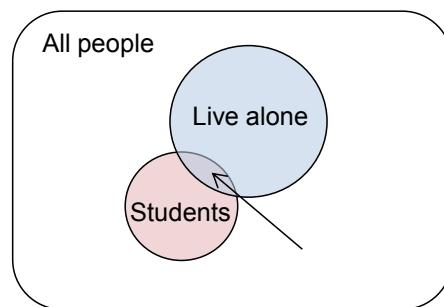
$$\iint p(x,y) dx dy = 1.$$

Joint probability is symmetric:

$$p(x,y) = p(y,x).$$

If  $X$  and  $Y$  represent events, the joint probability of  $X$  and  $Y$  is the probability that both occur, denoted  $p(X,Y)$  or  $p(X \cap Y)$ . In the Venn diagram representation (Fig. A4), we represent  $Y$  by another circle, intersecting the first one. The joint probability of  $X$  and  $Y$  is equal to the area of the intersection. It is always less than or equal to the area of each individual circle. This expresses the relations  $p(X,Y) \leq p(X)$  and  $p(X,Y) \leq p(Y)$ . These relations only hold for discrete variables.

Example: The probability that it rains on a given day and you will be at work on time is smaller than the probability that it rains.



**Figure A4.** The joint probability of the events “being a student” and “living alone” is represented by the area of the intersection, indicated by the arrow.

## 10.2 Marginalization

Marginalization is the operation of obtaining from a joint distribution over multiple variables the distribution over a subset of those variables. For example, if  $p(x,y)$  is the joint distribution of  $X$  and  $Y$ , then summing over  $Y$  produces the distribution of  $X$  alone:

$$\sum_y p(x,y) = p(x).$$

A daily-life example: Aida has carefully tracked what the probabilities are during the day that only her cat is present in the living room, only her dog, neither, or both. These probabilities are shown in Fig. A5; this table, called a *contingency table*, represents the probabilities of joint outcomes. The marginal probabilities are the probabilities that the cat is present or absent regardless of the dog, and the probabilities that the dog is present or absent regardless of the cat.

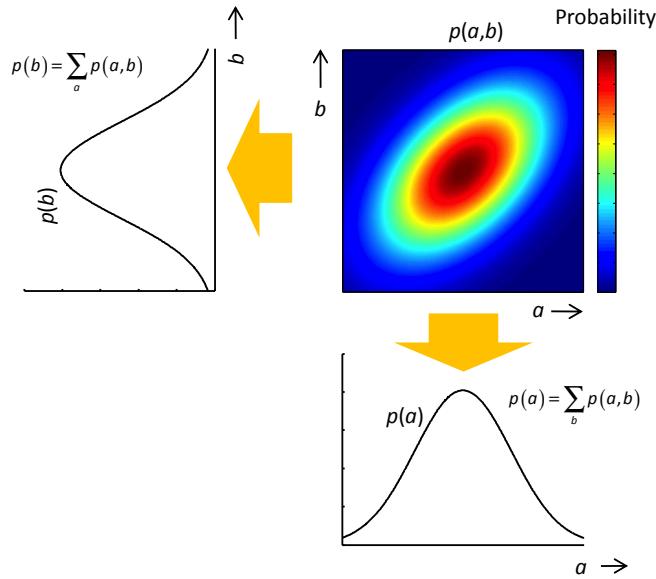
		CAT		
		absent	present	
DOG	absent	0.40	0.05	0.45
	present	0.30	0.25	0.55
		0.70	0.30	

**Figure A5.** The red numbers represent marginal probabilities of the joint probability distribution of the events “dog is present” and “cat is present”.

The continuous analog is obtained by replacing the sum by an integral:

$$\int p(x,y) dy = p(x).$$

This summation or integration is called “marginalization” because  $p(x)$  and  $p(y)$  are called the marginals of  $p(x,y)$ . If you think of  $(x,y)$  as a point in two-dimensional space, and the joint distribution providing  $z$ -values in this space, then the marginals are the distributions obtained by summing in either dimension (Fig. A6). This results in two one-dimensional distributions that live in the “margins” of the original two-dimensional distribution.



**Figure A6.** The color plot represents the joint probability distribution of two random variables  $A$  and  $B$ , the black curves the two marginals, obtained by summing the joint across of the the two variables.

Marginalization is an important, frequently used procedure. We encounter marginalization in several places throughout this book: in Chapter 1 (when discussing the denominator of the right-hand side of Bayes' rule), in Chapter 6 (models requiring marginalization), and in Chapter 8 (temporal integration), among others.

### 10.3 Conditional probability

The probability of  $x$  given  $y$  is denoted  $p_{X|Y}(x|y)$  or, with our notation convention from Section 5.7 in mind,  $p(x|y)$ . The “|” sign is read as “given” or “conditioned on”. It is defined as the probability of  $x$  and  $y$  as a pair, divided by the probability of  $y$ :

$$p(x|y) = \frac{p(x,y)}{p(y)}. \quad (\text{A.9})$$

Consider these three examples of conditional probability when  $X$  and  $Y$  are discrete random variables:

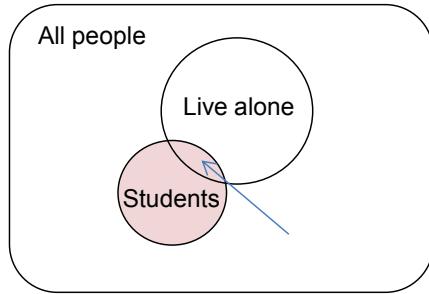
Example: if the probability that it rains today and you arrive at work on time is equal to 0.4, and the probability that it rains today is 0.5, then the probability that you arrive at work on time given that it rains is  $0.4/0.5 = 0.8$ .

Example: the probability that I roll a 6 on a die given that I roll an even number is equal to  $(1/6)/(1/2) = 1/3$ .

Example: In a given country, each state has a different proportion of taxi drivers. The probability that a person randomly selected from a particular state is a taxi driver,  $p(x|y)$ , is equal to the proportion of people in the country who live in that state *and* are taxi drivers,  $p(x,y)$ , divided by the proportion of people living in that state,  $p(y)$ .

From the contingency table in Fig. A5, conditional probabilities can readily be computed. For example, the probability that the cat is present given that the dog is present is 0.25 (cat and dog both present) divided by  $0.30+0.25 = 0.55$  (dog present).

The conditional probability  $p(X|Y)$  is the answer to the question: “of all outcomes that are consistent with event  $Y$ , what fraction is also consistent with event  $X$ ?” Conditional probability of an event always lies between 0 and 1. In the Venn diagram representation (Fig. A7),  $p(X|Y)$  is equal to the area of the intersection divided by the area of the second circle. Similarly, the probability that  $Y$  occurs given that  $X$  occurs is equal to the area of the intersection divided by the area of the first circle.



**Figure A7.** The conditional probability probability of “living alone” given “being a student” is represented by the area of the intersection divided by the area of the red disc.

It is easy to see, either from equation (A.9) or from the figure, that  $p(X|Y)$  is not equal to  $p(Y|X)$ . We already saw this in Chapter 1 (Figure 1.3). Mistakenly equating the two is known as the conditional probability fallacy or the prosecutor’s fallacy (see also the Box in Chapter 1).

Example: The probability is very high that a professional basketball player is tall. However, the probability is very low that a tall person is a professional basketball player. If  $X$  is “being a basketball pro” and  $Y$  is “being tall”, then this example illustrates that  $p(X|Y)$  is not equal to  $p(Y|X)$ .

The same holds for conditional probability distributions over continuous variables, i.e. the probability density  $p_{X|Y}(x|y)$  is not equal to  $p_{Y|X}(y|x)$ .

A (discrete or continuous) conditional probability distribution  $p(x|y)$  is a probability distribution over  $x$ , but not over  $y$ . The distinction arises from the fact that when summed over  $x$ ,  $p(x|y)$  adds up to 1, but when summed over  $y$  it does not necessarily add up to 1.

Exercise: Show formally that  $p(x|y)$  is normalized as a function of  $x$ .

Exercise: Give a counterexample that shows that  $p(x|y)$  is not normalized as a function of  $y$ .

Although  $p(x|y)$  is not normalized as a function of  $y$ , it is still a function of  $y$ . It is called the *likelihood function* of  $y$ . Remember: likelihood functions are not probability distributions, because they are not normalized, and they are always functions of the variable *after* the “|” sign. It would be wrong to talk about  $p(x|y)$  as “the likelihood of  $x$ ”. It might be useful at this point to read the box “A daily-life example of likelihood” in Section 2.3.2.

We will now combine the notion of marginalization with the definition of conditional probability.

Exercise: Show that

$$p(x) = \sum_y p(x|y)p(y). \quad (\text{A.10})$$

Eq. (A.10) and its continuous analog,  $p(x) = \int p(x|y)p(y)dy$ , are rules that we will use throughout the book. Continuing on the taxi driver example: suppose I am interested in the probability that a randomly selected citizen is a taxi driver. I know for each state the proportion of taxi drivers. I also know the proportion of all citizens living in each state. To obtain my answer, I multiply those two proportions for every state and then sum over all provinces.

We can condition every probability in Eq. (A.10) on a third random variable,  $z$  (this can be done with any rule in probability calculus). Then we get

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

Or in its integral form

$$p(x|z) = \int p(x|y,z)p(y|z)dy$$

Exercise: prove this formally using the definition of conditional probability and the marginalization rule.

Conditional distributions are not limited to a single random variable before and after the given sign. For example, one could consider the distribution of  $X$  and  $Y$  given  $Z$  and  $W$ , denoted  $p(x,y|z,w)$ . One can marginalize such conditional distributions:

Exercise: Show that  $\sum_y p(x, y | z) = p(x | z)$ .

Marginalization in the form of Eq. (A.10) is closely related to the *convolution* operation. Whenever the distribution  $p(x|y)$  is a function only of the difference  $x-y$ , Eq. (A.10) is the definition of the convolution of the two probability distributions.

## 10.4 Independence

Two random variables  $X$  and  $Y$  are called independent if their joint distribution factorizes into the marginals, i.e., if

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

for all  $x$  and  $y$ . For example, the probability that I roll a 6 on a die and toss heads on a coin is the product of both events taken separately. Independence can be depicted graphically as in Figure X: one can reconstruct the joint distribution by multiplying the marginals. The notion of independence is closely related to that of correlation: two independent random variables are also uncorrelated. The opposite is not true.

Exercise: Why not?

Exercise: If  $X$  and  $Y$  are independent, what can one say about the conditional distributions  $p_{X|Y}(x|y)$  and  $p_{Y|X}(y|x)$ ?

*Conditional independence*

If  $X$ ,  $Y$ , and  $Z$  denote three random variables, then  $X$  and  $Y$  are conditionally independent given  $Z$  if

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

for any values  $x$ ,  $y$ , and  $z$ . This is discussed in Chapter 4. Never confuse conditional independence with independence!

## 10.5 Bayes' rule

We saw before that the conditional probabilities  $p(x|y)$  and  $p(y|x)$  are not equal. Bayes' rule relates them to one other:

$$p(x | y) = \frac{p(y | x)p(x)}{p(y)} \tag{A.11}$$

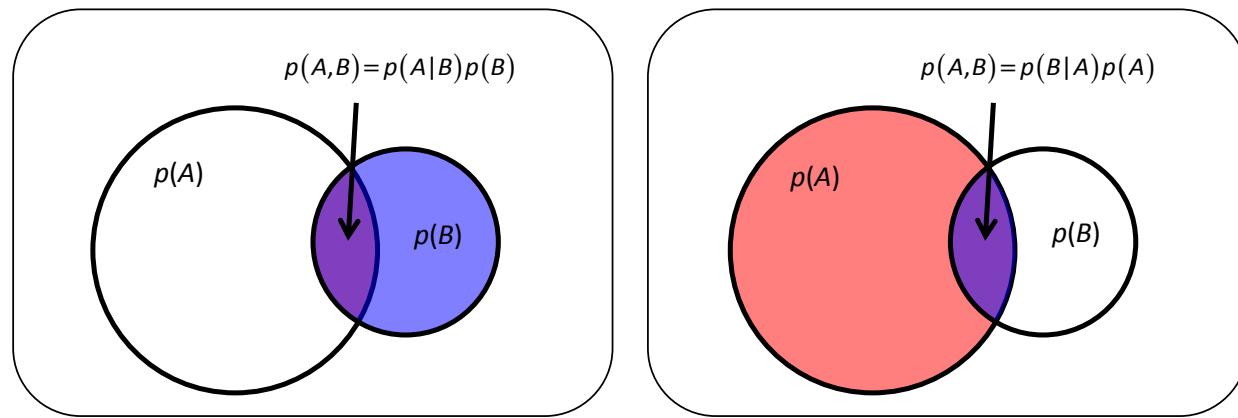
Here,  $p(y|x)$  as a function of  $x$  is the *likelihood function* over  $x$ ,  $p(x)$  is the *prior distribution* over  $x$ , and  $p(x|y)$  is the *posterior distribution* over  $x$ .

Exercise: Before reading on, try to prove Bayes' rule using the equations in the preceding sections.

Here is how the proof goes: From Eq. (A.9), we know that  $p(x,y)=p(x|y)p(y)$ . By renaming  $x$  and  $y$ , we also obtain  $p(y,x)=p(y|x)p(x)$ . Joint probability is symmetric,  $p(x,y)=p(y,x)$ . From these three equations, it follows that  $p(x|y)p(y)=p(y|x)p(x)$ . Dividing both sides by  $p(y)$  gives Bayes' rule.

Exercise: Prove that the right-hand side of Eq. (A.11) is normalized over  $x$ .

The Venn diagram interpretation of Bayes' rule for events  $X$  and  $Y$  is that the area of overlap can be calculated in two ways (Fig. A8): as a fraction of the  $X$ -circle area times the  $X$ -circle area, or as a fraction of the  $Y$ -circle area times the  $Y$ -circle area. Since the outcomes should be identical, this means that the two fractions can be expressed in terms of each other if one knows the ratio of areas of the  $X$ - and  $Y$ -circles.



**Figure A8.** Bayes' rule is obtained by writing the intersection area in two different ways and equating the two.

Example: Suppose that 1 in 100,000 people are professional basketball players, that 1 in 100 people are tall, and that 95% of basketball pros are tall. What is the probability that a tall person is a professional basketball player?

We solve this problem using a direct application of Bayes' rule: If  $X$  is “being a basketball pro” and  $Y$  is “being tall”, then  $p(X)=0.00001$ ,  $p(Y)=0.01$ , and  $p(Y|X)=0.95$ . It follows that  $p(X|Y)=0.95 \cdot 0.00001/0.01=0.00095$ , or about 1 in 1000.

Exercise: Prove a different form of Bayes' rule:

$$p(x|y) = \frac{p(y|x)p(x)}{\sum_x p(y|x)p(x)}$$

Since  $p(y)$  does not depend on  $x$ , it is a constant multiplicative factor in Eq. (A.11), and acts as a normalization factor. Since the normalization is determined by integrating the numerator,  $p(y|x)p(x)$ , it is often left out for convenience and replaced by a proportionality sign:

$$p(x|y) \propto p(y|x)p(x). \quad (\text{A.12})$$

Here, it is understood that the constant of proportionality is such that the left-hand side is normalized. See also the Box “Why the proportionality sign?” in Section 2.3.3. For binary variables, the normalization is usually written out, since it contains only two terms then.

Bayes' rule in the taxi driver example: you know what proportion of people in each state is a taxi driver. You know what proportion of the population lives in each state. You are told Mr. Lagrange is a taxi driver. What is your best guess of which state he is from? This is computed by multiplying the probability that someone lives in a certain state by the probability that someone in that state is a taxi driver, and comparing this product (the numerator of Bayes' rule) across states.

## 11 Functions of random variables (Advanced)

### 11.1 Functions of one variable: changing variables

In this section, we discuss the frequently occurring problem of transforming the distribution of a continuous random variable. If  $X$  is a random variable with probability distribution  $p_X(x)$ , and  $Y=f(X)$  is a new random variable obtained by applying the function or transformation  $f$  to  $X$ , what is the distribution of  $Y$ ? In this section, we will use subscripts like  $X$  in  $p_X(x)$  to avoid confusion, since there are multiple random variables.

Example:  $X$  is a random variable following a uniform distribution on  $[0,1]$ .  $Y=X^2$  is a new random variable obtained by squaring outcomes of  $X$ . What is the distribution of  $Y$ ?

An easy but wrong answer would be that because  $X$  follows a uniform distribution,  $Y$  does as well. It can be understood intuitively that this answer is wrong. When a number  $x$  lies between 0 and 1, squaring it will always make it smaller. Thus, even though the values of  $Y$  will also lie between 0 and 1, lower values in this range will have greater probability density than higher values do. The question can be answered correctly by considering the cumulative distribution functions of  $X$  and  $Y$ , which we will denote  $P_X(x)$  and  $P_Y(y)$ , respectively:

$$P_Y(y) = \Pr(Y \leq y) = \Pr(X^2 \leq y) = \Pr(X \leq \sqrt{y}) = P_X(\sqrt{y}).$$

Now, we use the fact that the probability density function is the derivative of the cdf, Eq. (A.2), to find the pdf of  $y$ , denoted  $p_Y(y)$ :

$$p_Y(y) = \frac{dP_Y}{dy} = \frac{d}{dy} P_X(\sqrt{y}) = \frac{dP_X}{dx} \Big|_{x=\sqrt{y}} \frac{d}{dy} \sqrt{y} = p_X(\sqrt{y}) \frac{1}{2\sqrt{y}} = \frac{1}{2\sqrt{y}},$$

where the third equality is obtained by applying the chain rule. The resulting distribution,  $p_Y(y)$ , is normalized (verify this) and conforms to our intuition: the probability density is higher for lower values of  $y$ .

We could have stated the same problem with  $p_X(x)$  being any distribution. The calculation is then identical except for the last step. Thus, we find  $p_Y(y) = p_X(\sqrt{y}) \frac{1}{2\sqrt{y}}$ . Thus, the distribution of the squared variable is a product of the original distribution evaluated at the value of  $x$  that maps to  $y$ ,  $p_X(\sqrt{y})$ , and an extra factor. The extra factor is equal to the derivative of the mapping from  $y$  to  $x$ . It would be wrong to leave out the extra factor, writing that  $p_Y(y) = p_X(\sqrt{y})$ , or to assume that the distribution of a squared variable is given by the square of the distribution,  $p_Y(y) = p_X(y)^2$ .

An extra factor like this appears not just in this example, but in our original, general problem. Suppose  $X$  is a random variable with probability distribution  $p_X(x)$ , and  $Y=f(X)$ , where  $f$  is a monotonically increasing function. What is the distribution of  $Y$ ? We first define the inverse function  $f^{-1}$  as the function of  $y$  that “undoes” the effect of  $f$ , in other words,  $f^{-1}(f(x))=x$ . This inverse function is well-defined because we assumed that  $f$  is monotonically increasing. Naïve but wrong ways to obtain the distributions of  $Y$  would be to substitute the inverse function into  $p_X$ ,  $p_Y(y)=p_X(f^{-1}(y))$ , or to assume that an operation applied to a distribution is the same as the distribution applied to the operation,  $p_Y(y)=f(p_X(y))$ . The correct approach is again to calculate the cumulative distribution of  $Y$ ,

$$P_Y(y) = \Pr(Y \leq y) = \Pr(f(X) \leq y) = \Pr(X \leq f^{-1}(y)) = P_X(f^{-1}(y)),$$

and from that the probability density function of  $y$ ,

$$p_Y(y) = \frac{dP_Y}{dy} = \frac{d}{dy} P_X(f^{-1}(y)) = \frac{dP_X}{dx} \bigg|_{x=f^{-1}(y)} \frac{d}{dy} f^{-1}(y) = p_X(f^{-1}(y)) \frac{d}{dy} f^{-1}(y). \quad (\text{A.13})$$

So far, we have considered a monotonically increasing function  $f$ . When  $f$  is instead monotonically decreasing, the final expression for  $p_Y(y)$  acquires an extra minus sign.

Exercise: Show this.

We can summarize both cases (monotonically increasing and decreasing) in a single equation:

$$p_Y(y) = p_X(f^{-1}(y)) \left| \frac{d}{dy} f^{-1}(y) \right|. \quad (\text{A.14})$$

Exercise: Use Eq. (A.14) to prove the first property of a normally distributed variable in the Box in Section 2.4.

Exercise: If  $X$  is an exponentially distributed random variable on the positive real line, what are the domain and distribution of  $Y=e^X$ ?

An informal but intuitive way of writing Eq. (A.14) is  $p_Y(y) = p_X(x) \left| \frac{dx}{dy} \right|$ , where it is

understood that  $x=f^{-1}(y)$ . How can the extra factor  $\left| \frac{dx}{dy} \right|$  be interpreted and why can it not be left

out? Recall that for a continuous variable, the probability of an event is defined as the integral of the probability density function between two points. The integral can be approximated (and is in fact defined as) a sum of areas of very narrow rectangles that together fill the interval. The factor  $\left| \frac{dx}{dy} \right|$  stems from the fact that a rectangle of width  $dy$  in the domain of the transformed variable  $Y$

does not always correspond to a rectangle of the same width in the domain of the original variable,  $X$ . The derivative can be regarded as a “magnification factor”. Equivalently, we can state that  $p_Y(y)|dy| = p_X(x)|dx|$ , because the left side of this equation is the probability that  $Y=f(X)$  lies in the infinitesimal region  $y \pm dy/2$ , which it does when  $x$  lies in the infinitesimal region  $x \pm dx/2$ . Note that transforming discrete random variables does not involve any factor

such as  $\left| \frac{dx}{dy} \right|$ , since the probability mass is concentrated in discrete points.

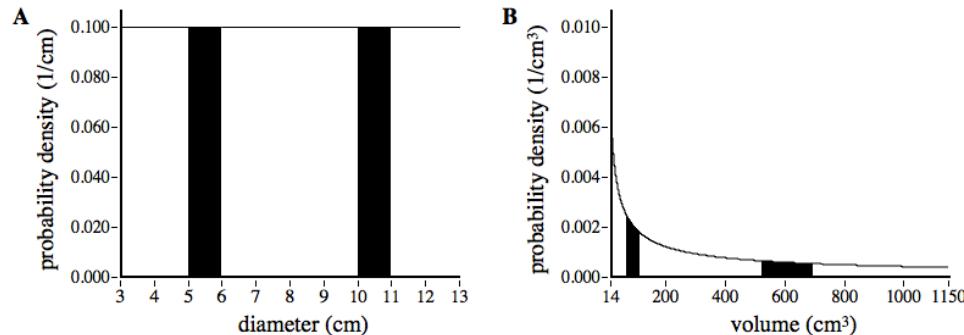
As a final illustration of the change-of-variables procedure, let's suppose that you are going to visit an apple orchard. You know very little about how fast apples grow or the duration of the

growing season in the area, and you don't know the type of apples in the orchard. If a friend asked you what you thought the size of the apples in the orchard was going to be, you might initially respond that you have no idea. Upon more careful consideration, drawing upon your limited knowledge of apples in general, suppose you state that you have a uniform prior density over the diameter of apples in the orchard, from 3 to 13 cm. What, then, is your prior density over apple *volume*?

Before we derive the answer, let's appreciate the problem. Your uniform prior over apple diameter means that, for example, you consider it equally probable that an apple's diameter will lie between 5 and 6 cm as between 10 and 11 cm. If we approximate apples as spheres, then the volume of an apple is

$$v = \frac{4\pi r^3}{3} = \frac{\pi w^3}{6}$$

where  $r$  is the radius and  $w=2r$  is the diameter of the apple. The volumes corresponding to diameters of 5, 6, 10, and 11 cm are therefore (to the nearest integer) 65, 113, 524, and 697 cm<sup>3</sup>, respectively. This means that you believe it is equally probable (10% probable, to be exact) for the volume of an apple to lie between 65 and 113 cm<sup>3</sup> - a range of 48 cm<sup>3</sup> - as it is to lie between 524 and 697 cm<sup>3</sup> - a range of 173 cm<sup>3</sup>. Your prior density over apple volume, then, is clearly not flat. Rather, the density will be higher at smaller volumes (see Figure A9).



**Figure A9.** Change of variables. **A.** A uniform prior over apple diameter, from 3 to 13 cm (a range of 10cm). The prior density is a line at height  $0.1\text{cm}^{-1}$ , because the total area under the density must equal 1. The probability that the diameter lies between 5 and 6 cm is 0.1, as is the probability the diameter lies between 10 and 11 cm (filled rectangular areas). **B.** The prior over apple volume. Each filled rectangular area is the probability that apple volume lies in the range corresponding to the apple diameters covered by the filled rectangles in A. Again, the total area under the density is 1, and the area of each filled rectangle is 0.1. (Note the differences in y-axis scales).

To derive your density over volume, we note that:

$$w = \left( \frac{6v}{\pi} \right)^{1/3}$$

From which it follows, after some rearrangement, that:

$$\frac{dw}{dv} = \left( \frac{2}{9\pi v^2} \right)^{1/3}$$

Therefore,

$$p_V(v) = p_W(w) \left| \frac{dw}{dv} \right| = p_W(w) \left( \frac{2}{9\pi v^2} \right)^{1/3} = \frac{1}{10cm} \left( \frac{2}{9\pi v^2} \right)^{1/3}$$

This is the curve plotted in Fig. A9b.

### Box: Ignorance and Reference Priors

An interesting consequence of the change-of-variables procedures, illustrated by the apple orchard example, is that it is not possible to be ignorant about every feature of a problem. For instance, it is not possible to be fully ignorant about apple size, generally defined. As we have just seen, if we are ignorant about apple diameter, in the sense that we consider a wide range of diameters to be equally probable, then we are consequently not ignorant about apple volume! When doing Bayesian statistical analysis, a researcher may want to incorporate as little prior opinion as possible into an analysis about which she feels she has almost no relevant background knowledge. How can she best do this, if by specifying her ignorance about a parameter, she is consequently specifying knowledge about related parameters? For instance, if a researcher has “no knowledge” of the standard deviation,  $\sigma$ , of a random variable, she may choose to use a flat prior over a very wide range of  $\sigma$ , but then she is implicitly specifying a non-uniform prior over the variance,  $\sigma^2$ . The search to develop appropriate default or reference priors for such situations is an interesting topic in the field of Bayesian statistical analysis.

## 11.2 Marginalization formulation

It is instructive to phrase the problem of transforming the distribution of a random variable as a formal problem of marginalization. This formulation is equivalent to the one in Section 11.1, but gives more insight in some ways. We assume again that  $X$  is a random variable with probability distribution  $p(x)$ , and  $Y=f(X)$ , where  $f$  is a monotonically increasing function. As we discussed in Section 8, a deterministic mapping such as  $f$  can be expressed as a delta distribution. Here, this distribution would take the form  $p_{Y|X}(y|x)=\delta(y-f(x))$ . Now we can compute the probability density at  $y$  formally using the marginalization identity from Eq. (A.10):

$$p_Y(y) = \int_{-\infty}^{\infty} p_{Y|X}(y|x) p_X(x) dx = \int_{-\infty}^{\infty} \delta(y-f(x)) p_X(x) dx. \quad (\text{A.15})$$

In words, the probability density of  $y$  is the total probability of all values of  $x$  whose image under  $f$  is  $y$ . We can evaluate this expression by making a transformation of variables:  $x=f^{-1}(t)$ , so that

$dx = \frac{df^{-1}}{dt} dt$ . Substituting, we find

$$p_Y(y) = \int_{-\infty}^{\infty} \delta(y - f^{-1}(t)) p_X(f^{-1}(t)) \frac{df^{-1}}{dt} dt$$

We can now use Eq. (A.6) to evaluate the integral:  $p_Y(y) = p_X(f^{-1}(y)) \frac{df^{-1}(y)}{dy}$ , which is the same as Eq. (A.13). Again, when  $f$  is monotonically decreasing instead of increasing, we obtain the same result but with a minus sign.

Exercise: Where does the minus sign come from in this formulation?

The advantage of this integral formulation is that the first equality in Eq. (A.15) is general and not limited to deterministic mappings from  $X$  to  $Y$ . Thus, the problem of transforming a random variable is simply a special case of a probabilistic mapping from  $Y$  to  $X$ , and the first equality in Eq. (A.15) can be applied for *any* conditional distribution  $p(y|x)$ .

A second advantage is that the expected value of any function  $g(Y)$  of a random variable  $Y=f(X)$  is now easy to transform:

$$\begin{aligned} E_Y(g(Y)) &= \int g(y) p_Y(y) dy \\ &= \int g(y) \left( \int \delta(y - f(x)) p_X(x) dx \right) dy \\ &= \int \left( \int g(y) \delta(y - f(x)) dy \right) p_X(x) dx \\ &= \int g(f(x)) p_X(x) dx \\ &= E_X(g(f(X))), \end{aligned}$$

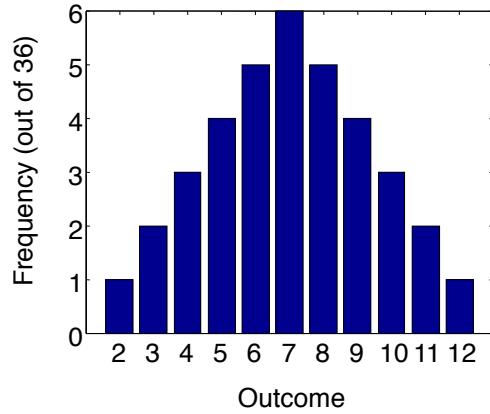
where from the second to the third line we swapped the order of integration. In other words, the combination  $p_Y(y)dy$  inside an integral is identical to  $p_X(x)dx$ , as long as  $y=f(x)$  is substituted elsewhere in the integral. (The integration limits might also change accordingly.)

Exercise: Use this to show that the mean of  $aX+b$  is  $aE(X)+b$ , and that its variance is  $a^2\text{Var } X$ .

A third advantage of the marginalization formulation is that it directly generalizes to functions of multiple variables, as we will now examine.

### 11.3 Functions of multiple variables

I roll two fair dice and add the outcomes. What is the probability distribution of the sum? Simple counting gives the answer: the outcome 2 can be reached in only one way (1+1) and therefore has probability 1/36. The outcome 3 can be reached in two ways (1+2 and 2+1) and therefore has probability of 2/36, etc. This results in the probability distribution shown in Fig. A10. How do we calculate this distribution formally?



**Figure A9.** The probability distribution of the sum of two dice rolls.

We call the random variables corresponding to both die rolls  $X$  and  $Y$ . Their sum is a new random variable,  $Z=X+Y$ . To calculate the distribution of  $Z$ , denoted  $p_Z(z)$ , we apply the discrete analog of Eq. (A.15):

$$\begin{aligned}
 p_Z(z) &= \sum_{x=1}^6 \sum_{y=1}^6 p_{Z|X,Y}(z|x,y) p_{X,Y}(x,y) \\
 &= \sum_{x=1}^6 \sum_{y=1}^6 p_{Z|X,Y}(z|x,y) p_X(x) p_Y(y) \\
 &= \sum_{x=1}^6 p_X(x) \sum_{y=1}^6 \delta(z-x-y) p_Y(y) \\
 &= \sum_{x=\max(1,z-6)}^{\min(6,z-1)} p_X(x) p_Y(z-x) \\
 &= \sum_{x=\max(1,z-6)}^{\min(6,z-1)} \frac{1}{6} \cdot \frac{1}{6} \\
 &= \frac{1}{36} (\min(6, z-1) - \max(1, z-6) + 1),
 \end{aligned}$$

where in the fourth equality we used the property of the delta function, Eq. (A.7), as well as the condition that for  $p_Y(y)$  to be nonzero, we must have  $1 \leq y \leq 6$ , therefore  $1 \leq z-x \leq 6$ , and therefore  $z-6 \leq x \leq z-1$ .

The same logic can be applied to a continuous distribution. Let  $X$  and  $Y$  be independent continuous variables with respective pdfs  $p_X(x)$  and  $p_Y(y)$ . We define a new variable  $Z=f(X, Y)$ , with  $f$  any function, and denote by  $f_x^{-1}$  the inverse function of  $f$  for given  $x$ :  $Y=f_x^{-1}(Z)$ . (Such an inverse function does not always exist, but in the examples in this book, it does.) Then the distribution of  $Z$  is

$$\begin{aligned}
p_Z(z) &= \iint p_{Z|X,Y}(z|x,y) p_X(x) p_Y(y) dx dy \\
&= \iint \delta(z - f(x,y)) p_X(x) p_Y(y) dx dy \\
&= \int dx p_X(x) \int dt \delta(z - t) p_Y(f_x^{-1}(t)) \left| \frac{df_x^{-1}}{dt} \right| \\
&= \int dx p_X(x) p_Y(f_x^{-1}(z)) \left| \frac{df_x^{-1}}{dz} \right|,
\end{aligned} \tag{A.16}$$

where from the second to the third line we have made the transformation of variables  $y=f_x^{-1}(t)$  (compare Section 11.2). One can think of the delta function as selecting a region of  $N$ -dimensional space – namely all points that map onto  $y$  – and of the integral as the total probability under  $p_X$  in that region.

Exercise: If  $X$  and  $Y$  are independent and have a uniform distribution on  $[0,1]$ , compute the distribution of  $Z=X+Y$ .

Exercise: Prove the second property of normally distributed variables from the Box in Section 2.4.

So far, we have computed the distribution of a sum random variable. We can also use Eq. (A.16) to compute the distribution of nonlinear combinations of random variables, such as a product or quotient. The distribution of the product (or quotient) of two variables is not equal to the product (or quotient) of their distributions, and often very far from it.

Exercise: If  $X$  and  $Y$  are independent and have a uniform distribution on  $[0,1]$ , show that the product random variable  $Z=XY$  has distribution  $p_Z(z) = -\log z$ . Verify that this distribution is normalized even though the density at 0 is infinity. This example illustrates how the distribution of a product is wildly different from the distributions of each of the factors.

Exercise: If  $X$  and  $Y$  are independent standard normal variables, show that the quotient random variable  $Z=Y/X$  has a Cauchy distribution, i.e.,  $p_Z(z)=1/(\pi(1+z^2))$ .

## 12 Drawing from a probability distribution

In probabilistic modeling, we often have to draw random numbers according to a specified probability distribution. These draws are also called *samples*. Drawing random numbers is by no means trivial, but fortunately, most software packages have built-in random number generators for the most common probability distributions. We can then use these functions to custom-write code for drawing from probability distributions that are not pre-programmed.

### 12.1 Drawing from a uniform distribution

The uniform distribution on  $[0,1]$  is the easiest distribution to draw from, and virtually every programming language and mathematics software package has a built-in command for it. In Matlab, for instance, the “rand” command will produce draws from a uniform distribution on  $[0,1]$ . It is straightforward to use this command to draw from a uniform distribution on any interval.

We can also use the rand command to draw from discrete probability distributions. The simplest case is drawing a value of a binary random variable. Without loss of generality, we denote its two possible values 0 and 1. Then, by rounding the output of rand, we get one of these values, each with equal probability: round(rand). An equivalent way of obtaining the same result is to compare the output of rand to 0.5: rand>0.5. The output of this operation is a boolean, which has numerical values 0 and 1.

Exercise: How would one draw from a binary probability distribution in which one value has probability 0.7 and the other 0.3?

In Matlab, drawing random numbers can be vectorized. For example, rand(1,Ntrials) will produce a row vector of length Ntrials, each entry of which is independently drawn from a uniform distribution on  $[0,1]$ . Whenever vectorization is possible, it is strongly recommended, as it will speed up the runtime of the code substantially.

### 12.2 Drawing from a normal distribution

Besides the uniform distribution, the distribution from which we will draw most often in this book is the normal distribution. In C++ and Matlab, one draws from this distribution using randn (an alternative is normrnd, which calls randn), which generates a random number drawn from a normal distribution with mean 0 and standard deviation 1. Like rand, randn can be vectorized.

Exercise: How would one use randn to draw a random number from a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ ? Use one of the properties in the Box in Section 2.4.

### 12.3 Drawing from a Poisson distribution

As we discussed in Section 9, neural spike counts are often drawn from a Poisson distribution. Matlab has the command `poissrnd` to draw from the Poisson distribution. It takes one input argument, which is the mean (or rate parameter) of the distribution. For example, `poissrnd(3.4)` will draw a random number from a Poisson distribution with mean 3.4. Note that the Poisson distribution is defined only for non-negative integers, and therefore `poissrnd` will only return non-negative integers. The argument (the mean of the distribution) must be a non-negative real number.

Caution: By adding a constant to a set of numbers drawn from a Gaussian distribution, or by multiplying them by a constant, one obtains a new set drawn from another Gaussian distribution. If one were to apply either operation to a sample from a Poisson distribution, one does not obtain a proper sample from another Poisson distribution.

In Matlab, one can simultaneously draw from Poisson distributions with different means. This is done by specifying the vector of means as an argument. For example, `poissrnd([3.4 0.1 10])` will generate three numbers, drawn from Poisson distributions with means 3.4, 0.1, and 10, respectively.

### 12.4 Drawing from other built-in distributions

While drawing from the uniform, Gaussian, and Poisson distribution will cover nearly all examples in this book, Matlab has built-in functions to implement random draws from many other distributions. Examples are the gamma distribution (`gmrnd`), the binomial distribution (`binornd`), and the exponential distribution (`exprnd`). For some other distributions, Matlab users have written freely available code.

### 12.5 Drawing from an arbitrary one-dimensional distribution

Every now and then, one has to draw from a distribution that is not built in and that nobody else has written suitable code for. To draw from an arbitrary probability distribution  $p(x)$ , first compute (analytically or numerically) its cumulative distribution function  $P(x)$ . The cumulative distribution function takes values between 0 and 1, and is steepest where  $p(x)$  is highest. Now use `rand` to draw a number from a uniform distribution on the interval  $[0,1]$ . Then find the value of  $x$  for which  $P(x)$  is closest to the drawn random number (or more sophisticatedly, use interpolation).

Example: Suppose one wants to draw from an exponential distribution on the positive real line,  $p(x)=\lambda e^{-\lambda x}$ . Then the cumulative distribution function is

$$P(x) = \int_0^x \lambda e^{-\lambda t} dt = -e^{-\lambda t} \Big|_0^x = 1 - e^{-\lambda x}.$$

$\log(1-c)$ . Therefore, the command `-log(1-rand)/lambda` will produce a sample from the exponential distribution with parameter `lambda`. Since  $1-c$  follows a uniform distribution on  $[0,1]$  when  $c$  does, this command is equivalent to `-log(rand)/lambda`. This is exactly how the built-in command `exprnd` is written (try “edit `exprnd`”)!

Exercise: A horribly unfair die produces the rolls 1 to 6 with probabilities (0.23, 0.26, 0.04, 0.26, 0.18, 0.03). Visualize how one can generate random rolls from this die. Write a Matlab script to do this. Verify that your code works as desired by drawing a million samples and comparing their histogram with the desired distribution.

The recipe for one-dimensional distributions also works for higher-dimensional distributions, as long as the distribution is discretized. For example, a two-dimensional distribution is typically discretized by evaluating it on a two-dimensional rectangular grid of points. First put all the points of the grid into a vector, and similarly for the distribution. Then apply the same procedure as for a discrete one-dimensional distribution. Then look up which two-dimensional point the vector entry obtained corresponds to.

Exercise: Matlab has the function “peaks” defined on a two-dimensional grid. First turn this into a probability mass function by normalization. Then generate a million samples from this distribution. Compare their histogram with the desired distribution.

More sophisticated approximate methods exist for drawing from arbitrary probability distributions in any dimension, most notably Markov chain Monte Carlo (MCMC) algorithms. These are well-documented in other books and we will not discuss them here.

## 12.6 Drawing from a mixture of distributions

Recall that a mixture of two distributions  $p_1(x)$  and  $p_2(x)$  is of the form

$$p(x) = w p_1(x) + (1-w) p_2(x),$$

with  $w$  a number between 0 and 1. One can draw from such a mixture distribution in two steps: first randomly pick the mixture component to draw from, according to the mixture proportions (use `rand`), then draw from the selected mixture component. This can be generalized to drawing from a mixture of any number of distributions using the general procedure from Section 12.5 to randomly select a mixture component. This general procedure can also be applied to the full distribution  $p(x)$ , but would be slower.

## 12.7 Proof of the Cramer-Rao bound (Advanced)

We will first prove an identity that will come in helpful:

$$\begin{aligned}
\left\langle \frac{\partial}{\partial s} \log p(\mathbf{r} | s) \right\rangle &= \int \left( \frac{\partial}{\partial s} \log p(\mathbf{r} | s) \right) p(\mathbf{r} | s) d\mathbf{r} \\
&= \int \left( \frac{1}{p(\mathbf{r} | s)} \frac{\partial}{\partial s} p(\mathbf{r} | s) \right) p(\mathbf{r} | s) d\mathbf{r} \\
&= \int \frac{\partial}{\partial s} p(\mathbf{r} | s) d\mathbf{r} \\
&= \frac{\partial}{\partial s} \int p(\mathbf{r} | s) d\mathbf{r} \\
&= \frac{\partial}{\partial s} 1 \\
&= 0.
\end{aligned} \tag{A.17}$$

Now we are ready to derive the Cramér-Rao bound. To do this, we invoke the Cauchy-Schwarz inequality, which states that for two random variables  $X$  and  $Y$ , the following holds:

$$\text{cov}(X, Y)^2 \leq \text{var}(X) \text{var}(Y). \tag{A.18}$$

In our case, we use  $X = \hat{s}(\mathbf{r})$  and  $Y = \frac{\partial}{\partial s} \log p(\mathbf{r} | s)$ . Then we can calculate the covariance in the left-hand side, using the fact that  $\langle Y \rangle = 0$  (Eq. (A.17)):

$$\begin{aligned}
\text{cov}(X, Y) &= \langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle \\
&= \langle (X - \langle X \rangle)Y \rangle \\
&= \langle XY \rangle - \langle X \rangle \langle Y \rangle \\
&= \langle XY \rangle \\
&= \left\langle \hat{s}(\mathbf{r}) \frac{\partial}{\partial s} \log p(\mathbf{r} | s) \right\rangle
\end{aligned} \tag{A.19}$$

Now we can use the helpful result we derived earlier, Eq. (A.17), and evaluate further:

$$\begin{aligned}
\text{cov}(X, Y) &= \left\langle \hat{s}(\mathbf{r}) \frac{\partial}{\partial s} \log p(\mathbf{r} | s) \right\rangle \\
&= \int \hat{s}(\mathbf{r}) \left( \frac{\partial}{\partial s} \log p(\mathbf{r} | s) \right) p(\mathbf{r} | s) d\mathbf{r} \\
&= \int \hat{s}(\mathbf{r}) \frac{1}{p(\mathbf{r} | s)} \left( \frac{\partial}{\partial s} p(\mathbf{r} | s) \right) p(\mathbf{r} | s) d\mathbf{r} \\
&= \int \hat{s}(\mathbf{r}) \left( \frac{\partial}{\partial s} p(\mathbf{r} | s) \right) d\mathbf{r} \\
&= \frac{\partial}{\partial s} \int \hat{s}(\mathbf{r}) p(\mathbf{r} | s) d\mathbf{r} \\
&= \frac{\partial}{\partial s} \langle \hat{s} \rangle.
\end{aligned} \tag{A.20}$$

We use the fact that  $\hat{s}$  is unbiased,  $\langle \hat{s} \rangle = s$ , to obtain  $\text{cov}(X, Y) = 1$ . Next, we evaluate the second factor on the right-hand side of Eq. (A.18):

$$\text{var}(Y) = \langle Y^2 \rangle = \left\langle \left( \frac{\partial}{\partial s} \log p(\mathbf{r} | s) \right)^2 \right\rangle = I(s). \tag{A.21}$$

Combining Eqs. (A.20), (A.21), and (A.18), we find Eq. **Error! Reference source not found.**.

### 13 Problems

**Problem 1.** In a flower garden, 20% of flowers are tulips. Of those, one quarter are red. What is the probability that a random flower in this garden is a tulip but not red? Although it is easily possible to solve this problem using intuition, we would like you to formally apply the rules of probability.

**Problem 2.** Four players, sitting around a table, are about to play a game. To determine who starts, one person throws two dice. The sum of the two numbers determines who starts, with the counting going clockwise starting with the roller being 1 (so with a sum of 5 or 9, the roller starts). What is the probability of starting for each player?

**Problem 3.** You and I each roll a die once.

- What is the probability that one of us rolls a 6 and the other an odd number?
- What is the probability that at least one of us rolls a 6?
- What is the probability that you roll higher than me?
- What is the probability that our total is higher than 8?

**Problem 4.** Which of the following is the correct definition of conditional probability?

- a)  $P(A|B) = P(A,B)P(B)$
- b)  $P(A|B) = P(B)/P(A,B)$
- c)  $P(A|B) = P(A,B)/P(B)$

**Problem 5.** Let  $X$ ,  $Y$ , and  $Z$  be random variables with possible values  $x$ ,  $y$ , and  $z$ . Prove the following.

- a) Conditional marginal:  $p(x|z) = \sum_y p(x|y,z) p(y|z).$
- b) Conditional Bayes' rule:  $p(x|y,z) = \frac{p(y|x,z) p(x|z)}{p(y|z)}.$

**Problem 6.** You are a student in a class of 30.

- a) What is the probability that a particular classmate shares your birthday?
- b) What is the probability that any classmate shares your birthday?
- c) What is the probability that any two students share the same birthday?

**Problem 7.** You and I alternately toss a coin. You start. If you toss heads, you win instantly. If you toss tails, it is my turn. If I toss tails, I win instantly. If I toss heads, it's your turn again. This repeats until one of us has won. What is your probability of winning this game?

**Problem 8. (Monty Hall problem)** You are on a game show. The host shows you three doors. Behind one of them, a prize is hidden. You choose one door. The host, who knows behind which door the prize lies, opens a remaining door that does not contain the prize. He then gives you the opportunity to switch your choice to the remaining unopened door, or stay with your original choice. Your door of choice gets opened and you receive the prize if it is there.

- a) To maximize the probability of receiving the prize, what should you do?
- b) If there are  $n$  doors and the host opens  $m$  of them (where  $m < n-1$ ), what is the probability of receiving the prize under the best strategy?
- c) Would the answer to (a) change if the host did not know which of the two remaining doors contained a prize, but the one he opens just happens not to contain the prize? Explain.
- d) Speculate on why most people believe it does not matter whether you stay or switch.

**Problem 9 (Requires familiarity with integration).** The probability density function of a Gaussian random variable  $X$  is  $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ . Show that the mean and variance of this random variable, defined as are equal to  $\mu$  and  $\sigma^2$ , respectively.

## Lab problems

### Problem L1.

- a) Matlab has the function “humps” defined on  $[0,1]$  in steps of 0.05. First turn this into a probability mass function by normalizing. Then generate a million samples from this distribution. Compare their histogram with the desired distribution.
- b) In Matlab, draw random samples from the probability distribution whose unnormalized form is  $p(x) \propto x^2 \exp(-x)$ . Verify that your code works as desired by drawing a million samples and comparing their histogram with the desired distribution.

Write code to draw a random number from a Von Mises distribution with mean  $\mu$  and concentration parameter  $\kappa$ . Verify that your code works as desired by drawing a million samples and comparing their histogram with the desired distribution.