# Bayesian models of cognition

Thomas L. Griffiths, Charles Kemp and Joshua B. Tenenbaum

## 1 Introduction

For over 200 years, philosophers and mathematicians have been using probability theory to describe human cognition. While the theory of probabilities was first developed as a means of analyzing games of chance, it quickly took on a larger and deeper significance as a formal account of how rational agents should reason in situations of uncertainty (Gigerenzer et al., 1989; Hacking, 1975). Our goal in this chapter is to illustrate the kinds of computational models of cognition that we can build if we assume that human learning and inference approximately follow the principles of Bayesian probabilistic inference, and to explain some of the mathematical ideas and techniques underlying those models.

Bayesian models are becoming increasingly prominent across a broad spectrum of the cognitive sciences. Just in the last few years, Bayesian models have addressed animal learning (Courville, Daw, & Touretzky, 2006), human inductive learning and generalization (Tenenbaum, Griffiths, & Kemp, 2006), visual scene perception (Yuille & Kersten, 2006), motor control (Kording & Wolpert, 2006), semantic memory (Steyvers, Griffiths, & Dennis, 2006), language processing and acquisition (Chater & Manning, 2006; Xu & Tenenbaum, in press), symbolic reasoning (Oaksford & Chater, 2001), causal learning and inference (Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Griffiths & Tenenbaum, 2005, 2007a), and social cognition (Baker, Tenenbaum, & Saxe, 2007), among other topics. Behind these different research programs is a shared sense of which are the most compelling computational questions that we can ask about the human mind. To us, the big question is this: how does the human mind go beyond the data of experience? In other words, how does the mind build rich, abstract, veridical models of the world given only the sparse and noisy data that we observe through our senses? This is by no means the only computationally interesting aspect of cognition that we can study, but it is surely one of the most central, and also one of the most challenging. It is a version of the classic problem of induction, which is as old as recorded Western thought and is the source of many deep problems and debates in modern philosophy of knowledge and philosophy of science. It is also at the heart of the difficulty in building machines with anything resembling human-like intelligence.

The Bayesian framework for probabilistic inference provides a general approach to understanding how problems of induction can be solved in principle, and perhaps how they might be solved in the human mind. Let us give a few examples. Vision researchers are interested in how the mind infers the intrinsic properties of a object (e.g., its color or shape) as well as its role in a visual scene (e.g., its spatial relation to other objects or its trajectory of motion). These features are severely underdetermined by the available image

data. For instance, the spectrum of light wavelengths reflected from an object's surface into the observer's eye is a product of two unknown spectra: the surface's color spectrum and the spectrum of the light illuminating the scene. Solving the problem of "color constancy" – inferring the object's color given only the light reflected from it, under any conditions of illumination – is akin to solving the equation $y = a \times b$ for $a$ given $y$, without knowing $b$. No deductive or certain inference is possible. At best we can make a reasonable guess, based on some expectations about which values of $a$ and $b$ are more likely a priori. This inference can be formalized in a Bayesian framework (Brainard & Freeman, 1997), and it can be solved reasonably well given prior probability distributions for natural surface reflectances and illumination spectra.

The problems of core interest in other areas of cognitive science may seem very different from the problem of color constancy in vision, and they are different in important ways, but they are also deeply similar. For instance, language researchers want to understand how people recognize words so quickly and so accurately from noisy speech, how we parse a sequence of words into a hierarchical representation of the utterance's syntactic phrase structure, or how a child infers the rules of grammar – an infinite generative system – from observing only a finite and rather limited set of grammatical sentences, mixed with more than a few incomplete or ungrammatical utterances. In each of these cases, the available data severely underconstrain the inferences that people make, and the best the mind can do is to make a good guess, guided – from a Bayesian standpoint – by prior probabilities about which world structures are most likely a priori. Knowledge of a language – its lexicon, its syntax and its pragmatic tendencies of use – provides probabilistic constraints and preferences on which words are most likely to be heard in a given context, or which syntactic parse trees a listener should consider in processing a sequence of spoken words. More abstract knowledge, in a sense what linguists have referred to as "universal grammar" (Chomsky, 1988), can generate priors on possible rules of grammar that guide a child in solving the problem of induction in language acquisition. Chater & Manning (2006) survey Bayesian models of language from this perspective.

Our focus in this chapter will be on problems in higher-level cognition: inferring causal structure from patterns of statistical correlation, learning about categories and hidden properties of objects, and learning the meanings of words. This focus is partly a pragmatic choice, as these topics are the subject of our own research and hence we know them best. But there are also deeper reasons for this choice. Learning about causal relations, category structures, or the properties or names of objects are problems that are very close to the classic problems of induction that have been much discussed and puzzled over in the Western philosophical tradition. Showing how Bayesian methods can apply to these problems thus illustrates clearly their importance in understanding phenomena of induction more generally. These are also cases where the important mathematical principles and techniques of Bayesian statistics can be applied in a relatively straightforward way. They thus provide an ideal training ground for readers new to Bayesian modeling.

Beyond their value as a general framework for solving problems of induction, Bayesian approaches can make several contributions to the enterprise of modeling human cognition. First, they provide a link between human cognition and the normative prescriptions of a theory of rational inductive inference. This connection eliminates many of the degrees of freedom from a cognitive model: Bayesian principles dictate how rational agents should

update their beliefs in light of new data, based on a set of assumptions about the nature of the problem at hand and the prior knowledge possessed by the agents. Bayesian models are typically formulated at Marr's (1982) level of "computational theory", rather than the algorithmic or process level that characterizes more traditional cognitive modeling paradigms, as described in other chapters of this volume: connectionist networks (see the chapter by McClelland), exemplar-based models (see the chapter by Logan), production systems and other cognitive architectures (see the chapter by Taatgen and Anderson), or dynamical systems (see the chapter by Shoener). Algorithmic or process accounts may be more satisfying in mechanistic terms, but they may also require assumptions about human processing mechanisms that are no longer needed when we assume that cognition is an approximately optimal response to the uncertainty and structure present in natural tasks and environments (Anderson, 1990). Finding effective computational models of human cognition then becomes a process of considering how best to characterize the computational problems that people face and the logic by which those computations can be carried out (Marr, 1982).

This focus implies certain limits on the phenomena that are valuable to study within a Bayesian paradigm. Some phenomena will surely be more satisfying to address at an algorithmic or neurocomputational level. For example, that a certain behavior takes people an average of 450 milliseconds to produce, measured from the onset of a visual stimulus, or that this reaction time increases when the stimulus is moved to a different part of the visual field or decreases when the same information content is presented auditorily, are not facts that a rational computational theory is likely to predict. Moreover, not all computational-level models of cognition may have a place for Bayesian analysis. Only problems of inductive inference, or problems that contain an inductive component, are naturally expressed in Bayesian terms. Deductive reasoning, planning, or problem solving, for instance, are not traditionally thought of in this way. However, Bayesian principles are increasingly coming to be seen as relevant to many cognitive capacities, even those not traditionally seen in statistical terms (Anderson, 1990; Oaksford & Chater, 2001), due to the need for people to make inherently underconstrained inferences from impoverished data in an uncertain world.

A second key contribution of probabilistic models of cognition is the opportunity for greater communication with other fields studying computational principles of learning and inference. These connections make it a uniquely exciting time to be exploring probabilistic models of the mind. The fields of statistics, machine learning, and artificial intelligence have recently developed powerful tools for defining and working with complex probabilistic models that go far beyond the simple scenarios studied in classical probability theory; we will present a taste of both the simplest models and more complex frameworks here. The more complex methods can support multiple hierarchically organized layers of inference, structured representations of abstract knowledge, and approximate methods of evaluation that can be applied efficiently to data sets with many thousands of entities. For the first time, we now have practical methods for developing computational models of human cognition that are based on sound probabilistic principles and that can also capture something of the richness and complexity of everyday thinking, reasoning and learning.

We can also exploit fertile analogies between specific learning and inference problems in the study of human cognition and in these other disciplines, to develop new cognitive models or new tools for working with existing models. We will discuss some of these relationships in this chapter, but there are many other cases. For example, prototype

and exemplar models of categorization (Reed, 1972; Medin & Schaffer, 1978; Nosofsky, 1986) can both be seen as rational solutions to a standard classification task in statistical pattern recognition: an object is generated from one of several probability distributions (or "categories") over the space of possible objects, and the goal is to infer which distribution is most likely to have generated that object (Duda, Hart, & Stork, 2000). In rational probabilistic terms, these methods differ only in how these category-specific probability distributions are represented and estimated (Ashby & Alfonso-Reese, 1995; Nosofsky, 1998).

Finally, probabilistic models can be used to advance and perhaps resolve some of the great theoretical debates that divide traditional approaches to cognitive science. The history of computational models of cognition exhibits an enduring tension between models that emphasize symbolic representations and deductive inference, such as first order logic or phrase structure grammars, and models that emphasize continuous representations and statistical learning, such as connectionist networks or other associative systems. Probabilistic models can be defined with either symbolic or continuous representations, or hybrids of both, and help to illustrate how statistical learning can be combined with symbolic structure. More generally, we think that the most promising routes to understanding human intelligence in computational terms will involve deep interactions between these two traditionally opposing approaches, with sophisticated statistical inference machinery operating over structured symbolic knowledge representations. Contemporary probabilistic methods give us the first general-purpose set of tools for building such structured statistical models, and we will see several simple examples of these models in this chapter.

The tension between symbols and statistics is perhaps only exceeded by the tension between accounts that focus on the importance of innate, domain-specific knowledge in explaining human cognition, and accounts that focus on domain-general learning mechanisms. Again, probabilistic models provide a middle ground where both approaches can productively meet, and they suggest various routes to resolving the tensions between these approaches by combining the important insights of both. Probabilistic models highlight the role of prior knowledge in accounting for how people learn as much as they do from limited observed data, and provide a framework for explaining precisely how prior knowledge interacts with data in guiding generalization and action. They also provide a tool for exploring the kinds of knowledge that people bring to learning and reasoning tasks, allowing us to work forwards from rational analyses of tasks and environments to predictions about behavior, and to work backwards from subjects' observed behavior to viable assumptions about the knowledge they could bring to the task. Crucially, these models do not require that the prior knowledge be innate. Bayesian inference in hierarchical probabilistic models can explain how abstract prior knowledge may itself be learned from data, and then put to use to guide learning in subsequent tasks and new environments.

This chapter will discuss both the basic principles that underlie Bayesian models of cognition and several advanced techniques for probabilistic modeling and inference that have come out of recent work in computer science and statistics. Our first step is to summarize the logic of Bayesian inference which is at the heart of many probabilistic models. We then turn to a discussion of three recent innovations that make it easier to define and use probabilistic models of complex domains: graphical models, hierarchical Bayesian models, and Markov chain Monte Carlo. We illustrate the central ideas behind each of these techniques by considering a detailed cognitive modeling application, drawn from causal learning, property

induction, and language modeling respectively.

## 2 The basics of Bayesian inference

Many aspects of cognition can be formulated as solutions to problems of induction. Given some observed data about the world, the mind draws conclusions about the underlying process or structure that gave rise to these data, and then uses that knowledge to make predictive judgments about new cases. Bayesian inference is a rational engine for solving such problems within a probabilistic framework, and consequently is the heart of most probabilistic models of cognition.

### *2.1 Bayes' rule*

Bayesian inference grows out of a simple formula known as *Bayes' rule* (Bayes, 1763/1958). When stated in terms of abstract random variables, Bayes' rule is no more than an elementary result of probability theory. Assume we have two random variables, $A$ and $B$.[1] One of the principles of probability theory (sometimes called the *chain rule*) allows us to write the *joint probability* of these two variables taking on particular values $a$ and $b$, $P(a, b)$, as the product of the *conditional probability* that $A$ will take on value $a$ given $B$ takes on value $b$, $P(a|b)$, and the *marginal probability* that $B$ takes on value $b$, $P(b)$. Thus, we have

$$P(a, b) = P(a|b)P(b). \tag{1}$$

There was nothing special about the choice of $A$ rather than $B$ in factorizing the joint probability in this way, so we can also write

$$P(a, b) = P(b|a)P(a). \tag{2}$$

It follows from Equations 1 and 2 that $P(a|b)P(b) = P(b|a)P(a)$, which can be rearranged to give

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)}. \tag{3}$$

This expression is Bayes' rule, which indicates how we can compute the conditional probability of $b$ given $a$ from the conditional probability of $a$ given $b$.

While Equation 3 seems relatively innocuous, Bayes' rule gets its strength, and its notoriety, when we make some assumptions about the variables we are considering and the meaning of probability. Assume that we have an agent who is attempting to infer the process that was responsible for generating some data, $d$. Let $h$ be a hypothesis about this process. We will assume that the agent uses probabilities to represent degrees of belief in $h$ and various alternative hypotheses $h'$. Let $P(h)$ indicate the probability that the agent ascribes to $h$ being the true generating process, prior to (or independent of) seeing the data $d$. This quantity is known as the *prior probability*. How should that agent change his beliefs in light of the evidence provided by $d$? To answer this question, we need a procedure for

[1] We will use uppercase letters to indicate random variables, and matching lowercase variables to indicate the values those variables take on. When defining probability distributions, the random variables will remain implicit. For example, $P(a)$ refers to the probability that the variable $A$ takes on the value $a$, which could also be written $P(A = a)$. We will write joint probabilities in the form $P(a, b)$. Other notations for joint probabilities include $P(a\&b)$ and $P(a \cap b)$.

computing the *posterior probability*, $P(h|d)$, or the degree of belief in $h$ conditioned on the observation of $d$.

Bayes' rule provides just such a procedure, if we treat both the hypotheses that agents entertain and the data that they observe as random variables, so that the rules of probabilistic inference can be applied to relate them. Replacing $a$ with $d$ and $b$ with $h$ in Equation 3 gives

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}, \tag{4}$$

the form in which Bayes' rule is most commonly presented in analyses of learning or induction. The posterior probability is proportional to the product of the prior probability and another term $P(d|h)$, the probability of the data given the hypothesis, commonly known as the *likelihood*. Likelihoods are the critical bridge from priors to posteriors, re-weighting each hypothesis by how well it predicts the observed data.

In addition to telling us how to compute with conditional probabilities, probability theory allows us to compute the probability distribution associated with a single variable (known as the *marginal probability*) by summing over other variables in a joint distribution: e.g., $P(b) = \sum_a P(a, b)$. This is known as *marginalization*. Using this principle, we can rewrite Equation 4 as

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')}, \tag{5}$$

where $\mathcal{H}$ is the set of all hypotheses considered by the agent, sometimes referred to as the *hypothesis space*. This formulation of Bayes' rule makes it clear that the posterior probability of $h$ is directly proportional to the product of its prior probability and likelihood, relative to the sum of these same scores – products of priors and likelihoods – for all alternative hypotheses under consideration. The sum in the denominator of Equation 5 ensures that the resulting posterior probabilities are normalized to sum to one.

A simple example may help to illustrate the interaction between priors and likelihoods in determining posterior probabilities. Consider three possible medical conditions that could be posited to explain why a friend is coughing (the observed data $d$): $h_1$ = "cold", $h_2$ = "lung cancer", $h_3$ = "stomach flu". The first hypothesis seems intuitively to be the best of the three, for reasons that Bayes' rule makes clear. The probability of coughing given that one has lung cancer, $P(d|h_2)$ is high, but the prior probability of having lung cancer $P(h_2)$ is low. Hence the posterior probability of lung cancer $P(h_2|d)$ is low, because it is proportional to the product of these two terms. Conversely, the prior probability of having stomach flu $P(h_3)$ is relatively high (as medical conditions go), but its likelihood $P(d|h_3)$, the probability of coughing given that one has stomach flu, is relatively low. So again, the posterior probability of stomach flu, $P(h_3|d)$, will be relatively low. Only for hypothesis $h_1$ are both the prior $P(h_1)$ and the likelihood $P(d|h_1)$ relatively high: colds are fairly common medical conditions, and coughing is a symptom frequently found in people who have colds. Hence the posterior probability $P(h_1|d)$ of having a cold given that one is coughing is substantially higher than the posteriors for the competing alternative hypotheses – each of which is less likely for a different sort of reason.

*2.2 Comparing hypotheses*

The mathematics of Bayesian inference is most easily introduced in the context of comparing two simple hypotheses. For example, imagine that you are told that a box contains two coins: one that produces heads 50% of the time, and one that produces heads 90% of the time. You choose a coin, and then flip it ten times, producing the sequence `HHHHHHHHHH`. Which coin did you pick? How would your beliefs change if you had obtained `HHTHTHTTHT` instead?

To formalize this problem in Bayesian terms, we need to identify the hypothesis space, $\mathcal{H}$, the prior probability of each hypothesis, $P(h)$, and the probability of the data under each hypothesis, $P(d|h)$. We have two coins, and thus two hypotheses. If we use $\theta$ to denote the probability that a coin produces heads, then $h_0$ is the hypothesis that $\theta = 0.5$, and $h_1$ is the hypothesis that $\theta = 0.9$. Since we have no reason to believe that one coin is more likely to be picked than the other, it is reasonable to assume equal prior probabilities: $P(h_0) = P(h_1) = 0.5$. The probability of a particular sequence of coinflips containing $N_H$ heads and $N_T$ tails being generated by a coin which produces heads with probability $\theta$ is

$$P(d|\theta) = \theta^{N_H}(1-\theta)^{N_T}. \tag{6}$$

Formally, this expression follows from assuming that each flip is drawn independently from a Bernoulli distribution with parameter $\theta$; less formally, that heads occurs with probability $\theta$ and tails with probability $1-\theta$ on each flip. The likelihoods associated with $h_0$ and $h_1$ can thus be obtained by substituting the appropriate value of $\theta$ into Equation 6.

We can take the priors and likelihoods defined in the previous paragraph, and plug them directly into Equation 5 to compute the posterior probabilities for both hypotheses, $P(h_0|d)$ and $P(h_1|d)$. However, when we have just two hypotheses it is often easier to work with the *posterior odds*, or the ratio of these two posterior probabilities. The posterior odds in favor of $h_1$ is

$$\frac{P(h_1|d)}{P(h_0|d)} = \frac{P(d|h_1)}{P(d|h_0)}\frac{P(h_1)}{P(h_0)}, \tag{7}$$

where we have used the fact that the denominator of Equation 4 or 5 is constant over all hypotheses. The first and second terms on the right hand side are called the *likelihood ratio* and the *prior odds* respectively. We can use Equation 7 (and the priors and likelihoods defined above) to compute the posterior odds of our two hypotheses for any observed sequence of heads and tails: for the sequence `HHHHHHHHHH`, the odds are approximately 357:1 in favor of $h_1$; for the sequence `HHTHTHTTHT`, approximately 165:1 in favor of $h_0$.

The form of Equation 7 helps to clarify how prior knowledge and new data are combined in Bayesian inference. The two terms on the right hand side each express the influence of one of these factors: the prior odds are determined entirely by the prior beliefs of the agent, while the likelihood ratio expresses how these odds should be modified in light of the data $d$. This relationship is made even more transparent if we examine the expression for the log posterior odds,

$$\log\frac{P(h_1|d)}{P(h_0|d)} = \log\frac{P(d|h_1)}{P(d|h_0)} + \log\frac{P(h_1)}{P(h_0)} \tag{8}$$

in which the extent to which one should favor $h_1$ over $h_0$ reduces to an additive combination of a term reflecting prior beliefs (the log prior odds) and a term reflecting the contribution

of the data (the log likelihood ratio). Based upon this decomposition, the log likelihood ratio in favor of $h_1$ is often used as a measure of the evidence that $d$ provides for $h_1$.

*2.3 Parameter estimation*

The analysis outlined above for two simple hypotheses generalizes naturally to any finite set, although posterior odds may be less useful when there are multiple alternatives to be considered. Bayesian inference can also be applied in contexts where there are (uncountably) infinitely many hypotheses to evaluate – a situation that arises often. For example, instead of choosing between just two possible values for the probability $\theta$ that a coin produces heads, we could consider any real value of $\theta$ between 0 and 1. What then should we infer about the value of $\theta$ from a sequence such as HHHHHHHHHH?

Under one classical approach, inferring $\theta$ is treated as a problem of estimating a fixed parameter of a probabilistic model, to which the standard solution is *maximum-likelihood* estimation (see, e.g., Rice, 1995). Maximum-likelihood estimation is simple and often sensible, but can also be problematic – particularly as a way to think about human inference. Our coinflipping example illustrates some of these problems. The maximum-likelihood estimate of $\theta$ is the value $\hat{\theta}$ that maximizes the probability of the data as given in Equation 6. It is straightforward to show that $\hat{\theta} = \frac{N_H}{N_H+N_T}$, which gives $\hat{\theta} = 1.0$ for the sequence HHHHHHHHHH.

It should be immediately clear that the single value of $\theta$ which maximizes the probability of the data might not provide the best basis for making predictions about future data. Inferring that $\theta$ is exactly 1 after seeing the sequence HHHHHHHHHH implies that we should predict that the coin will never produce tails. This might seem reasonable after observing a long sequence consisting solely of heads, but the same conclusion follows for an all-heads sequences of *any* length (because $N_T$ is always 0, so $\frac{N_H}{N_H+N_T}$ is always 1). Would you really predict that a coin would produce only heads after seeing it produce a head on just one or two flips?

A second problem with maximum-likelihood estimation is that it does not take into account other knowledge that we might have about $\theta$. This is largely by design: maximum-likelihood estimation and other classical statistical techniques have historically been promoted as "objective" procedures that do not require prior probabilities, which were seen as inherently and irremediably subjective. While such a goal of objectivity might be desirable in certain scientific contexts, cognitive agents typically do have access to relevant and powerful prior knowledge, and they use that knowledge to make stronger inferences from sparse and ambiguous data than could be rationally supported by the data alone. For example, given the sequence HHH produced by flipping an apparently normal, randomly chosen coin, many people would say that the coin's probability of producing heads is nonetheless around 0.5 – perhaps because we have strong prior expectations that most coins are nearly fair.

Both of these problems are addressed by a Bayesian approach to inferring $\theta$. If we assume that $\theta$ is a random variable, then we can apply Bayes' rule to obtain

$$p(\theta|d) = \frac{P(d|\theta)p(\theta)}{P(d)}, \tag{9}$$

where

$$P(d) = \int_0^1 P(d|\theta)p(\theta)\, d\theta. \tag{10}$$

The key difference from Bayesian inference with finitely many hypotheses is that our beliefs about the hypotheses (both priors and posteriors) are now characterized by *probability densities* (notated by a lowercase "p") rather than probabilities strictly speaking, and the sum over hypotheses becomes an integral.

The posterior distribution over $\theta$ contains more information than a single point estimate: it indicates not just which values of $\theta$ are probable, but also how much uncertainty there is about those values. Collapsing this distribution down to a single number discards information, so Bayesians prefer to maintain distributions wherever possible (this attitude is similar to Marr's (1982, p. 106) "principle of least commitment"). However, there are two methods that are commonly used to obtain a point estimate from a posterior distribution. The first method is *maximum a posteriori (MAP)* estimation: choosing the value of $\theta$ that maximizes the posterior probability, as given by Equation 9. The second method is computing the *posterior mean* of the quantity in question: a weighted average of all possible values of the quantity, where the weights are given by the posterior distribution. For example, the posterior mean value of the coin weight $\theta$ is computed as follows:

$$\bar{\theta} = \int_0^1 \theta \, p(\theta|d) \, d\theta. \tag{11}$$

In the case of coinflipping, the posterior mean also corresponds to the *posterior predictive distribution*: the probability that the next toss of the coin will produce heads, given the observed sequence of previous flips.

Different choices of the prior, $p(\theta)$, will lead to different inferences about the value of $\theta$. A first step might be to assume a *uniform* prior over $\theta$, with $p(\theta)$ being equal for all values of $\theta$ between 0 and 1 (more formally, $p(\theta) = 1$ for $\theta \in [0, 1]$). With this choice of $p(\theta)$ and the Bernoulli likelihood from Equation 6, Equation 9 becomes

$$p(\theta) = \frac{\theta^{N_H}(1-\theta)^{N_T}}{\int_0^1 \theta^{N_H}(1-\theta)^{N_T} \, d\theta} \tag{12}$$

where the denominator is just the integral from Equation 10. Using a little calculus to compute this integral, the posterior distribution over $\theta$ produced by a sequence $d$ with $N_H$ heads and $N_T$ tails is

$$p(\theta|d) = \frac{(N_H + N_T + 1)!}{N_H! \; N_T!} \; \theta^{N_H}(1-\theta)^{N_T}. \tag{13}$$

This is actually a distribution of a well known form: a beta distribution with parameters $N_H + 1$ and $N_T + 1$, denoted $\text{Beta}(N_H + 1, N_T + 1)$ (e.g., Pitman, 1993). Using this prior, the MAP estimate for $\theta$ is the same as the maximum-likelihood estimate, $\frac{N_H}{N_H+N_T}$, but the posterior mean is slightly different, $\frac{N_H+1}{N_H+N_T+2}$. Thus, the posterior mean is sensitive to the consideration that we might not want to put as much evidential weight on seeing a single head as on a sequence of ten heads in a row: on seeing a single head, the posterior mean predicts that the next toss will produce a head with probability $\frac{2}{3}$, while a sequence of ten heads leads to the prediction that the next toss will produce a head with probability $\frac{11}{12}$.

We can also use priors that encode stronger beliefs about the value of $\theta$. For example, we can take a $\text{Beta}(V_H + 1, V_T + 1)$ distribution for $p(\theta)$, where $V_H$ and $V_T$ are positive

integers. This distribution gives

$$p(\theta) = \frac{(V_H + V_T + 1)!}{V_H! V_T!} \theta^{V_H} (1-\theta)^{V_T} \tag{14}$$

having a mean at $\frac{V_H+1}{V_H+V_T+2}$, and gradually becoming more concentrated around that mean as $V_H + V_T$ becomes large. For instance, taking $V_H = V_T = 1000$ would give a distribution that strongly favors values of $\theta$ close to 0.5. Using such a prior with the Bernoulli likelihood from Equation 6 and applying the same kind of calculations as above, we obtain the posterior distribution

$$p(\theta|d) = \frac{(N_H + N_T + V_H + V_T + 1)!}{(N_H + V_H)!\ (N_T + V_T)!}\ \theta^{N_H+V_H} (1-\theta)^{N_T+V_T}, \tag{15}$$

which is Beta$(N_H + V_H + 1, N_T + V_T + 1)$. Under this posterior distribution, the MAP estimate of $\theta$ is $\frac{N_H+V_H}{N_H+N_T+V_H+V_T}$, and the posterior mean is $\frac{N_H+V_H+1}{N_H+N_T+V_H+V_T+2}$. Thus, if $V_H = V_T = 1000$, seeing a sequence of ten heads in a row would induce a posterior distribution over $\theta$ with a mean of $\frac{1011}{2012} \approx 0.5025$. In this case, the observed data matter hardly at all. A prior that is much weaker but still biased towards approximately fair coins might take $V_H = V_T = 5$. Then an observation of ten heads in a row would lead to a posterior mean of $\frac{16}{22} \approx .727$, significantly tilted towards heads but still closer to a fair coin than the observed data would suggest on their own. We can say that such a prior acts to "smooth" or "regularize" the observed data, damping out what might be misleading fluctuations when the data are far from the learner's initial expectations. On a larger scale, these principles of Bayesian parameter estimation with informative "smoothing" priors have been applied to a number of cognitively interesting machine-learning problems, such as Bayesian learning in neural networks (Mackay, 2003).

Our analysis of coin flipping with informative priors has two features of more general interest. First, the prior and posterior are specified using distributions of the same form (both being beta distributions). Second, the parameters of the prior, $V_H$ and $V_T$, act as "virtual examples" of heads and tails, which are simply pooled with the real examples tallied in $N_H$ and $N_T$ to produce the posterior, as if both the real and virtual examples had been observed in the same data set. These two properties are not accidental: they are characteristic of a class of priors called *conjugate priors* (e.g., Bernardo & Smith, 1994). The likelihood determines whether a conjugate prior exists for a given problem, and the form that the prior will take. The results we have given in this section exploit the fact that the beta distribution is the conjugate prior for the Bernoulli or binomial likelihood (Equation 6) – the uniform distribution on $[0, 1]$ is also a beta distribution, being Beta$(1, 1)$. Conjugate priors exist for many of the distributions commonly used in probabilistic models, such as Gaussian, Poisson, and multinomial distributions, and greatly simplify many Bayesian calculations. Using conjugate priors, posterior distributions can be computed analytically, and the interpretation of the prior as contributing virtual examples is intuitive.

While conjugate priors are elegant and practical to work with, there are also important forms of prior knowledge that they cannot express. For example, they can capture the notion of smoothness in simple linear predictive systems but not in more complex nonlinear predictors such as multilayer neural networks. Crucially for modelers interested in higher-level cognition, conjugate priors cannot capture knowledge that the causal process

generating the observed data could take on one of several qualitatively different forms. Still, they can sometimes be used to address questions of selecting models of different complexity, as we do in the next section, when the different models under consideration have the same qualitative form. A major area of current research in Bayesian statistics and machine learning focuses on building more complex models that maintain the benefits of working with conjugate priors, building on the techniques for model selection that we discuss next (e.g., Neal, 1992, 1998; Blei, Griffiths, Jordan, & Tenenbaum, 2004; Griffiths & Ghahramani, 2005).

### *2.4 Model selection*

Whether there were a finite number or not, the hypotheses that we have considered so far were relatively homogeneous, each offering a single value for the parameter $\theta$ characterizing our coin. However, many problems require comparing hypotheses that differ in their complexity. For example, the problem of inferring whether a coin is fair or biased based upon an observed sequence of heads and tails requires comparing a hypothesis that gives a single value for $\theta$ – if the coin is fair, then $\theta = 0.5$ – with a hypothesis that allows $\theta$ to take on any value between 0 and 1.

Using observed data to choose between two probabilistic models that differ in their complexity is often called the problem of *model selection* (Myung & Pitt, 1997; Myung, Forster, & Browne, 2000). One familiar statistical approach to this problem is via hypothesis testing, but this approach is often complex and counter-intuitive. In contrast, the Bayesian approach to model selection is a seamless application of the methods discussed so far. Hypotheses that differ in their complexity can be compared directly using Bayes' rule, once they are reduced to probability distributions over the observable data (see Kass & Raftery, 1995).

To illustrate this principle, assume that we have two hypotheses: $h_0$ is the hypothesis that $\theta = 0.5$, and $h_1$ is the hypothesis that $\theta$ takes a value drawn from a uniform distribution on $[0, 1]$. If we have no a priori reason to favor one hypothesis over the other, we can take $P(h_0) = P(h_1) = 0.5$. The probability of the data under $h_0$ is straightforward to compute, using Equation 6, giving $P(d|h_0) = 0.5^{N_H+N_T}$. But how should we compute the likelihood of the data under $h_1$, which does not make a commitment to a single value of $\theta$?

The solution to this problem is to compute the marginal probability of the data under $h_1$. As discussed above, given a joint distribution over a set of variables, we can always sum out variables until we obtain a distribution over just the variables that interest us. In this case, we define the joint distribution over $d$ and $\theta$ given $h_1$, and then integrate over $\theta$ to obtain

$$P(d|h_1) = \int_0^1 P(d|\theta, h_1)p(\theta|h_1)\, d\theta \tag{16}$$

where $p(\theta|h_1)$ is the distribution over $\theta$ assumed under $h_1$ – in this case, a uniform distribution over $[0, 1]$. This does not require any new concepts – it is exactly the same kind of computation as we needed to perform to compute the denominator for the posterior distribution over $\theta$ (Equation 10). Performing this computation, we obtain $P(d|h_1) = \frac{N_H!\ N_T!}{(N_H+N_T+1)!}$, where again the fact that we have a conjugate prior provides us with a neat analytic result. Having computed this likelihood, we can apply Bayes' rule just as we did for two simple

hypotheses. Figure 1a shows how the log posterior odds in favor of $h_1$ change as $N_H$ and $N_T$ vary for sequences of length 10.

The ease with which hypotheses differing in complexity can be compared using Bayes' rule conceals the fact that this is actually a very challenging problem. Complex hypotheses have more degrees of freedom that can be adapted to the data, and can thus always be made to fit the data better than simple hypotheses. For example, for any sequence of heads and tails, we can always find a value of $\theta$ that would give higher probability to that sequence than does the hypothesis that $\theta = 0.5$. It seems like a complex hypothesis would thus have an inherent unfair advantage over a simple hypothesis. The Bayesian solution to the problem of comparing hypotheses that differ in their complexity takes this into account. More degrees of freedom provide the opportunity to find a better fit to the data, but this greater flexibility also makes a worse fit possible. For example, for $d$ consisting of the sequence `HHTHTTHHHT`, $P(d|\theta, h_1)$ is greater than $P(d|h_0)$ for $\theta \in (0.5, 0.694]$, but is less than $P(d|h_0)$ outside that range. Marginalizing over $\theta$ averages these gains and losses: a more complex hypothesis will be favored only if its greater complexity consistently provides a better account of the data. To phrase this principle another way, a Bayesian learner judges the fit of a parameterized model not by how well it fits using the *best* parameter values, but by how well it fits using *randomly selected* parameters, where the parameters are drawn from a prior specified by the model ($p(\theta|h_1)$ in Equation 16) (Ghahramani, 2004). This penalization of more complex models is known as the "Bayesian Occam's razor" (Jeffreys & Berger, 1992; Mackay, 2003), and is illustrated in Figure 1b.

### *2.5 Summary*

Bayesian inference stipulates how rational learners should update their beliefs in the light of evidence. The principles behind Bayesian inference can be applied whenever we are making inferences from data, whether the hypotheses involved are discrete or continuous, or have one or more unspecified free parameters. However, developing probabilistic models that can capture the richness and complexity of human cognition requires going beyond these basic ideas. In the remainder of the chapter we will summarize several recent tools that have been developed in computer science and statistics for defining and using complex probabilistic models, and provide examples of how they can be used in modeling human cognition.

## 3. Graphical models

Our discussion of Bayesian inference above was formulated in the language of "hypotheses" and "data". However, the principles of Bayesian inference, and the idea of using probabilistic models, extend to much richer settings. In its most general form, a probabilistic model simply defines the joint distribution for a system of random variables. Representing and computing with these joint distributions becomes challenging as the number of variables grows, and their properties can be difficult to understand. Graphical models provide an efficient and intuitive framework for working with high-dimensional probability distributions, which is applicable when these distributions can be viewed as the product of smaller components defined over local subsets of variables.

A graphical model associates a probability distribution with a graph. The nodes of the graph represent the variables on which the distribution is defined, the edges between the

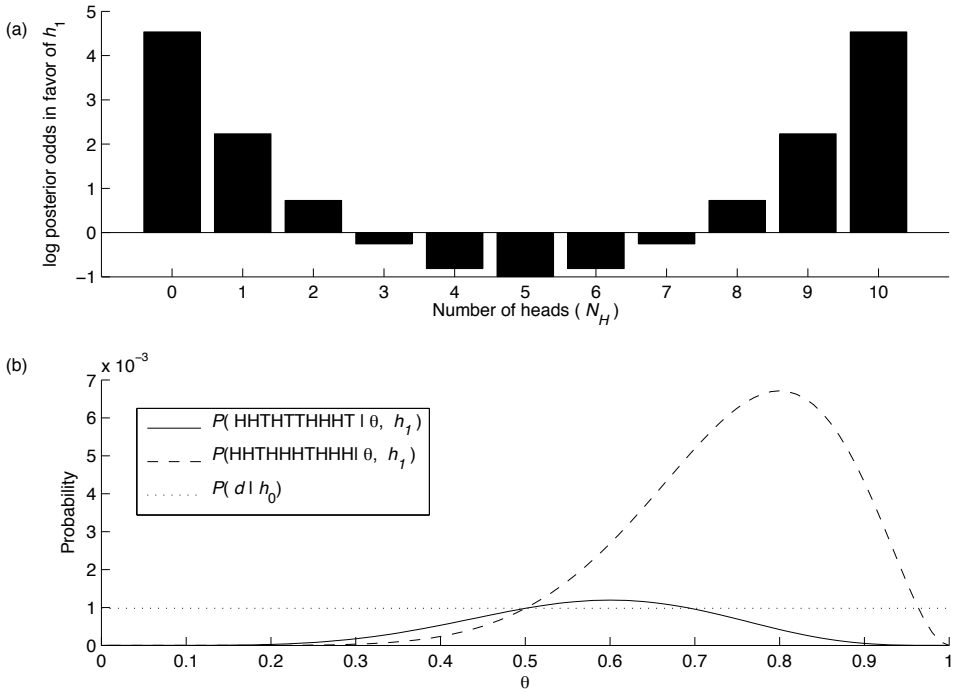


*Figure 1.* Comparing hypotheses about the weight of a coin. (a) The vertical axis shows log posterior odds in favor of $h_1$, the hypothesis that the probability of heads ($\theta$) is drawn from a uniform distribution on $[0, 1]$, over $h_0$, the hypothesis that the probability of heads is 0.5. The horizontal axis shows the number of heads, $N_H$, in a sequence of 10 flips. As $N_H$ deviates from 5, the posterior odds in favor of $h_1$ increase. (b) The posterior odds shown in (a) are computed by averaging over the values of $\theta$ with respect to the prior, $p(\theta)$, which in this case is the uniform distribution on $[0, 1]$. This averaging takes into account the fact that hypotheses with greater flexibility – such as the free-ranging $\theta$ parameter in $h_1$ – can produce both better and worse predictions, implementing an automatic "Bayesian Occam's razor". The solid line shows the probability of the sequence `HHTHTTHHHT` for different values of $\theta$, while the dotted line is the probability of any sequence of length 10 under $h_0$ (equivalent to $\theta = 0.5$). While there are some values of $\theta$ that result in a higher probability for the sequence, on average the greater flexibility of $h_1$ results in lower probabilities. Consequently, $h_0$ is favored over $h_1$ (this sequence has $N_H = 6$). In contrast, a wide range of values of $\theta$ result in higher probability for for the sequence `HHTHHHTHHH`, as shown by the dashed line. Consequently, $h_1$ is favored over $h_0$ (this sequence has $N_H = 8$).

nodes reflect their probabilistic dependencies, and a set of functions relating nodes and their neighbors in the graph are used to define a joint distribution over all of the variables based on those dependencies. There are two kinds of graphical models, differing in the nature of the edges that connect the nodes. If the edges simply indicate a dependency between variables, without specifying a direction, then the result is an *undirected graphical model*. Undirected graphical models have long been used in statistical physics, and many probabilistic neural network models, such as Boltzmann machines (Ackley, Hinton, & Sejnowski, 1985), can be interpreted as models of this kind. If the edges indicate the direction of a dependency, the result is a *directed graphical model*. Our focus here will be on directed graphical models, which are also known as Bayesian networks or Bayes nets (Pearl, 1988). Bayesian networks can often be given a causal interpretation, where an edge between two nodes indicates that one node is a direct cause of the other, which makes them particularly appealing for modeling higher-level cognition.

### *3.1 Bayesian networks*

A Bayesian network represents the probabilistic dependencies relating a set of variables. If an edge exists from node $A$ to node $B$, then $A$ is referred to as a "parent" of $B$, and $B$ is a "child" of $A$. This genealogical relation is often extended to identify the "ancestors" and "descendants" of a node. The directed graph used in a Bayesian network has one node for each random variable in the associated probability distribution, and is constrained to be *acyclic*: one can never return to the same node by following a sequence of directed edges. The edges express the probabilistic dependencies between the variables in a fashion consistent with the *Markov condition*: conditioned on its parents, each variable is independent of all other variables except its descendants (Pearl, 1988; Spirtes, Glymour, & Schienes, 1993). As a consequence of the Markov condition, any Bayesian network specifies a canonical factorization of a full joint probability distribution into the product of local conditional distributions, one for each variable conditioned on its parents. That is, for a set of variables $X_1, X_2, \ldots, X_N$, we can write $P(x_1, x_2, \ldots, x_N) = \prod_i P(x_i|\mathrm{Pa}(X_i))$ where $\mathrm{Pa}(X_i)$ is the set of parents of $X_i$.

Bayesian networks provide an intuitive representation for the structure of many probabilistic models. For example, in the previous section we discussed the problem of estimating the weight of a coin, $\theta$. One detail that we left implicit in that discussion was the assumption that successive coin flips are independent, given a value for $\theta$. This conditional independence assumption is expressed in the graphical model shown in Figure 2a, where $x_1, x_2, \ldots, x_N$ are the outcomes (heads or tails) of $N$ successive tosses. Applying the Markov condition, this structure represents the probability distribution

$$P(x_1, x_2, \ldots, x_N, \theta) = p(\theta) \prod_{i=1}^{N} P(x_i|\theta) \tag{17}$$

in which the $x_i$ are independent given the value of $\theta$. Other dependency structures are possible. For example, the flips could be generated in a Markov chain, a sequence of random variables in which each variable is independent of all of its predecessors given the variable that immediately precedes it (e.g., Norris, 1997). Using a Markov chain structure, we could represent a hypothesis space of coins that are particularly biased towards alternating or

maintaining their last outcomes, letting the parameter $\theta$ be the probability that the outcome $x_i$ takes the same value as $x_{i-1}$ (and assuming that $x_1$ is heads with probability 0.5). This distribution would correspond to the graphical model shown in Figure 2b. Applying the Markov condition, this structure represents the probability distribution

$$P(x_1, x_2, \ldots, x_N, \theta) = p(\theta)P(x_1)\prod_{i=2}^{N} P(x_i|x_{i-1}\theta), \tag{18}$$

in which each $x_i$ depends only on $x_{i-1}$, given $\theta$. More elaborate structures are also possible: any directed acyclic graph on $x_1, x_2, \ldots, x_N$ and $\theta$ corresponds to a valid set of assumptions about the dependencies among these variables.

When introducing the basic ideas behind Bayesian inference, we emphasized the fact that hypotheses correspond to different assumptions about the process that could have generated some observed data. Bayesian networks help to make this idea transparent. Every Bayesian network indicates a sequence of steps that one could follow in order to generate samples from the joint distribution over the random variables in the network. First, one samples the values of all variables with no parents in the graph. Then, one samples the variables with parents taking known values, one after another. For example, in the structure shown in Figure 2b, we would sample $\theta$ from the distribution $p(\theta)$, then sample $x_1$ from the distribution $P(x_1)$, then successively sample $x_i$ from $P(x_i|x_{i-1}, \theta)$ for $i = 2, \ldots, N$. A set of probabilistic steps that can be followed to generate the values of a set of random variables is known as a *generative model*, and the directed graph associated with a probability distribution provides an intuitive representation for the steps that are involved in such a model.

For the generative models represented by Figure 2a or 2b, we have assumed that all variables except $\theta$ are observed in each sample from the model, or each data point. More generally, generative models can include a number of steps that make reference to unobserved or *latent* variables. Introducing latent variables can lead to apparently complicated dependency structures among the observable variables. For example, in the graphical model shown in Figure 2c, a sequence of latent variables $z_1, z_2, \ldots, z_N$ influences the probability that each respective coin flip in a sequence $x_1, x_2, \ldots, x_N$ comes up heads (in conjunction with a set of parameters $\phi$). The latent variables form a Markov chain, with the value of $z_i$ depending only on the value of $z_{i-1}$ (in conjunction with the parameters $\theta$). This model, called a *hidden Markov model*, is widely used in computational linguistics, where $z_i$ might be the syntactic class (such as noun or verb) of a word, $\theta$ encodes the probability that a word of one class will appear after another (capturing simple syntactic constraints on the structure of sentences), and $\phi$ encodes the probability that each word will be generated from a particular syntactic class (e.g., Charniak, 1993; Jurafsky & Martin, 2000; Manning & Schütze, 1999). The dependencies among the latent variables induce dependencies among the observed variables – in the case of language, the constraints on transitions between syntactic classes impose constraints on which words can follow one another.

### *3.2 Representing probability distributions over propositions*

Our treatment of graphical models in the previous section – as representations of the dependency structure among variables in generative models for data – follows their

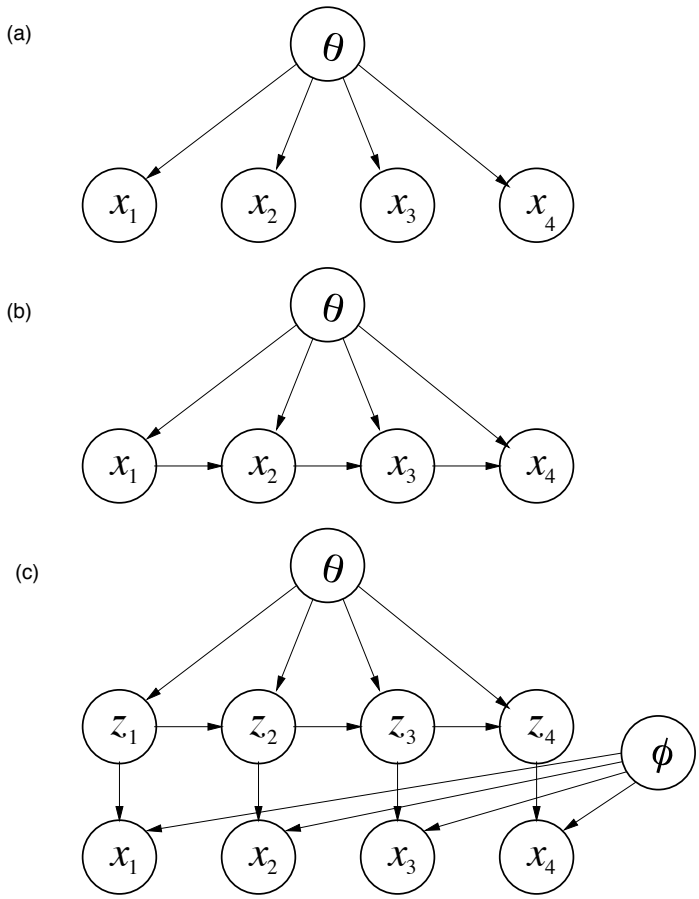


*Figure 2.* Graphical models showing different kinds of processes that could generate a sequence of coinflips. (a) Independent flips, with parameters $\theta$ determining the probability of heads. (b) A Markov chain, where the probability of heads depends on the result of the previous flip. Here the parameters $\theta$ define the probability of heads after a head and after a tail. (c) A hidden Markov model, in which the probability of heads depends on a latent state variable $z_i$. Transitions between values of the latent state are set by parameters $\theta$, while other parameters $\phi$ determine the probability of heads for each value of the latent state. This kind of model is commonly used in computational linguistics, where the $x_i$ might be the sequence of words in a document, and the $z_i$ the syntactic classes from which they are generated.

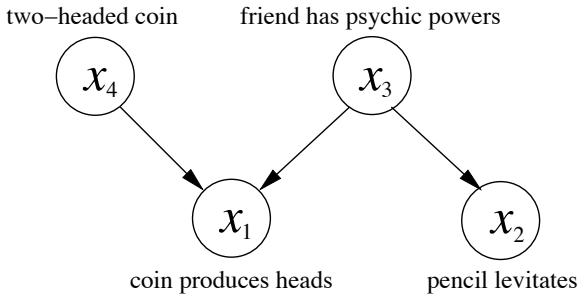


*Figure 3.* Directed graphical model (Bayesian network) showing the dependencies among variables in the "psychic friend" example discussed in the text.

standard uses in the fields of statistics and machine learning. Graphical models can take on a different interpretation in artificial intelligence, when the variables of interest represent the truth value of certain propositions (Russell & Norvig, 2002). For example, imagine that a friend of yours claims to possess psychic powers – in particular, the power of psychokinesis. He proposes to demonstrate these powers by flipping a coin, and influencing the outcome to produce heads. You suggest that a better test might be to see if he can levitate a pencil, since the coin producing heads could also be explained by some kind of sleight of hand, such as substituting a two-headed coin. We can express all possible outcomes of the proposed tests, as well as their causes, using the binary random variables $X_1$, $X_2$, $X_3$, and $X_4$ to represent (respectively) the truth of the coin being flipped and producing heads, the pencil levitating, your friend having psychic powers, and the use of a two-headed coin. Any set of beliefs about these outcomes can be encoded in a joint probability distribution, $P(x_1, x_2, x_3, x_4)$. For example, the probability that the coin comes up heads ($x_1 = 1$) should be higher if your friend actually does have psychic powers ($x_3 = 1$). Figure 3 shows a Bayesian network expressing a possible pattern of dependencies among these variables. For example, $X_1$ and $X_2$ are assumed to be independent given $X_3$, indicating that once it was known whether or not your friend was psychic, the outcomes of the coin flip and the levitation experiments would be completely unrelated. By the Markov condition, we can write $P(x_1, x_2, x_3, x_4) = P(x_1|x_3, x_4)P(x_2|x_3)P(x_3)P(x_4)$.

In addition to clarifying the dependency structure of a set of random variables, Bayesian networks provide an efficient way to represent and compute with probability distributions. In general, a joint probability distribution on $N$ binary variables requires $2^N - 1$ numbers to specify (one for each set of joint values taken by the variables, minus one because of the constraint that probability distributions sum to 1). In the case of the psychic friend example, where there are four variables, this would be $2^4 - 1 = 15$ numbers. However, the factorization of the joint distribution over these variables allows us to use fewer numbers in specifying the distribution over these four variables. We only need one number for each variable conditioned on each possible set of values its parents can take, or $2^{|\mathrm{Pa}(X_i)|}$ numbers for each variable $X_i$ (where $|\mathrm{Pa}(X_i)|$ is the size of the parent set of $X_i$). For our "psychic friend" network, this adds up to 8 numbers rather than 15, because $X_3$ and $X_4$ have no parents (contributing one number each), $X_2$ has one parent (contributing two numbers), and $X_1$ has two parents (contributing four numbers). Recognizing the struc-

ture in this probability distribution can also greatly simplify the computations we want to perform. When variables are independent or conditionally independent of others, it reduces the number of terms that appear in sums over subsets of variables necessary to compute marginal beliefs about a variable or conditional beliefs about a variable given the values of one or more other variables. A variety of algorithms have been developed to perform these probabilistic inferences efficiently on complex models, by recognizing and exploiting conditional independence structures in Bayesian networks (Pearl, 1988; Mackay, 2003). These algorithms form the heart of many modern artificial intelligence systems, making it possible to reason efficiently under uncertainty (Korb & Nicholson, 2003; Russell & Norvig, 2002).

### *3.3 Causal graphical models*

In a standard Bayesian network, edges between variables indicate only statistical dependencies between them. However, recent work has explored the consequences of augmenting directed graphical models with a stronger assumption about the relationships indicated by edges: that they indicate direct causal relationships (Pearl, 2000; Spirtes et al., 1993). This assumption allows causal graphical models to represent not just the probabilities of events that one might observe, but also the probabilities of events that one can produce through intervening on a system. The inferential implications of an event can differ strongly, depending on whether it was observed passively or under conditions of intervention. For example, observing that nothing happens when your friend attempts to levitate a pencil would provide evidence against his claim of having psychic powers; but secretly intervening to hold the pencil down while your friend attempts to levitate it would make the pencil's non-levitation unsurprising and uninformative about his powers.

In causal graphical models, the consequences of intervening on a particular variable can be assessed by removing all incoming edges to that variable and performing probabilistic inference in the resulting "mutilated" model (Pearl, 2000). This procedure produces results that align with our intuitions in the psychic powers example: intervening on $X_2$ breaks its connection with $X_3$, rendering the two variables independent. As a consequence, $X_2$ cannot provide evidence about the value of $X_3$. Several recent papers have investigated whether people are sensitive to the consequences of intervention, generally finding that people differentiate between observational and interventional evidence appropriately (Hagmayer, Sloman, Lagnado, & Waldmann, in press; Lagnado & Sloman, 2004; Steyvers et al., 2003). Introductions to causal graphical models that consider applications to human cognition are provided by Glymour (2001) and Sloman (2005).

The prospect of using graphical models to express the probabilistic consequences of causal relationships has led researchers in several fields to ask whether these models could serve as the basis for learning causal relationships from data. Every introductory class in statistics teaches that "correlation does not imply causation", but the opposite is true: patterns of causation do imply patterns of correlation. A Bayesian learner should thus be able to work backwards from observed patterns of correlation (or statistical dependency) to make probabilistic inferences about the underlying causal structures likely to have generated those observed data. We can use the same basic principles of Bayesian inference developed in the previous section, where now the data are samples from an unknown causal graphical model and the hypotheses to be evaluated are different candidate graphical models. For technical introductions to the methods and challenges of learning causal graphical models,

Table 1: Contingency Table Representation used in Elemental Causal Induction

| | Effect Present ($e^+$) | Effect Absent ($e^-$) |
|---|---|---|
| Cause Present ($c^+$) | $N(e^+, c^+)$ | $N(e^-, c^+)$ |
| Cause Absent ($c^-$) | $N(e^+, c^-)$ | $N(e^-, c^-)$ |

see Heckerman (1998) and Glymour and Cooper (1999).

As in the previous section, it is valuable to distinguish between the problems of parameter estimation and model selection. In the context of causal learning, model selection becomes the problem of determining the graph structure of the causal model – which causal relationships exist – and parameter estimation becomes the problem of determining the strength and polarity of the causal relations specified by a given graph structure. We will illustrate the differences between these two aspects of causal learning, and how graphical models can be brought into contact with empirical data on human causal learning, with a task that has been extensively studied in the cognitive psychology literature: judging the status of a single causal relationship between two variables based on contingency data.

### *3.4 Example: Causal induction from contingency data*

Much psychological research on causal induction has focused upon this simple causal learning problem: given a candidate cause, $C$, and a candidate effect, $E$, people are asked to give a numerical rating assessing the degree to which $C$ causes $E$.[2] We refer to tasks of this sort as "elemental causal induction" tasks. The exact wording of the judgment question varies and until recently was not the subject of much attention, although as we will see below it is potentially quite important. Most studies present information corresponding to the entries in a $2 \times 2$ contingency table, as in Table 1. People are given information about the frequency with which the effect occurs in the presence and absence of the cause, represented by the numbers $N(e^+, c^+), N(e^-, c^-)$ and so forth. In a standard example, $C$ might be injecting a chemical into a mouse, and $E$ the expression of a particular gene. $N(e^+, c^+)$ would be the number of injected mice expressing the gene, while $N(e^-, c^-)$ would be the number of uninjected mice not expressing the gene.

The leading psychological models of elemental causal induction are measures of association that can be computed from simple combinations of the frequencies in Table 1. A classic model first suggested by Jenkins and Ward (1965) asserts that the degree of causation is best measured by the quantity

$$\Delta P = \frac{N(e^+, c^+)}{N(e^+, c^+) + N(e^-, c^+)} - \frac{N(e^+, c^-)}{N(e^+, c^-) + N(e^-, c^-)} = P(e^+|c^+) - P(e^+|c^-), \quad (19)$$

where $P(e^+|c^+)$ is the empirical conditional probability of the effect given the presence of the cause, estimated from the contingency table counts $N(\cdot)$. $\Delta P$ thus reflects the change in the probability of the effect occurring as a consequence of the occurrence of the cause.

[2] As elsewhere in this chapter, we will represent variables such as $C, E$ with capital letters, and their instantiations with lowercase letters, with $c^+, e^+$ indicating that the cause or effect is present, and $c^-, e^-$ indicating that the cause or effect is absent.

More recently, Cheng (1997) has suggested that people's judgments are better captured by a measure called "causal power",

$$\text{power} = \frac{\Delta P}{1 - P(e^+|c^-)}. \tag{20}$$

which takes $\Delta P$ as a component, but predicts that $\Delta P$ will have a greater effect when $P(e^+|c^-)$ is large.

Several experiments have been conducted with the aim of evaluating $\Delta P$ and causal power as models of human jugments. In one such study, Buehner and Cheng (1997, Experiment 1B; this experiment also appears in Buehner, Cheng, & Clifford, 2003) asked people to evaluate causal relationships for 15 sets of contingencies expressing all possible combinations of $P(e^+|c^-)$ and $\Delta P$ in increments of 0.25. The results of this experiment are shown in Figure 4, together with the predictions of $\Delta P$ and causal power. As can be seen from the figure, both $\Delta P$ and causal power capture some of the trends in the data, producing correlations of $r = 0.89$ and $r = 0.88$ respectively. However, since the trends predicted by the two models are essentially orthogonal, neither model provides a complete account of the data.[3]

$\Delta P$ and causal power seem to capture some important elements of human causal induction, but miss others. We can gain some insight into the assumptions behind these models, and identify some possible alternative models, by considering the computational problem behind causal induction using the tools of causal graphical models and Bayesian inference. The task of elemental causal induction can be seen as trying to infer which causal graphical model best characterizes the relationship between the variables $C$ and $E$. Figure 5 shows two possible causal structures relating $C$, $E$, and another variable $B$ which summarizes the influence of all of the other "background" causes of $E$ (which are assumed to be constantly present). The problem of learning which causal graphical model is correct has two aspects: inferring the right causal structure, a problem of model selection, and determining the right parameters assuming a particular structure, a problem of parameter estimation.

In order to formulate the problems of model selection and parameter estimation more precisely, we need to make some further assumptions about the nature of the causal graphical models shown in Figure 5. In particular, we need to define the form of the conditional probability distribution $P(E|B, C)$ for the different structures, often called the *parameterization* of the graphs. Sometimes the parameterization is trivial – for example, $C$ and $E$ are independent in Graph 0, so we just need to specify $P_0(E|B)$, where the subscript indicates that this probability is associated with Graph 0. This can be done using a single numerical parameter $w_0$ which provides the probability that the effect will be present in the presence of the background cause, $P_0(e^+|b^+; w_0) = w_0$. However, when a node has multiple parents, there are many different ways in which the functional relationship between causes and effects could be defined. For example, in Graph 1 we need to account for how the causes $B$ and $C$ interact in producing the effect $E$.

A simple and widely used parameterization for Bayesian networks of binary variables is the noisy-OR distribution (Pearl, 1988). The noisy-OR can be given a natural interpre-

[3]See Griffiths and Tenenbaum (2005) for the details of how these correlations were evaluated, using a power-law transformation to allow for nonlinearities in participants' judgment scales.

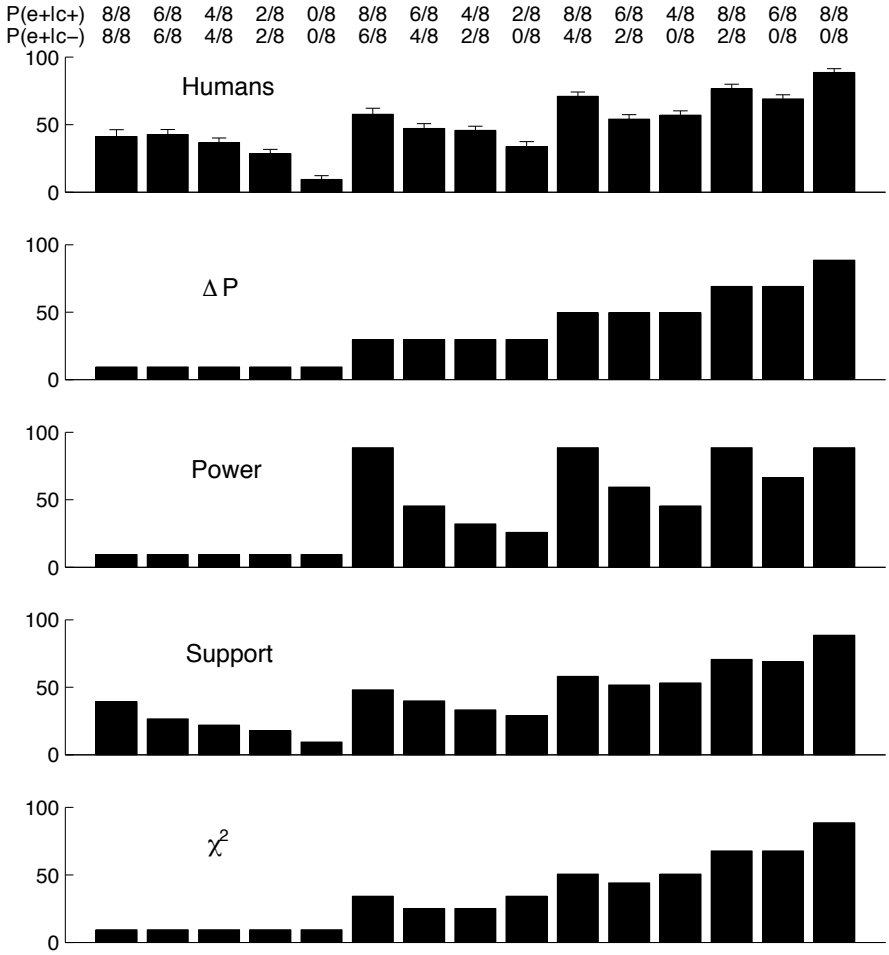


*Figure 4.* Predictions of models compared with the performance of human participants from Buehner and Cheng (1997, Experiment 1B). Numbers along the top of the figure show stimulus contingencies, error bars indicate one standard error.

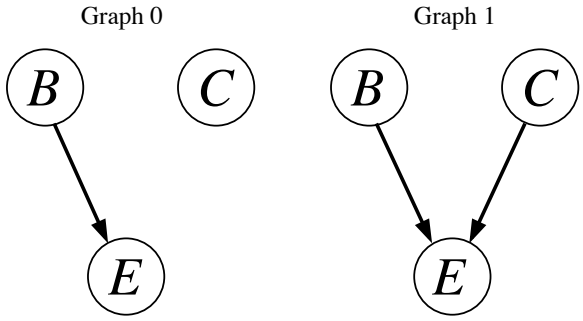


*Figure 5.* Directed graphs involving three variables, $B, C, E$, relevant to elemental causal induction. $B$ represents background variables, $C$ a potential causal variable, and $E$ the effect of interest. Graph 1is assumed in computing $\Delta P$ and causal power. Computing causal support involves comparing the structure of Graph 1 to that of Graph 0in which $C$ and $E$ are independent.

tation in terms of causal relations between multiple causes and a single joint effect. For Graph 1, these assumptions are that $B$ and $C$ are both generative causes, increasing the probability of the effect; that the probability of $E$ in the presence of just $B$ is $w_0$, and in the presence of just $C$ is $w_1$; and that, when both $B$ and $C$ are present, they have independent opportunities to produce the effect. This parameterization can be represented in a compact mathematical form as

$$P_1(e^+|b, c; w_0, w_1) = 1 - (1 - w_0)^b(1 - w_1)^c, \tag{21}$$

where $w_0, w_1$ are parameters associated with the strength of $B, C$ respectively. The variable $c$ is 1 if the cause is present ($c^+$) or 0 if the cause if is absent ($c^-$), and likewise for the variable $b$ with the background cause. This expression gives $w_0$ for the probability of $E$ in the presence of $B$ alone, and $w_0 + w_1 - w_0w_1$ for the probability of $E$ in the presence of both $B$ and $C$. This parameterization is called a noisy-OR because if $w_0$ and $w_1$ are both 1, Equation 21 reduces to the logical OR function: the effect occurs if and only if $B$ or $C$ are present, or both. With $w_0$ and $w_1$ in the range $[0, 1]$, the noisy-OR softens this function but preserves its essentially disjunctive interaction: the effect occurs if and only if $B$ causes it (which happens with probability $w_0$) or $C$ causes it (which happens with probability $w_1$), or both.

An alternative to the noisy-OR might be a linear parameterization of Graph 1, asserting that the probability of $E$ occurring is a linear function of $B$ and $C$. This corresponds to assuming that the presence of a cause simply increases the probability of an effect by a constant amount, regardless of any other causes that might be present. There is no distinction between generative and preventive causes. The result is

$$P_1(e^+|b, c; w_0, w_1) = w_0 \cdot b + w_1 \cdot c. \tag{22}$$

This parameterization requires that we constrain $w_0 + w_1$ to lie between 0 and 1 to ensure that Equation 22 results in a legal probability distribution. Because of this dependence between parameters that seem intuitively like they should be independent, such a linear parameterization is not normally used in Bayesian networks. However, it is relevant for understanding models of human causal induction.

Given a particular causal graph structure and a particular parameterization – for example, Graph 1 parameterized with a noisy-OR function – inferring the strength parameters that best characterize the causal relationships in that model is straightforward. We can use any of the parameter-estimation methods discussed in the previous section (such as maximum-likelihood or MAP estimation) to find the values of the parameters ($w_0$ and $w_1$ in Graph 1) that best fit a set of observed contingencies. Tenenbaum and Griffiths (2001; Griffiths & Tenenbaum, 2005) showed that the two psychological models of causal induction introduced above – $\Delta P$ and causal power – both correspond to maximum-likelihood estimates of the causal strength parameter $w_1$, but under different assumptions about the parameterization of Graph 1. $\Delta P$ results from assuming the linear parameterization, while causal power results from assuming the noisy-OR.

This view of $\Delta P$ and causal power helps to reveal their underlying similarities and differences: they are similar in being maximum-likelihood estimates of the strength parameter describing a causal relationship, but differ in the assumptions that they make about

the form of that relationship. This analysis also suggests another class of models of causal induction that has not until recently been explored: models of learning causal graph structure, or causal model selection rather than parameter estimation. Recalling our discussion of model selection, we can express the evidence that a set of contingencies $d$ provide in favor of the existence of a causal relationship (i.e., Graph 1 over Graph 0) as the log-likelihood ratio in favor of Graph 1. Terming this quantity "causal support", we have

$$\text{support} = \log \frac{P(d|\text{Graph 1})}{P(d|\text{Graph 0})} \tag{23}$$

where $P(d|\text{Graph 1})$ and $P(d|\text{Graph 0})$ are computed by integrating over the parameters associated with the different structures

$$P(d|\text{Graph 1}) = \int_0^1 \int_0^1 P_1(d|w_0, w_1, \text{Graph 1})\ P(w_0, w_1|\text{Graph 1})\ dw_0\ dw_1 \tag{24}$$

$$P(d|\text{Graph 0}) = \int_0^1 P_0(d|w_0, \text{Graph 0})\ P(w_0|\text{Graph 0})\ dw_0. \tag{25}$$

Tenenbaum and Griffiths (2001; Griffiths & Tenenbaum, 2005) proposed this model, and specifically assumed a noisy-OR parameterization for Graph 1 and uniform priors on $w_0$ and $w_1$. Equation 25 is identical to Equation 16 and has an analytic solution. Evaluating Equation 24 is more of a challenge, but one that we will return to later in this chapter when we discuss Monte Carlo methods for approximate probabilistic inference.

The results of computing causal support for the stimuli used by Buehner and Cheng (1997) are shown in Figure 4. Causal support provides an excellent fit to these data, with $r = 0.97$. The model captures the trends predicted by both $\Delta P$ and causal power, as well as trends that are predicted by neither model. These results suggest that when people evaluate contingency, they may be taking into account the evidence that those data provide for a causal relationship as well as the strength of the relationship they suggest. The figure also shows the predictions obtained by applying the $\chi^2$ measure to these data, a standard hypothesis-testing method of assessing the evidence for a relationship (and a common ingredient in non-Bayesian approaches to structure learning, e.g. Spirtes et al., 1993). These predictions miss several important trends in the human data, suggesting that the ability to assert expectations about the nature of a causal relationship that go beyond mere dependency (such as the assumption of a noisy-OR parameterization), is contributing to the success of this model. Causal support predicts human judgments on several other datasets that are problematic for $\Delta P$ and causal power, and also accommodates causal learning based upon the rate at which events occur (see Griffiths & Tenenbaum, 2005, for more details).

The Bayesian approach to causal induction can be extended to cover a variety of more complex cases, including learning in larger causal networks (Steyvers et al., 2003), learning about dynamic causal relationships in physical systems (Tenenbaum & Griffiths, 2003), choosing which interventions to perform in the aid of causal learning (Steyvers et al., 2003), learning about hidden causes (Griffiths, Baraff, & Tenenbaum, 2004) and distinguishing hidden common causes from mere coincidences (Griffiths & Tenenbaum, 2007a), and online learning from sequentially presented data (Danks, Griffiths, & Tenenbaum, 2003).

Modeling learning in these more complex cases often requires us to work with stronger and more structured prior distributions than were needed above to explain elemental causal induction. This prior knowledge can be usefully described in terms of intuitive domain theories (Carey, 1985; Wellman & Gelman, 1992; Gopnik & Meltzoff, 1997), systems of abstract concepts and principles that specify the kinds of entities that can exist in a domain, their properties and possible states, and the kinds of causal relations that can exist between them. We have begun to explore how these abstract causal theories can be formalized as probabilistic generators for hypothesis spaces of causal graphical models, using probabilistic forms of generative grammars, predicate logic, or other structured representations (Griffiths, 2005; Griffiths & Tenenbaum, 2007b; Mansinghka, Kemp, Tenenbaum, & Griffiths, 2006; Tenenbaum et al., 2006; Tenenbaum, Griffiths, & Niyogi, 2007; Tenenbaum & Niyogi, 2003). Given observations of causal events relating a set of objects, these probabilistic theories generate the relevant variables for representing those events, a constrained space of possible causal graphs over those variables, and the allowable parameterizations for those graphs. They also generate a prior distribution over this hypothesis space of candidate causal models, which provides the basis for Bayesian causal learning in the spirit of the methods described above.

We see it as an advantage of the Bayesian approach that it forces modelers to make clear their assumptions about the form and content of learners' prior knowledge. The framework lets us test these assumptions empirically and study how they vary across different settings, by specifying a rational mapping from prior knowledge to learners' behavior in any given task. It may also seem unsatisfying, though, by passing on the hardest questions of learning to whatever mechanism is responsible for establishing learners' prior knowledge. This is the problem we address in the next section, using the techniques of hierarchical Bayesian models.

## 4 Hierarchical Bayesian models

The predictions of a Bayesian model can often depend critically on the prior distribution that it uses. Our early coinflipping examples provided a simple and clear case of the effects of priors. If a coin is tossed once and comes up heads, then a learner who began with a uniform prior on the bias of the coin should predict that the next toss will produce heads with probability $\frac{2}{3}$. If the learner began instead with the belief that the coin is likely to be fair, she should predict that the next toss will produce heads with probability close to $\frac{1}{2}$.

Within statistics, Bayesian approaches have at times been criticized for necessarily requiring some form of prior knowledge. It is often said that good statistical analyses should "let the data speak for themselves", hence the motivation for maximum-likelihood estimation and other classical statistical methods that do not require a prior to be specified. Cognitive models, however, will usually aim for the opposite goal. Most human inferences are guided by background knowledge, and cognitive models should formalize this knowledge and show how it can be used for induction. From this perspective, the prior distribution used by a Bayesian model is critical, since an appropriate prior can capture the background knowledge that humans bring to a given inductive problem. As mentioned in the previous section, prior distributions can capture many kinds of knowledge: priors for causal reasoning, for example, may incorporate theories of folk physics, or knowledge about the powers and liabilities of different ontological kinds.

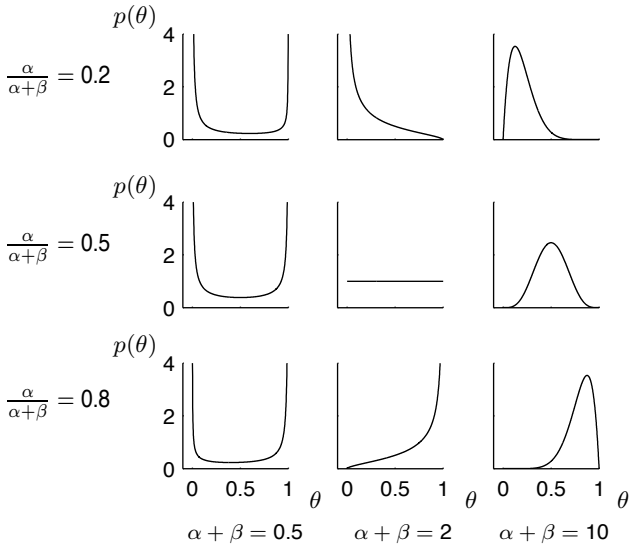


*Figure 6.* The beta distribution serves as a prior on the bias $\theta$ of a coin. The mean of the distribution is $\frac{\alpha}{\alpha+\beta}$, and the shape of the distribution depends on $\alpha + \beta$.

Since background knowledge plays a central role in many human inferences, it is important to ask how this knowledge might be acquired. In a Bayesian framework, the acquisition of background knowledge can be modeled as the acquisition of a prior distribution. We have already seen one piece of evidence that prior distributions can be learned: given two competing models, each of which uses a different prior distribution, Bayesian model selection can be used choose between them. Here we provide a more comprehensive treatment of the problem of learning prior distributions, and show how this problem can be addressed using hierarchical Bayesian models (Good, 1980; Gelman, Carlin, Stern, & Rubin, 1995). Although we will focus on just two applications, the hierarchical Bayesian approach has been applied to several other cognitive problems (Lee, 2006; Tenenbaum et al., 2006; Mansinghka et al., 2006), and many additional examples of hierarchical models can be found in the statistical literature (Gelman et al., 1995; Goldstein, 2003).

Consider first the case where the prior distribution to be learned has known form but unknown parameters. For example, suppose that the prior distribution on the bias of a coin is Beta$(\alpha, \beta)$, where the parameters $\alpha$ and $\beta$ are unknown. We previously considered cases where the parameters $\alpha$ and $\beta$ were positive integers, but in general these parameters can be positive real numbers.[4] As with integer-valued parameters, the mean of the beta distribution is $\frac{\alpha}{\alpha+\beta}$, and $\alpha + \beta$ determines the shape of the distribution. The distribution

[4]The general form of the beta distribution is

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{26}$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx$ is the generalized factorial function (also known as the *gamma function*), with $\Gamma(n) = (n-1)!$ for any integer argument $n$ and smoothly interpolating between the factorials for real-valued arguments (e.g., Boas, 1983).

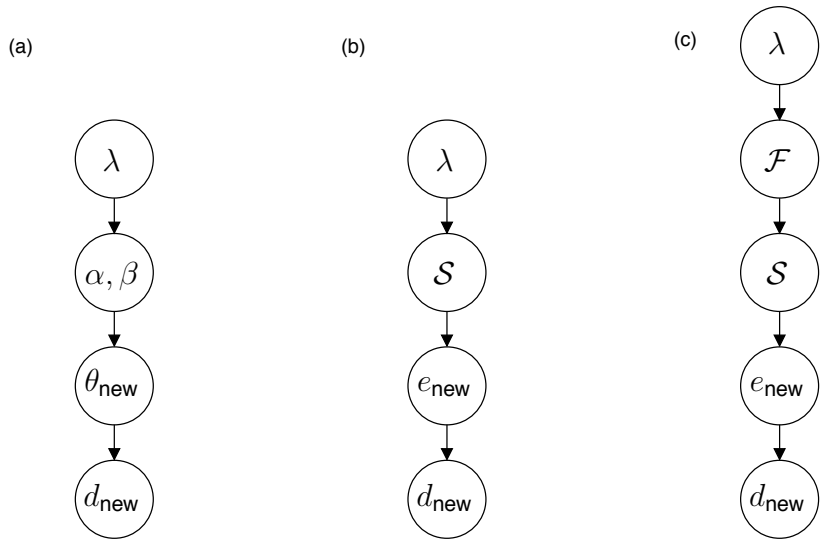


*Figure 7.* Three hierarchical Bayesian models. (a) A model for inferring $\theta_{\text{new}}$, the bias of a coin. $d_{\text{new}}$ specifies the number of heads and tails observed when the coin is tossed. $\theta_{\text{new}}$ is is drawn from a beta distribution with parameters $\alpha$ and $\beta$. The prior distribution on these parameters has a single hyperparameter, $\lambda$. (b) A model for inferring $e_{\text{new}}$, the extension of a novel property. $d_{\text{new}}$ is a sparsely observed version of $e_{\text{new}}$, and $e_{\text{new}}$ is assumed to be drawn from a prior distribution induced by structured representation $\mathcal{S}$. The hyperparameter $\lambda$ specifies a prior distribution over a hypothesis space of structured representations. (c) A model that can discover the form $\mathcal{F}$ of the structure $\mathcal{S}$. The hyperparameter $\lambda$ now specifies a prior distribution over a hypothesis space of structural forms.

is tightly peaked around its mean when $\alpha + \beta$ is large, flat when $\alpha = \beta = 1$, and U-shaped when $\alpha + \beta$ is small (Figure 6). Observing the coin being tossed provides some information about the values of $\alpha$ and $\beta$, and a learner who begins with prior distributions on the values of these parameters can update these distributions as each new coin toss is observed. The prior distributions on $\alpha$ and $\beta$ may be defined in terms of one or more hyperparameters. The hierarchical model in Figure 7a uses three levels, where the hyperparameter at the top level ($\lambda$) is fixed. In principle, however, we can develop hierarchical models with any number of levels — we can can continue adding hyperparameters and priors on these hyperparameters until we reach a level where we are willing to assume that the hyperparameters are fixed in advance.

At first, the upper levels in hierarchical models like Figure 7a might seem too abstract to be of much practical use. Yet these upper levels play a critical role — they allow knowledge to be shared across contexts that are related but distinct. In our coin tossing example, these contexts correspond to observations of many different coins, each of which has a bias sampled from the same prior distribution Beta$(\alpha, \beta)$. It is possible to learn something about $\alpha$ and $\beta$ by tossing a single coin, but the best way to learn about $\alpha$ and $\beta$ is probably to experiment with many different coins. If most coins tend to come up heads about half the time, we might infer that $\alpha$ and $\beta$ are both large, and are close to each other in size. Suppose, however, that we are working in a factory that produces trick coins for

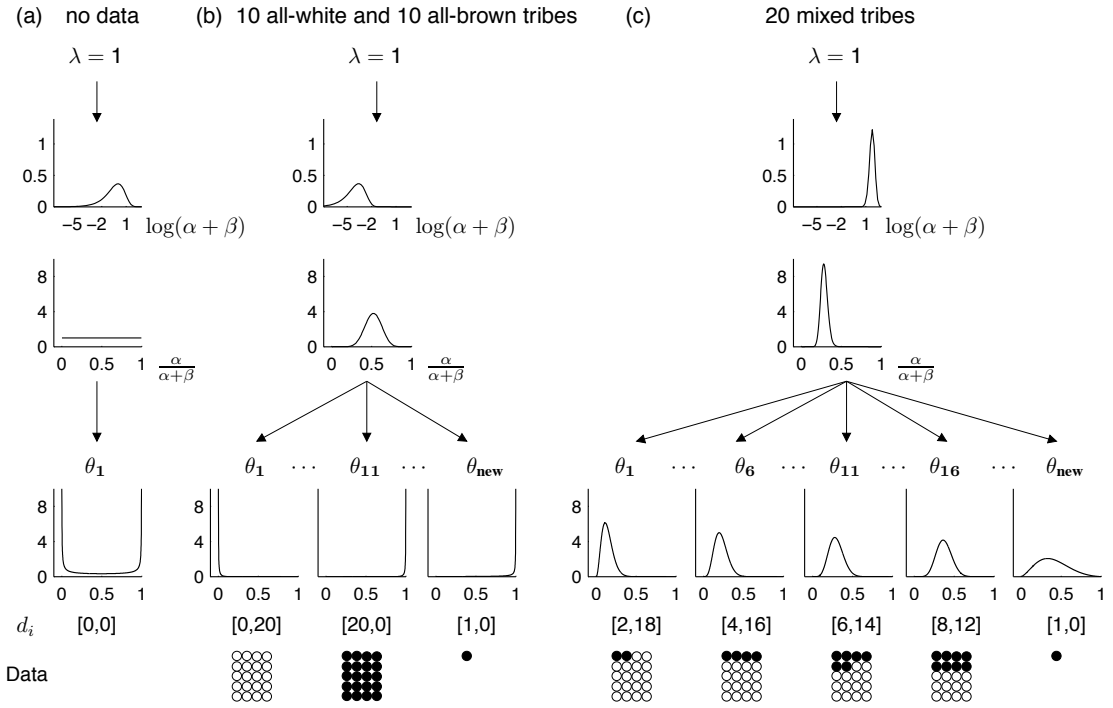


*Figure 8.* Inferences about the distribution of features within tribes. (a) Prior distributions on $\theta$, $\log(\alpha+\beta)$ and $\frac{\alpha}{\alpha+\beta}$. (b) Posterior distributions after observing 10 all-white tribes and 10 all-brown tribes. (c) Posterior distributions after observing 20 tribes. Black circles indicate obese indiviuals, and the rate of obesity varies among tribes.

magicians. If 80% of coins come up heads almost always, and the remainder come up tails almost always, we might infer that $\alpha$ and $\beta$ are both very small, and that $\frac{\alpha}{\alpha+\beta} \approx 0.8$.

More formally, suppose that we have observed many coins being tossed, and that $d_i$ is the tally of heads and tails produced by the $i$th coin. The $i$th coin has bias $\theta_i$, and each bias $\theta_i$ is sampled from a beta distribution with parameters $\alpha$ and $\beta$. The hierarchical model in Figure 8 captures these assumptions, and is known by statisticians as a beta-binomial model (Gelman et al., 1995). To learn about the prior distribution $\text{Beta}(\alpha, \beta)$ we must formalize our expectations about the values of $\alpha$ and $\beta$. We will assume that the mean of the beta distribution $\frac{\alpha}{\alpha+\beta}$ is uniformly drawn from the interval $[0, 1]$, and that the sum of the parameters $\alpha + \beta$ is drawn from an exponential distribution with hyperparameter $\lambda$. Given the hierarchical model in Figure 8, inferences about any of the $\theta_i$ can be made by integrating out $\alpha$ and $\beta$:

$$p(\theta_i|d_1, d_2, \ldots, d_n) = \int p(\theta_i|\alpha, \beta, d_i)p(\alpha, \beta|d_1, d_2, \ldots, d_n)d\alpha d\beta \tag{27}$$

and this integral can be approximated using the Markov chain Monte Carlo methods described in the next section (see also Kemp, Perfors, & Tenenbaum, in press).

*4.1 Example: Learning about feature variability*

Humans acquire many kinds of knowledge about categories and their features. Some kinds of knowledge are relatively concrete: for instance, children learn that balls tend to be round, and that televisions tend to be box-shaped. Other kinds of knowledge are more abstract, and represent discoveries about categories in general. For instance, 30-month-old children display a *shape bias*: they appear to know that the objects in any given category tend to have the same shape, even if they differ along other dimensions, such as color and texture (Heibeck & Markman, 1987; Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). The shape bias is one example of abstract knowledge about feature variability, and Kemp et al. (in press) have argued that knowledge of this sort can be acquired by hierarchical Bayesian models.

A task carried out by Nisbett, Krantz, Jepson, and Kunda (1983) shows how knowledge about feature variability can support inductive inferences from very sparse data. These researchers asked participants to imagine that they were exploring an island in the Southeastern Pacific, that they had encountered a single member of the Barratos tribe, and that this individual was brown and obese. Based on this single example, participants concluded that most Barratos were brown, but gave a much lower estimate of the proportion of obese Barratos. These inferences can be explained by the beliefs that skin color is a feature that is consistent within tribes, and that obesity tends to vary within tribes, and the model in Figure 8 can explain how these beliefs might be acquired.

Kemp et al. (in press) describe a model that can reason simultaneously about multiple features, but for simplicity we will consider skin color and obesity separately. Consider first the case where $\theta_i$ represents the proportion of brown-skinned individuals within tribe $i$, and suppose that we have observed 20 members from each of 20 tribes. Half the tribes are brown and the other half are white, but all of the individuals in a given tribe have the same skin color. Given these observations, the posterior distribution on $\alpha + \beta$ indicates that $\alpha + \beta$ is likely to be small (Figure 8b). Recall that small values of $\alpha + \beta$ imply that most of the $\theta_i$ will be close to 0 or close to 1 (Figure 6): in other words, that skin color tends to be homogeneous within tribes. Learning that $\alpha + \beta$ is small allows the model to make strong predictions about a sparsely observed new tribe: having observed a single brown-skinned member of a new tribe, the posterior distribution on $\theta_{\text{new}}$ indicates that most members of the tribe are likely to be brown (Figure 8b). Note that the posterior distribution on $\theta_{\text{new}}$ is almost as sharply peaked as the posterior distribution on $\theta_{11}$: the model has realized that observing one member of a new tribe is almost as informative as observing 20 members of that tribe.

Consider now the case where $\theta_i$ represents the proportion of obese individuals within tribe $i$. Suppose that obesity is a feature that varies within tribes: a quarter of the 20 tribes observed have an obesity rate of 10%, and the remaining three quarters have rates of 20%, 30%, and 40% respectively (Figure 8c). Given these observations, the posterior distributions on $\alpha + \beta$ and $\frac{\alpha}{\alpha+\beta}$ (Figure 8c) indicate that obesity varies within tribes ($\alpha + \beta$ is high), and that the base rate of obesity is around 25% ($\frac{\alpha}{\alpha+\beta}$ is around 0.25). Again, we can use these posterior distributions to make predictions about a new tribe, but now the model requires many observations before it concludes that most members of the new tribe are obese. Unlike the case in Figure 8b, the model has learned that a single observation

of a new tribe is not very informative, and the distribution on $\theta_{\text{new}}$ is now similar to the average of the $\theta$ values for all previously observed tribes.

In Figures 8b and 8c, a hierarchical model is used to simultaneously learn about high-level knowledge ($\alpha$ and $\beta$) and low-level knowledge (the values of $\theta_i$). Any hierarchical model, however, can be used for several different purposes. If $\alpha$ and $\beta$ are fixed in advance, the model supports top-down learning: knowledge about $\alpha$ and $\beta$ can guide inferences about the $\theta_i$. If the $\theta_i$ are fixed in advance, the model supports bottom-up learning, and the $\theta_i$ can guide inferences about $\alpha$ and $\beta$. The ability to support top-down and bottom-up inferences is a strength of the hierarchical approach, but simultaneous learning at multiple levels of abstraction is often required to account for human inferences. Note, for example, that judgments about the Barratos depend critically on learning at two levels: learning at the level of $\theta$ is needed to incorporate the observation that the new tribe has at least one obese, brown-skinned member, and learning at the level of $\alpha$ and $\beta$ is needed to discover that skin-color is homogeneous within tribes but that obesity is not.

### *4.2 Example: Property induction*

We have just seen that hierarchical Bayesian models can explain how the parameters of a prior distribution might be learned. Prior knowledge in human cognition, however, is often better characterized using more structured representations. Here we present a simple case study that shows how a hierarchical Bayesian model can acquire structured prior knowledge.

Structured prior knowledge plays a role in many inductive inferences, but we will consider the problem of property induction. In a typical task of this sort, learners find out that one or more members of a domain have a novel property, and decide how to extend the property to the remaining members of the domain. For instance, given that gorillas carry enzyme X132, how likely is it that chimps also carry this enzyme? (Rips, 1975; Osherson, Smith, Wilkie, Lopez, & Shafir, 1990). For our purposes, inductive problems like these are interesting because they rely on relatively rich prior knowledge, and because this prior knowledge often appears to be learned. For example, humans learn at some stage that gorillas are more closely related to chimps than to squirrels, and taxonomic knowledge of this sort guides inferences about novel anatomical and physiological properties.

The problem of property induction can be formalized as an inference about the extension of a novel property (Kemp & Tenenbaum, 2003). Suppose that we are working with a finite set of animal species. Let $e_{\text{new}}$ be a binary vector which represents the true extension of the novel property (Figures 7 and 9). For example, the element in $e_{\text{new}}$ that corresponds to gorillas will be 1 (represented as a black circle in Figure 9) if gorillas have the novel property, and 0 otherwise. Let $d_{\text{new}}$ be a partially observed version of extension $e_{\text{new}}$ (Figure 9). We are interested in the posterior distribution on $e_{\text{new}}$ given the sparse observations in $d_{\text{new}}$. Using Bayes' rule, this distribution can be written as

$$P(e_{\text{new}}|d_{\text{new}}, \mathcal{S}) = \frac{P(d_{\text{new}}|e_{\text{new}})P(e_{\text{new}}|\mathcal{S})}{P(d_{\text{new}}|\mathcal{S})} \tag{28}$$

where $\mathcal{S}$ captures the structured prior knowledge which is relevant to the novel property. The first term in the numerator, $P(d_{\text{new}}|e_{\text{new}})$, depends on the process by which the observations in $d_{\text{new}}$ were sampled from the true extension $e_{\text{new}}$. We will assume for simplicity that the

entries in $d_{\text{new}}$ are sampled at random from the vector $e_{\text{new}}$. The denominator can be computed by summing over all possible values of $e_{\text{new}}$:

$$P(d_{\text{new}}|\mathcal{S}) = \sum_{e_{\text{new}}} P(d_{\text{new}}|e_{\text{new}})P(e_{\text{new}}|\mathcal{S}). \quad (29)$$

For reasoning about anatomy, physiology, and other sorts of generic biological properties (e.g., "has enzyme X132"), the prior $P(e_{\text{new}}|\mathcal{S})$ will typically capture knowledge about taxonomic relationships between biological species. For instance, it seems plausible *a priori* that gorillas and chimps are the only familiar animals that carry a certain enzyme, but less probable that this enzyme will only be found in gorillas and squirrels.

Prior knowledge about taxonomic relationships between living kinds can be captured using a tree-structured representation like the taxonomy shown in Figure 9. We will therefore assume that the structured prior knowledge $\mathcal{S}$ takes the form of a tree, and define a prior distribution $P(e_{\text{new}}|\mathcal{S})$ using a stochastic process over this tree. The stochastic process assigns some prior probability to all possible extensions, but the most likely extensions are those that are smooth with respect to tree $\mathcal{S}$. An extension is smooth if nearby species in the tree tend to have the same status — either both have the novel property, or neither does. One example of a stochastic process that tends to generate properties smoothly over the tree is a mutation process, inspired by biological evolution: the property is randomly chosen to be on or off at the root of the tree, and then has some small probability of switching state at each point of each branch of the tree (Huelsenbeck & Ronquist, 2001; Kemp, Perfors, & Tenenbaum, 2004).

For inferences about generic biological properties, the problem of acquiring prior knowledge has now been reduced to the problem of finding an appropriate tree $\mathcal{S}$. Human learners acquire taxonomic representations in part by observing properties of entities: noticing, for example, that gorillas and chimps have many properties in common and should probably appear nearby in a taxonomic structure. This learning process can be formalized using the hierarchical Bayesian model in Figure 9. We assume that a learner has partially observed the extensions of $n$ properties, and that these observations are collected in vectors labeled $d_1$ through $d_n$. The true extensions $e_i$ of these properties are generated from the same tree-based prior that is assumed to generate $e_{\text{new}}$, the extension of the novel property. Learning the taxonomy now amounts to making inferences about the tree $\mathcal{S}$ that is most likely to have generated all of these partially observed properties. Again we see that a hierarchical formulation allows information to be shared across related contexts. Here, information about $n$ partially observed properties is used to influence the prior distribution for inferences about $e_{\text{new}}$. To complete the hierarchical model in Figure 9 it is necessary to specify a prior distribution on trees $\mathcal{S}$: for simplicity, we can use a uniform distribution over tree topologies, and an exponential distribution with parameter $\lambda$ over the branch lengths.

Inferences about $e_{\text{new}}$ can now be made by integrating out the underlying tree $\mathcal{S}$:

$$P(e_{\text{new}}|d_1, \ldots, d_n, d_{\text{new}}) = \int P(e_{\text{new}}|d_{\text{new}}, \mathcal{S})p(\mathcal{S}|d_1, \ldots, d_n, d_{\text{new}})d\mathcal{S} \quad (30)$$

where $P(e_{\text{new}}|d_{\text{new}}, \mathcal{S})$ is defined in Equation 28. This integral can be approximated by using Markov chain Monte Carlo methods of the kind discussed in the next section to draw a sample of trees from the distribution $p(\mathcal{S}|d_1, \ldots, d_n, d_{\text{new}})$ (Huelsenbeck & Ronquist, 2001).

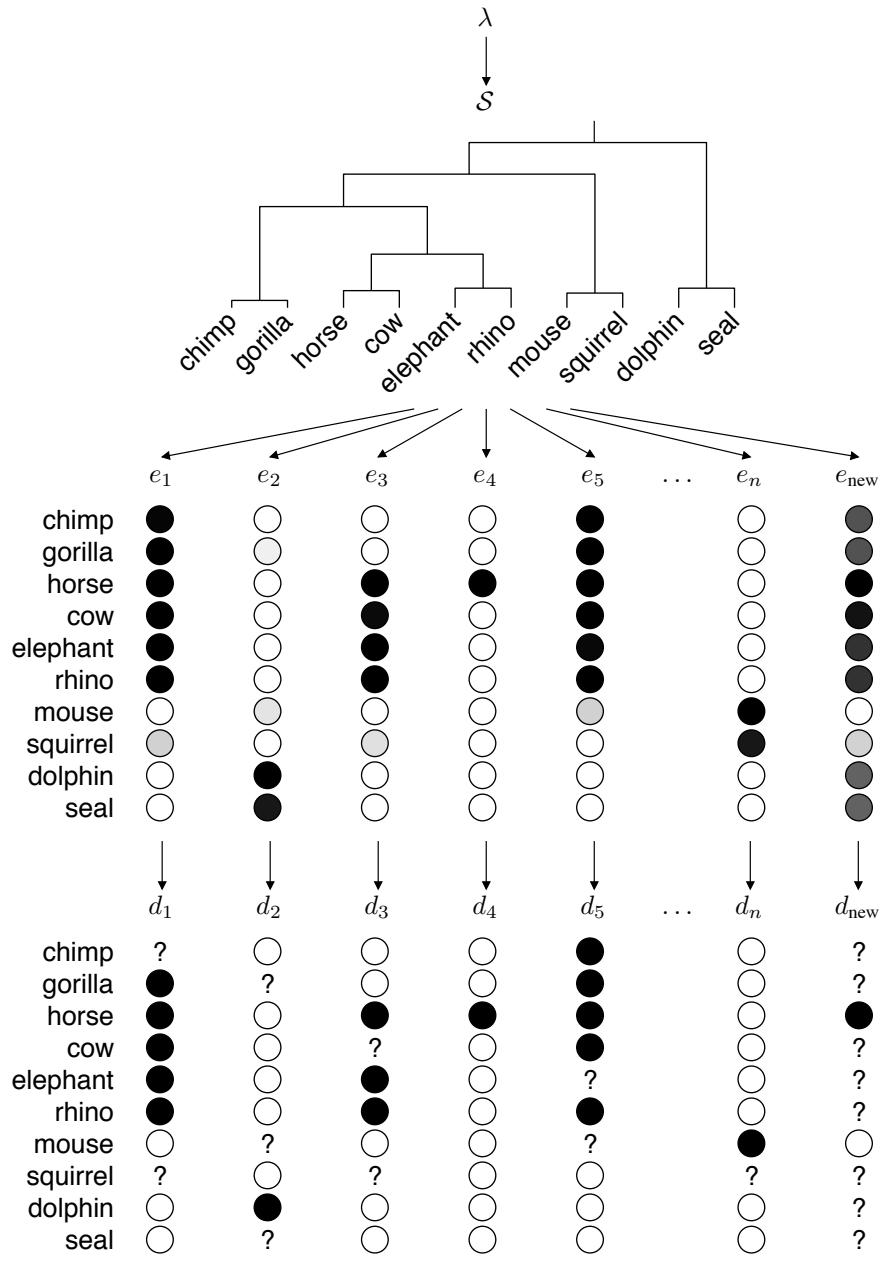


*Figure 9.* Learning a tree-structured prior for property induction. Given a collection of sparsely observed properties $d_i$ (a black circle indicates that a species has a given property), we can compute a posterior distribution on structure $S$ and posterior distributions on each extension $e_i$. Since the distribution over $S$ is difficult to display, we show a single tree with high posterior probability. Since each distribution on $e_i$ is difficult to display, we show instead the posterior probability that each species has each property (dark circles indicate probabilities close to 1).

If preferred, a single tree with high posterior probability can be identified, and this tree can be used to make predictions about the extension of the novel property. Kemp et al. (2004) follow this second strategy, and show that a single tree is sufficient to accurately predict human inferences about the extensions of novel biological properties.

The model in Figures 7b and 9 assumes that the extensions $e_i$ are generated over some true but unknown tree $\mathcal{S}$. Tree structures may be useful for capturing taxonomic relationships between biological species, but different kinds of structured representations such as chains, rings, or sets of clusters are useful in other settings. Understanding which kind of representation is best for a given context is sometimes thought to rely on innate knowledge: Atran (1998), for example, argues that the tendency to organize living kinds into tree structures reflects an "innately determined cognitive module." The hierarchical Bayesian approach challenges the inevitability of this conclusion by showing how a model might discover which kind of representation is best for a given data set. We can create such a model by adding an additional level to the model in Figure 7b. Suppose that variable $\mathcal{F}$ indicates whether $\mathcal{S}$ is a tree, a chain, a ring, or an instance of some other structural form. Given a prior distribution over a hypothesis space of possible forms, the model in Figure 7c can simultaneously discover the form $\mathcal{F}$ and the instance of that form $\mathcal{S}$ that best account for a set of observed properties. Kemp et al. (2004) formally define a model of this sort, and show that it chooses appropriate representations for several domains. For example, the model chooses a tree-structured representation given information about animals and their properties, but chooses a linear representation (the liberal-conservative spectrum) when supplied with information about the voting patterns of Supreme Court judges.

The models in Figure 7b and 7c demonstrate that the hierarchical Bayesian approach can account for the acquisition of structured prior knowledge. Many domains of human knowledge, however, are organized into representations that are richer and more sophisticated than the examples we have considered. The hierarchical Bayesian approach provides a framework that can help to explore the use and acquisition of richer prior knowledge, such as the intuitive causal theories we described at the end of Section 3. For instance, Mansinghka, Kemp, Tenenbaum, and Griffiths (2006) describe a two-level hierarchical model in which the lower level represents a space of causal graphical models, while the higher level specifies a simple abstract theory: it assumes that the variables in the graph come in one or more classes, with the prior probability of causal relations between them depending on these classes. The model can then be used to infer the number of classes, which variables are in which classes, and the probability of causal links existing between classes directly from data, at the same time as it learns the specific causal relations that hold between individual pairs of variables. Given data from a causal network that embodies some such regularity, the model of Mansinghka et al. (2006) infers the correct network structure from many fewer examples than would be required under a generic uniform prior, because it can exploit the constraint of a learned theory of the network's abstract structure. While the theories that can be learned using our best hierarchical Bayesian models are still quite simple, these frameworks provide a promising foundation for future work and an illustration of how structured knowledge representations and sophisticated statistical inference can interact productively in cognitive modeling.

## 5 Markov chain Monte Carlo

The probability distributions one has to evaluate in applying Bayesian inference can quickly become very complicated, particularly when using hierarchical Bayesian models. Graphical models provide some tools for speeding up probabilistic inference, but these tools tend to work best when most variables are directly dependent on a relatively small number of other variables. Other methods are needed to work with large probability distributions that exhibit complex interdependencies among variables. In general, ideal Bayesian computations can only be approximated for these complex models, and many methods for approximate Bayesian inference and learning have been developed (Bishop, 2006; Mackay, 2003). In this section we introduce the Markov chain Monte Carlo approach, a general-purpose toolkit for inferring the values of latent variables, estimating parameters and learning model structure, which can work with a very wide range of probabilistic models. The main drawback of this approach is that it can be slow, but given sufficient time it can yield accurate inferences for models that cannot be handled by other means.

The basic idea behind Monte Carlo methods is to represent a probability distribution by a set of samples from that distribution. Those samples provide an idea of which values have high probability (since high probability values are more likely to be produced as samples), and can be used in place of the distribution itself when performing various computations. When working with Bayesian models of cognition, we are typically interested in understanding the posterior distribution over a parameterized model – such as a causal network with its causal strength parameters – or over a class of models – such as the space of all causal network structures on a set of variables, or all taxonomic tree structures on a set of objects. Samples from the posterior distribution can be useful in discovering the best parameter values for a model or the best models in a model class, and for estimating how concentrated the posterior is on those best hypotheses (i.e., how confident a learner should be in those hypotheses).

Sampling can also be used to approximate averages over the posterior distribution. For example, in computing the posterior probability of a parameterized model given data, it is necessary to compute the model's marginal likelihood, or the average probability of the data over all parameter settings of the model (as in Equation 16 for determining whether we have a fair or weighted coin). Averaging over all parameter settings is also necessary for ideal Bayesian prediction about future data points (as in computing the posterior predictive distribution for a weighted coin, Equation 11). Finally, we could be interested in averaging over a space of model structures, making predictions about model features that are likely to hold regardless of which structure is correct. For example, we could estimate how likely it is that one variable $A$ causes variable $B$ in a complex causal network of unknown structure, by computing the probability that a link $A \rightarrow B$ exists in a high-probability sample from the posterior over network structures (Friedman & Koller, 2000).

Monte Carlo methods were originally developed primarily for approximating these sophisticated averages – that is, approximating a sum over all of the values taken on by a random variable with a sum over a random sample of those values. Assume that we want to evaluate the average (also called the *expected value*) of a function $f(\mathbf{x})$ over a probability distribution $p(\mathbf{x})$ defined on a set of $k$ random variables taking on values $\mathbf{x} = (x_1, x_2, \ldots, x_k)$. This can be done by taking the integral of $f(\mathbf{x})$ over all value of $\mathbf{x}$, weighted by their

probability $p(\mathbf{x})$. Monte Carlo provides an alternative, relying upon the law of large numbers to justify the approximation

$$\int f(\mathbf{x})p(\mathbf{x})\ d\mathbf{x} \approx \sum_{i=1}^{m} f(\mathbf{x}^{(i)}) \tag{31}$$

where the $\mathbf{x}^{(i)}$ are a set of $m$ samples from the distribution $p(\mathbf{x})$. The accuracy of this approximation increases as $m$ increases.

To show how the Monte Carlo approach to approximate numerical integration is useful for evaluating Bayesian models, recall our model of causal structure-learning known as causal support. In order to compute the evidence that a set of contingencies $d$ provides in favor of a causal relationship, we needed to evaluate the integral

$$P(d|\text{Graph 1}) = \int_0^1 \int_0^1 P_1(d|w_0, w_1, \text{Graph 1})\ P(w_0, w_1|\text{Graph 1})\ dw_0\ dw_1 \tag{32}$$

where $P_1(d|w_0, w_1, \text{Graph 1})$ is derived from the noisy-OR parameterization, and $P(w_0, w_1|\text{Graph 1})$ is assumed to be uniform over all values of $w_0$ and $w_1$ between 0 and 1. If we view $P_1(d|w_0, w_1, \text{Graph 1})$ simply as a function of $w_0$ and $w_1$, it is clear that we can approximate this integral using Monte Carlo. The analogue of Equation 31 is

$$P(d|\text{Graph 1}) \approx \sum_{i=1}^{m} P_1(d|w_0^{(i)}, w_1^{(i)}, \text{Graph 1}) \tag{33}$$

where the $w_0^{(i)}$ and $w_1^{(i)}$ are a set of $m$ samples from the distribution $P(w_0, w_1|\text{Graph 1})$. A version of this simple approximation was used to compute the values of causal support shown in Figure 4 (for details, see Griffiths & Tenenbaum, 2005).

One limitation of classical Monte Carlo methods is that it is not easy to automatically generate samples from most probability distributions. There are a number of ways to address this problem, including methods such as rejection sampling and importance sampling (see, e.g., Neal, 1993). One of the most flexible methods for generating samples from a probability distribution is Markov chain Monte Carlo (MCMC), which can be used to construct samplers for arbitrary probability distributions even if the normalizing constants of those distributions are unknown. MCMC algorithms were originally developed to solve problems in statistical physics (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953), and are now widely used across physics, statistics, machine learning, and related fields (e.g., Newman & Barkema, 1999; Gilks, Richardson, & Spiegelhalter, 1996; Mackay, 2003; Neal, 1993).

As the name suggests, Markov chain Monte Carlo is based upon the theory of Markov chains – sequences of random variables in which each variable is conditionally independent of all previous variables given its immediate predecessor (as in Figure 2b). The probability that a variable in a Markov chain takes on a particular value conditioned on the value of the preceding variable is determined by the *transition kernel* for that Markov chain. One well known property of Markov chains is their tendency to converge to a *stationary distribution*: as the length of a Markov chain increases, the probability that a variable in that chain takes on a particular value converges to a fixed quantity determined by the choice

of transition kernel. If we sample from the Markov chain by picking some initial value and then repeatedly sampling from the distribution specified by the transition kernel, we will ultimately generate samples from the stationary distribution.

In MCMC, a Markov chain is constructed such that its stationary distribution is the distribution from which we want to generate samples. If the target distribution is $p(\mathbf{x})$, then the Markov chain would be defined on sequences of values of $\mathbf{x}$. The transition kernel $K(\mathbf{x}^{(i+1)}|\mathbf{x}^{(i)})$ gives the probability of moving from state $\mathbf{x}^{(i)}$ to state $\mathbf{x}^{(i+1)}$. In order for the stationary distribution of the Markov chain to be the target distribution $p(\mathbf{x})$, the transition kernel must be chosen so that $p(\mathbf{x})$ is invariant to the kernel. Mathematically this is expressed by the condition

$$p(\mathbf{x}^{(i+1)}) = \sum_{\mathbf{x}} p(\mathbf{x}) K(\mathbf{x}|\mathbf{x}'). \tag{34}$$

If this is the case, once the probability that the chain is in a particular state is equal to $p(\mathbf{x})$, it will continue to be equal to $p(\mathbf{x})$ – hence the term "stationary distribution". Once the chain converges to its stationary distribution, averaging a function $f(\mathbf{x})$ over the values of $\mathbf{x}^{(i)}$ will approximate the average of that function over the probability distribution $p(\mathbf{x})$.

Fortunately, there is a simple procedure that can be used to construct a transition kernel that will satisfy Equation 34 for any choice of $p(\mathbf{x})$, known as the *Metropolis-Hastings algorithm* (Hastings, 1970; Metropolis et al., 1953). The basic idea is to define $K(\mathbf{x}^{(i+1)}|\mathbf{x}^{(i)})$ as the result of two probabilistic steps. The first step uses an arbitrary *proposal distribution*, $q(\mathbf{x}^*|\mathbf{x}^{(i)})$, to generate a proposed value $\mathbf{x}^*$ for $\mathbf{x}^{(i+1)}$. The second step is to decide whether to accept this proposal. This is done by computing the *acceptance probability*, $A(\mathbf{x}^*|\mathbf{x}^{(i)})$, defined to be

$$A(\mathbf{x}^*|\mathbf{x}^{(i)}) = \min\left[\frac{p(\mathbf{x}^*)q(\mathbf{x}^{(i)}|\mathbf{x}^*)}{p(\mathbf{x}^{(i)})q(\mathbf{x}^*|\mathbf{x}^{(i)})}, 1\right]. \tag{35}$$

If a random number generated from a uniform distribution over $[0, 1]$ is less than $A(\mathbf{x}^*|\mathbf{x}^{(i)})$, the proposed value $\mathbf{x}^*$ is accepted as the value of $\mathbf{x}^{(i+1)}$. Otherwise, the Markov chain remains at its previous value, and $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)}$. An illustration of the use of the Metropolis-Hastings algorithm to generate samples from a Gaussian distribution (which is easy to sample from in general, but convenient to work with in this case) appears in Figure 10.

One advantage of the Metropolis-Hastings algorithm is that it requires only limited knowledge of the probability distribution $p(\mathbf{x})$. Inspection of Equation 35 reveals that, in fact, the Metropolis-Hastings algorithm can be applied even if we only know some quantity proportional to $p(\mathbf{x})$, since only the ratio of these quantities affects the algorithm. If we can sample from distributions related to $p(\mathbf{x})$, we can use other Markov chain Monte Carlo methods. In particular, if we are able to sample from the conditional probability distribution for each variable in a set given the remaining variables, $p(x_j|x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)$, we can use another popular algorithm, *Gibbs sampling* (Geman & Geman, 1984; Gilks et al., 1996), which is known in statistical physics as the heatbath algorithm (Newman & Barkema, 1999). The Gibbs sampler for a target distribution $p(\mathbf{x})$ is the Markov chain defined by drawing each $x_j$ from the conditional distribution $p(x_j|x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k)$.

Markov chain Monte Carlo can be a good way to obtaining samples from probability distributions that would otherwise be difficult to compute with, including the posterior

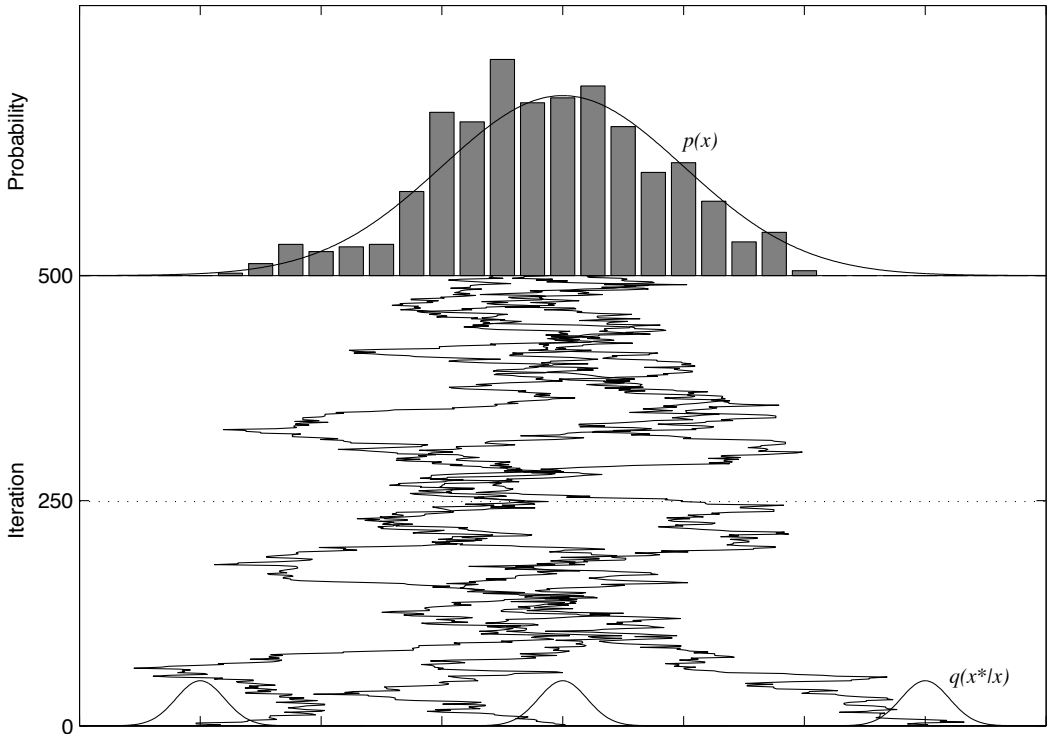


*Figure 10.* The Metropolis-Hastings algorithm. The solid lines shown in the bottom part of the figure are three sequences of values sampled from a Markov chain. Each chain began at a different location in the space, but used the same transition kernel. The transition kernel was constructed using the procedure described in the text for the Metropolis-Hastings algorithm: the proposal distribution, $q(x^*|x)$, was a Gaussian distribution with mean $x$ and standard deviation 0.2 (shown centered on the starting value for each chain at the bottom of the figure), and the acceptance probabilities were computed by taking $p(x)$ to be Gaussian with mean 0 and standard deviation 1 (plotted with a solid line in the top part of the figure). This guarantees that the stationary distribution associated with the transition kernel is $p(x)$. Thus, regardless of the initial value of each chain, the probability that the chain takes on a particular value will converge to $p(x)$ as the number of iterations increases. In this case, all three chains move to explore a similar part of the space after around 100 iterations. The histogram in the top part of the figure shows the proportion of time the three chains spend visiting each part in the space after 250 iterations (marked with the dotted line), which closely approximates $p(x)$. Samples from the Markov chains can thus be used similarly to samples from $p(x)$.

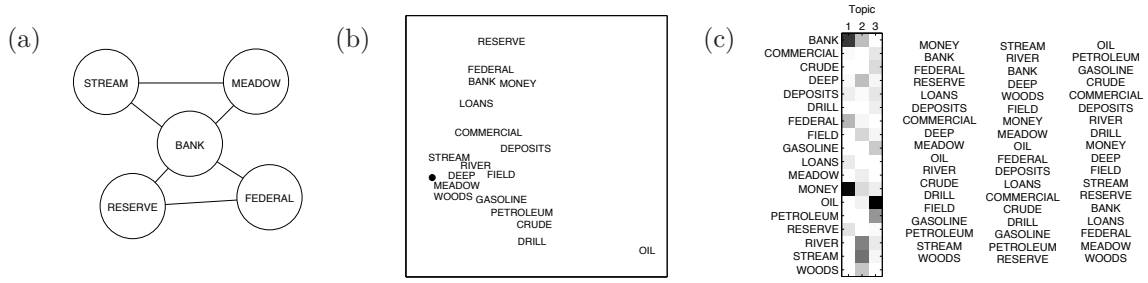


*Figure 11.* Approaches to semantic representation. (a) In a semantic network, words are represented as nodes, and edges indicate semantic relationships. (b) In a semantic space, words are represented as points, and proximity indicates semantic association. These are the first two dimensions of a solution produced by Latent Semantic Analysis (Landauer & Dumais, 1997). The black dot is the origin. (c) In the topic model, words are represented as belonging to a set of probabilistic topics. The matrix shown on the left indicates the probability of each word under each of three topics. The three columns on the right show the words that appear in those topics, ordered from highest to lowest probability.

distributions associated with complex probabilistic models. To illustrate how MCMC can be applied in the context of a Bayesian model of cognition, we will show how Gibbs sampling can be used to extract a statistical representation of the meanings of words from a collection of text documents.

### *5.1 Example: Inferring topics from text*

Several computational models have been proposed to account for the large-scale structure of semantic memory, including semantic networks (e.g., Collins & Loftus, 1975; Collins & Quillian, 1969) and semantic spaces (e.g., Landauer & Dumais, 1997; Lund & Burgess, 1996). These approaches embody different assumptions about the way that words are represented. In semantic networks, words are nodes in a graph where edges indicate semantic relationships, as shown in Figure 11 (a). In semantic space models, words are represented as points in high-dimensional space, where the distance between two words reflects the extent to which they are semantically related, as shown in Figure 11 (b).

Probabilistic models provide an opportunity to explore alternative representations for the meaning of words. One such representation is exploited in topic models, in which words are represented in terms of the set of topics to which they belong (Blei, Ng, & Jordan, 2003; Hofmann, 1999; Griffiths & Steyvers, 2004). Each topic is a probability distribution over words, and the content of the topic is reflected in the words to which it assigns high probability. For example, high probabilities for WOODS and STREAM would suggest a topic refers to the countryside, while high probabilities for FEDERAL and RESERVE would suggest a topic refers to finance. Each word will have a probability under each of these different topics, as shown in Figure 11 (c). For example, MEADOW has a relatively high probability under the countryside topic, but a low probability under the finance topic, similar to WOODS and STREAM.

Representing word meanings using probabilistic topics makes it possible to use

Bayesian inference to answer some of the critical problems that arise in processing language. In particular, we can make inferences about which semantically related concepts are likely to arise in the context of an observed set of words or sentences, in order to facilitate subsequent processing. Let $z$ denote the dominant topic in a particular context, and $w_1$ and $w_2$ be two words that arise in that context. The semantic content of these words is encoded through a set of probability distributions that identify their probability under different topics: if there are $T$ topics, then these are the distributions $P(w|z)$ for $z = \{1, \ldots, T\}$. Given $w_1$, we can infer which topic $z$ was likely to have produced it by using Bayes' rule,

$$P(z|w_1) = \frac{P(w_1|z)P(z)}{\sum_{z'=1}^{T} P(w_1|z')P(z')} \tag{36}$$

where $P(z)$ is a prior distribution over topics. Having computed this distribution over topics, we can make a prediction about future words by summing over the possible topics,

$$P(w_2|w_1) = \sum_{z=1}^{T} P(w_2|z)P(z|w_1). \tag{37}$$

A topic-based representation can also be used to disambiguate words: if BANK occurs in the context of STREAM, it is more likely that it was generated from the bucolic topic than the topic associated with finance.

Probabilistic topic models are an interesting alternative to traditional approaches to semantic representation, and in many cases actually provide better predictions of human behavior (Griffiths & Steyvers, 2003; Griffiths, Steyvers, & Tenenbaum, in press). However, one critical question in using this kind of representation is that of which topics should be used. Fortunately, work in machine learning and information retrieval has provided an answer to this question. As with popular semantic space models (Landauer & Dumais, 1997; Lund & Burgess, 1996), the representation of a set of words in terms of topics can be inferred automatically from the text contained in large document collections. The key to this process is viewing topic models as generative models for documents, making it possible to use standard methods of Bayesian statistics to identify a set of topics that likely to have generated an observed collection of documents. Figure 12 shows a sample of topics inferred from the TASA corpus (Landauer & Dumais, 1997), a collection of passages excerpted from educational texts used in curricula from the first year of school to the first year of college.

We can specify a generative model for documents by assuming that each document is a mixture of topics, with each word in that document being drawn from a particular topic, and the topics varying in probability across documents. For any particular document, we write the probability of a word $w$ in that document as

$$P(w) = \sum_{z=1}^{T} P(w|z)P(z), \tag{38}$$

where $P(w|z)$ is the probability of word $w$ under topic $z$, which remains constant across all documents, and $P(z)$ is the probability of topic $j$ in this document. We can summarize these probabilities with two sets of parameters, taking $\phi_w^{(z)}$ to indicate $P(w|z)$, and $\theta_z^{(d)}$ to indicate $P(z)$ in a particular document $d$. The procedure for generating a collection of

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| PRINTING | **PLAY** | TEAM | JUDGE | HYPOTHESIS | STUDY | **CLASS** | ENGINE |
| PAPER | PLAYS | GAME | TRIAL | EXPERIMENT | **TEST** | MARX | FUEL |
| PRINT | STAGE | BASKETBALL | **COURT** | SCIENTIFIC | STUDYING | ECONOMIC | ENGINES |
| PRINTED | AUDIENCE | PLAYERS | CASE | OBSERVATIONS | HOMEWORK | CAPITALISM | STEAM |
| TYPE | THEATER | PLAYER | JURY | SCIENTISTS | NEED | CAPITALIST | GASOLINE |
| PROCESS | ACTORS | **PLAY** | ACCUSED | EXPERIMENTS | **CLASS** | SOCIALIST | AIR |
| INK | DRAMA | PLAYING | GUILTY | SCIENTIST | MATH | SOCIETY | **POWER** |
| PRESS | SHAKESPEARE | SOCCER | DEFENDANT | EXPERIMENTAL | TRY | SYSTEM | COMBUSTION |
| IMAGE | ACTOR | PLAYED | JUSTICE | **TEST** | TEACHER | **POWER** | DIESEL |
| PRINTER | THEATRE | BALL | **EVIDENCE** | METHOD | WRITE | RULING | EXHAUST |
| PRINTS | PLAYWRIGHT | TEAMS | WITNESSES | HYPOTHESES | PLAN | SOCIALISM | MIXTURE |
| PRINTERS | PERFORMANCE | BASKET | CRIME | TESTED | ARITHMETIC | HISTORY | GASES |
| COPY | DRAMATIC | FOOTBALL | LAWYER | **EVIDENCE** | ASSIGNMENT | POLITICAL | CARBURETOR |
| COPIES | COSTUMES | SCORE | WITNESS | BASED | PLACE | SOCIAL | GAS |
| FORM | COMEDY | **COURT** | ATTORNEY | OBSERVATION | STUDIED | STRUGGLE | COMPRESSION |
| OFFSET | TRAGEDY | GAMES | HEARING | SCIENCE | CAREFULLY | REVOLUTION | JET |
| GRAPHIC | **CHARACTERS** | TRY | INNOCENT | FACTS | DECIDE | WORKING | BURNING |
| SURFACE | SCENES | COACH | DEFENSE | DATA | IMPORTANT | PRODUCTION | AUTOMOBILE |
| PRODUCED | OPERA | GYM | CHARGE | RESULTS | NOTEBOOK | CLASSES | STROKE |
| **CHARACTERS** | PERFORMED | SHOT | CRIMINAL | EXPLANATION | REVIEW | BOURGEOIS | INTERNAL |

*Figure 12.* A sample of topics from a 1700 topic solution derived from the TASA corpus. Each column contains the 20 highest probability words in a single topic, as indicated by $P(w|z)$. Words in boldface occur in different senses in neighboring topics, illustrating how the model deals with polysemy and homonymy. These topics were discovered in a completely unsupervised fashion, using just word-document co-occurrence frequencies.

documents is then straightforward. First, we generate a set of topics, sampling $\phi^{(z)}$ from some prior distribution $p(\phi)$. Then for each document $d$, we generate the weights of those topics, sampling $\theta^{(d)}$ from a distribution $p(\theta)$. Assuming that we know in advance how many words will appear in the document, we then generate those words in turn. A topic $z$ is chosen for each word that will be in the document by sampling from the distribution over topics implied by $\theta^{(d)}$. Finally, the identity of the word $w$ is determined by sampling from the distribution over words $\phi^{(z)}$ associated with that topic.

To complete the specification of our generative model, we need to specify distributions for $\phi$ and $\theta$ so that we can make inferences about these parameters from a corpus of documents. As in the case of coinflipping, calculations can be simplified by using a conjugate prior. Both $\phi$ and $\theta$ are arbitrary distributions over a finite set of outcomes, or *multinomial distributions*, and the conjugate prior for the multinomial distribution is the Dirichlet distribution. Just as the multinomial distribution is a multivariate generalization of the Bernoulli distribution we used in the coinflipping example, the Dirichlet distribution is a multivariate generalization of the beta distribution. We assume that the number of "virtual examples" of instances of each topic appearing in each document is set by a parameter $\alpha$, and likewise use a parameter $\beta$ to represent the number of instances of each word in each topic. Figure 13 shows a graphical model depicting the dependencies among these variables. This model, known as Latent Dirichlet Allocation, was introduced in machine learning by Blei, Ng, and Jordan (2003).

We extract a set of topics from a collection of documents in a completely unsupervised fashion, using Bayesian inference. Since the Dirichlet priors are conjugate to the multinomial distributions $\phi$ and $\theta$, we can compute the joint distribution $P(\mathbf{w}, \mathbf{z})$ by integrating out $\phi$ and $\theta$, just as we did in the model selection example above (Equation 16). We can then ask questions about the posterior distribution over $\mathbf{z}$ given $\mathbf{w}$, given by Bayes rule:

$$P(\mathbf{z}|\mathbf{w}) = \frac{P(\mathbf{w}, \mathbf{z})}{\sum_{\mathbf{z}} P(\mathbf{w}, \mathbf{z})}. \tag{39}$$

Since the sum in the denominator is intractable, having $T^n$ terms, and we are forced to

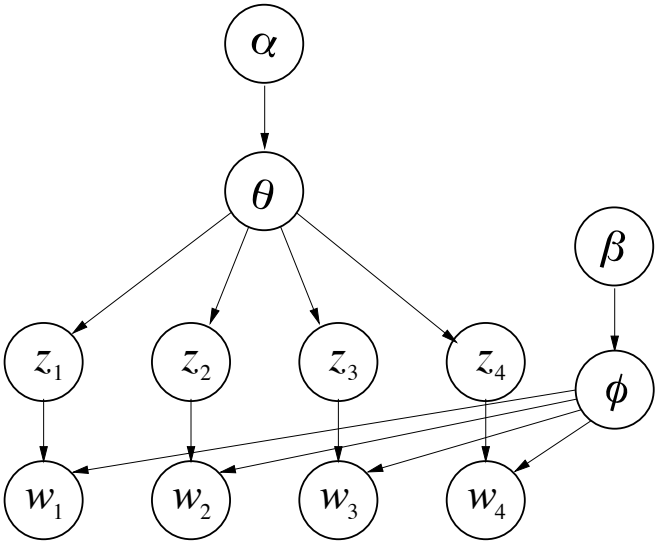


*Figure 13.* Graphical model for Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003). The distribution over words given topics, $\phi$, and the distribution over topics in a document, $\theta$, are generated from Dirichlet distributions with parameters $\beta$ and $\alpha$ respectively. Each word in the document is generated by first choosing a topic $z_i$ from $\theta$, and then choosing a word according to $\phi^{(z_i)}$.

evaluate this posterior using Markov chain Monte Carlo. In this case, we use Gibbs sampling to investigate the posterior distribution over assignments of words to topics, $\mathbf{z}$.

The Gibbs sampling algorithm consists of choosing an initial assignment of words to topics (for example, choosing a topic uniformly at random for each word), and then sampling the assignment of each word $z_i$ from the conditional distribution $P(z_i|\mathbf{z}_{-i}, \mathbf{w})$. Each iteration of the algorithm is thus a probabilistic shuffling of the assignments of words to topics. This procedure is illustrated in Figure 14. The figure shows the results of applying the algorithm (using just three topics) to a small portion of the TASA corpus. This portion features 30 documents that use the word MONEY, 30 documents that use the word OIL, and 30 documents that use the word RIVER. The vocabulary is restricted to 18 words, and the entries indicate the frequency with which the 731 tokens of those words appeared in the 90 documents. Each word token in the corpus, $w_i$, has a topic assignment, $z_i$, at each iteration of the sampling procedure. In the figure, we focus on the tokens of three words: MONEY, BANK, and STREAM. Each word token is initially assigned a topic at random, and each iteration of MCMC results in a new set of assignments of tokens to topics. After a few iterations, the topic assignments begin to reflect the different usage patterns of MONEY and STREAM, with tokens of these words ending up in different topics, and the multiple senses of BANK.

The details behind this particular Gibbs sampling algorithm are given in Griffiths and Steyvers (2004), where the algorithm is used to analyze the topics that appear in a large database of scientific documents. The conditional distribution for $z_i$ that is used in the algorithm can be derived using an argument similar to our derivation of the posterior

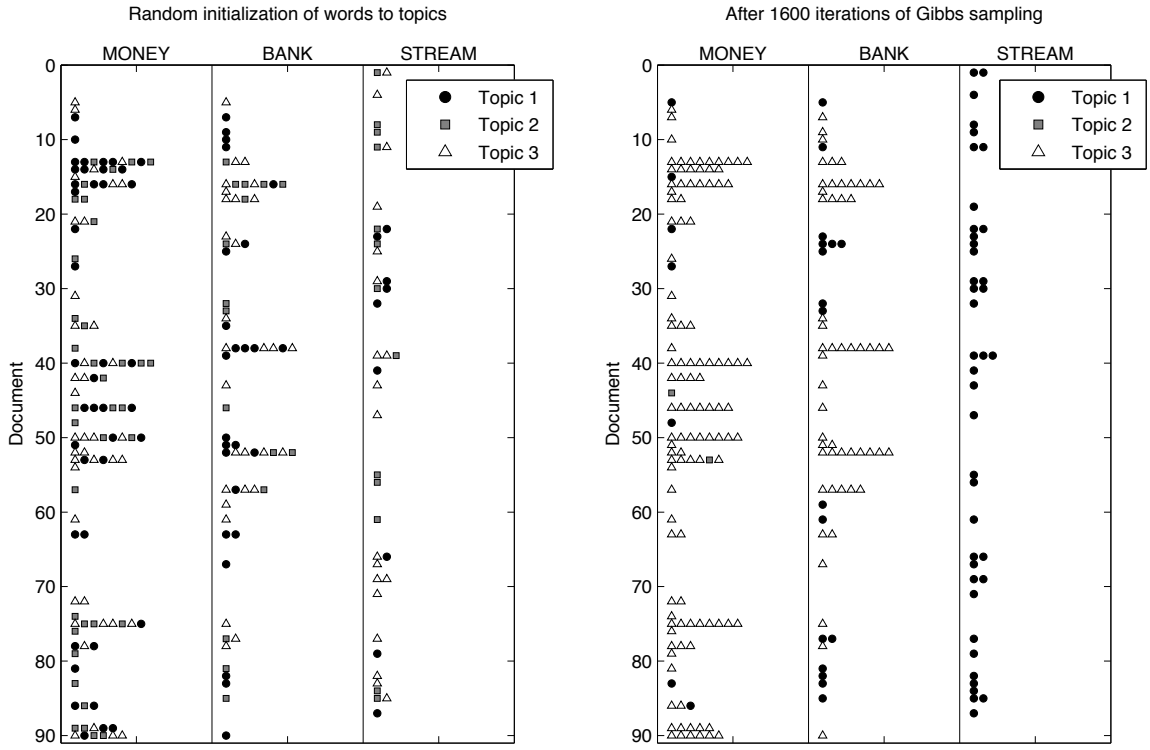


*Figure 14.* Illustration of the Gibbs sampling algorithm for learning topics. Each word token $w_i$ appearing in the corpus has a topic assignment, $z_i$. The figure shows the assignments of all tokens of three types – MONEY, BANK, and STREAM – before and after running the algorithm. Each marker corresponds to a single token appearing in a particular document, and shape and color indicates assignment: topic 1 is a black circle, topic 2 is a gray square, and topic 3 is a white triangle. Before running the algorithm, assignments are relatively random, as shown in the left panel. After running the algorithm, tokens of MONEY are almost exclusively assigned to topic 3, tokens of STREAM are almost exclusively assigned to topic 1, and tokens of BANK are assigned to whichever of topic 1 and topic 3 seems to dominate a given document. The algorithm consists of iteratively choosing an assignment for each token, using a probability distribution over tokens that guarantees convergence to the posterior distribution over assignments.

predictive distribution in coinflipping, giving

$$P(z_i|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n^{(w_i)}_{-i,z_i} + \beta}{n^{(\cdot)}_{-i,z_i} + W\beta} \frac{n^{(d_i)}_{-i,z_i} + \alpha}{n^{(d_i)}_{-i,\cdot} + T\alpha}, \tag{40}$$

where $\mathbf{z}_{-i}$ is the assignment of all $z_k$ such that $k \neq i$, and $n^{(w_i)}_{-i,z_i}$ is the number of words assigned to topic $z_i$ that are the same as $w_i$, $n^{(\cdot)}_{-i,z_i}$ is the total number of words assigned to topic $z_i$, $n^{(d_i)}_{-i,z_i}$ is the number of words from document $d_i$ assigned to topic $z_i$, and $n^{(d_i)}_{-i,\cdot}$ is the total number of words in document $d_i$, all not counting the assignment of the current word $w_i$. The two terms in this expression have intuitive interpretations, being the posterior predictive distributions on words within a topic and topics within a document given the current assignments $\mathbf{z}_{-i}$ respectively. The result of the MCMC algorithm is a set of samples from $P(\mathbf{z}|\mathbf{w})$, reflecting the posterior distribution over topic assignments given a collection of documents. A single sample can be used to evaluate the topics that appear in a corpus, as shown in Figure 12, or the assignments of words to topics, as shown in Figure 14. We can also compute quantities such as the strength of association between words (given by Equation 37) by averaging over many samples.[5]

While other inference algorithms exist that can be used with this generative model (e.g., Blei et al., 2003; Minka & Lafferty, 2002), the Gibbs sampler is an extremely simple (and reasonably efficient) way to investigate the consequences of using topics to represent semantic relationships between words. Griffiths and Steyvers (2002, 2003) suggested that topic models might provide an alternative to traditional approaches to semantic representation, and showed that they can provide better predictions of human word association data than Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997). Topic models can also be applied to a range of other tasks that draw on semantic association, such as semantic priming and sentence comprehension (Griffiths et al., in press).

The key advantage that topic models have over semantic space models is postulating a more structured representation – different topics can capture different senses of words, allowing the model to deal with polysemy and homonymy in a way that is automatic and transparent. For instance, similarity in semantic space models must obey a version of the triangle inequality for distances: if there is high similarity between words $w_1$ and $w_2$, and between words $w_2$ and $w_3$, then $w_1$ and $w_3$ must be at least fairly similar. But word associations often violate this rule. For instance, ASTEROID is highly associated with BELT, and BELT is highly associated with BUCKLE, but ASTEROID and BUCKLE have little association. LSA thus has trouble representing these associations. Out of approximately 4500 words in a large-scale set of word association norms (Nelson, McEvoy, & Schreiber, 1998), LSA judges that BELT is the 13th most similar word to ASTEROID, that BUCKLE is

[5]When computing quantities such as $P(w_2|w_1)$, as given by Equation 37, we need a way of finding the parameters $\phi$ that characterize the distribution over words associated with each topic. This can be done using ideas similar to those applied in our coinflipping example: for each samples of $\mathbf{z}$ we can estimate $\phi$ as

$$\hat{\phi}^{(w)}_z = \frac{n^{(w)}_z + \beta}{n^{(\cdot)}_z + W\beta} \tag{41}$$

which is the posterior predictive distribution over new words $w$ for topic $z$ conditioned on $\mathbf{w}$ and $\mathbf{z}$.

the second most similar word to BELT, and consequently BUCKLE is the 41st most similar word to ASTEROID – more similar than TAIL, IMPACT, or SHOWER. In contrast, using topics makes it possible to represent these associations faithfully, because BELT belongs to multiple topics, one highly associated with ASTEROID but not BUCKLE, and another highly associated with BUCKLE but not ASTEROID.

The relative success of topic models in modeling semantic similarity is thus an instance of the capacity for probabilistic models to combine structured representations with statistical learning – a theme that has run through all of the examples we have considered in this chapter. The same capacity makes it easy to extend these models to capture other aspects of language. As generative models, topic models can be modified to incorporate richer semantic representations such as hierarchies (Blei et al., 2004), as well as rudimentary syntax (Griffiths, Steyvers, Blei, & Tenenbaum, 2005), and extensions of the Markov chain Monte Carlo algorithm described in this section make it possible to sample from the posterior distributions induced by these models.

## 6. Conclusion

Our aim in this chapter has been to survey the conceptual and mathematical foundations of Bayesian models of cognition, and to introduce several advanced techniques that are driving state-of-the-art research. We have had space to discuss only a few specific and rather simple cognitive models based on these ideas, but much more can be found in the current literature referenced in the introduction. These Bayesian models of cognition represent just one side of a larger movement that seeks to understand intelligence in terms of rational probabilistic inference. Related ideas are providing new paradigms for the study of neural coding and computation (Doya, Ishii, Pouget, & Rao, 2007), children's cognitive development (Gopnik & Tenenbaum, in press), machine learning (Bishop, 2006) and artificial intelligence (Russell & Norvig, 2002).

We hope that this chapter conveys some sense of what all this excitement is about – or at least why we find this line of work exciting. Bayesian models give us ways to approach deep questions of human cognition that have not been previously amenable to rigorous formal study. How can human minds make predictions and generalizations from such limited data, and so often be correct? How can structured representations of abstract knowledge constrain and guide sophisticated statistical inferences from sparse data? What specific forms of knowledge support human inductive inference, across different domains and tasks? How can these structured knowledge representations themselves be acquired from experience? And how can the necessary computations be carried out or approximated tractably for complex models that might approach the scale of interesting chunks of human cognition? We are still far from having good answers to these questions, but as this chapter shows, we are beginning to see what answers might look like and to have the tools needed to start building them.

## Acknowledgements

This chapter is based in part on tutorials given by the authors at the Annual Meeting of the Cognitive Science Society in 2004 and 2006, and on portions of a tutorial on probabilistic inference written by Thomas L. Griffiths and Alan Yuille that appeared as an online

supplement to the special issue of *Trends in Cognitive Sciences* on Probabilistic Models of Cognition (Volume 10, Issue 7). We thank the participants in those tutorials and the special issue for their feedback on this material. The writing of this chapter was supported in part by grants from the James S. McDonnell Foundation Causal Learning Research Collaborative, the DARPA BICA program, the National Science Foundation (TLG), the Air Force Office of Scientific Research (JBT, TLG), the William Asbjornsen Albert fellowship (CK), and the Paul E. Newton Career Development Chair (JBT).

## References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, *9*, 147-169.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*, 216-233.

Atran, S. (1998). Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences*, *21*, 547-609.

Baker, C. L., Tenenbaum, J. B., & Saxe, R. R. (2007). Goal inference as inverse planning. In *Proceedings of the 29th annual meeting of the cognitive science society.*

Bayes, T. (1763/1958). Studies in the history of probability and statistics: IX. Thomas Bayes's Essay towards solving a problem in the doctrine of chances. *Biometrika*, *45*, 296-315.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory.* New York: Wiley.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* New York: Springer.

Blei, D., Griffiths, T., Jordan, M., & Tenenbaum, J. (2004). Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems 16.* Cambridge, MA: MIT Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Boas, M. L. (1983). *Mathematical methods in the physical sciences* (2nd ed.). New York: Wiley.

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

Buehner, M., & Cheng, P. W. (1997). Causal induction: The Power PC theory versus the Rescorla-Wagner theory. In M. Shafto & P. Langley (Eds.), *Proceedings of the 19th Annual Conference of the Cognitive Science Society* (p. 55-61). Hillsdale, NJ: Lawrence Erlbaum Associates.

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1119-1140.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Charniak, E. (1993). *Statistical language learning.* Cambridge, MA: MIT Press.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*, 335-344.

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367-405.

Chomsky, N. (1988). *Language and problems of knowledge: The managua lectures.* MIT Press.

Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review*, *82*, 407-428.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behaviour*, *8*, 240-247.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*, 294-300.

Danks, D., Griffiths, T. L., & Tenenbaum, J. B. (2003). Dynamical causal learning. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances Neural Information Processing Systems 15* (p. 67-74). Cambridge, MA: MIT Press.

Doya, K., Ishii, S., Pouget, A., & Rao, R. P. N. (Eds.). (2007). *The Bayesian brain: Probabilistic approaches to neural coding.* Cambridge, MA: MIT Press.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification.* New York: Wiley.

Friedman, N., & Koller, D. (2000). Being Bayesian about network structure. In *Proceedings of the 16th annual conference on uncertainty in ai* (p. 201-210). Stanford, CA.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* New York: Chapman & Hall.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721-741.

Ghahramani, Z. (2004). Unsupervised learning. In O. Bousquet, G. Raetsch, & U. von Luxburg (Eds.), *Advanced lectures on machine learning.* Berlin: Springer-Verlag.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989). *The empire of chance.* Cambridge: Cambridge University Press.

Gilks, W., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice.* Suffolk, UK: Chapman and Hall.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology.* Cambridge, MA: MIT Press.

Glymour, C., & Cooper, G. (1999). *Computation, causation, and discovery.* Cambridge, MA: MIT Press.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.).

Good, I. J. (1980). Some history of the hierarchical Bayesian methodology. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp. 489–519). Valencia: Valencia University Press.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories.* Cambridge, MA: MIT Press.

Gopnik, A., & Tenenbaum, J. B. (in press). Bayesian networks, Bayesian learning, and cognitive development. *Developmental Science.*

Griffiths, T. L. (2005). *Causes, coincidences, and theories.* Unpublished doctoral dissertation, Stanford University.

Griffiths, T. L., Baraff, E. R., & Tenenbaum, J. B. (2004). Using physical theories to infer hidden causal structure. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th annual meeting of the cognitive science society* (p. 446-451). Mahwah, NJ: Erlbaum.

Griffiths, T. L., & Ghahramani, Z. (2005). *Infinite latent feature models and the Indian buffet process* (Tech. Rep. No. 2005-001). Gatsby Computational Neuroscience Unit.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Griffiths, T. L., & Steyvers, M. (2003). Prediction and semantic association. In *Neural information processing systems 15.* Cambridge, MA: MIT Press.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science*, *101*, 5228-5235.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. In *Advances in Neural Information Processing Systems 17.* Cambridge, MA: MIT Press.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (in press). Topics in semantic association. *Psychological Review.*

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354-384.

Griffiths, T. L., & Tenenbaum, J. B. (2007a). From mere coincidences to meaningful discoveries. *Cognition*, *103*, 180-226.

Griffiths, T. L., & Tenenbaum, J. B. (2007b). Two proposals for causal grammars. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation.* Oxford: Oxford University Press.

Hacking, I. (1975). *The emergence of probability.* Cambridge: Cambridg University Press.

Hagmayer, Y., Sloman, S. A., Lagnado, D. A., & Waldmann, M. R. (in press). Causal reasoning through intervention. In *Causal learning: Psychology, philosophy, and computation.* Oxford: Oxford University Press.

Hastings, W. K. (1970). Monte Carlo methods using Markov chains and their applications. *Biometrika*, *57*, 97-109.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In M. I. Jordan (Ed.), *Learning in graphical models* (p. 301-354). Cambridge, MA: MIT Press.

Heibeck, T., & Markman, E. (1987). Word learning in children: an examination of fast mapping. *Child Development*, *58*, 1021–1024.

Hofmann, T. (1999). Probablistic latent semantic indexing. In *Proceedings of the Twenty-Second Annual International SIGIR Conference.*

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*(8), 754-755.

Jeffreys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, *80*(1), 64-72.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs*, *79.*

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing.* Upper Saddle River, NJ: Prentice Hall.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2004). Learning domain structures. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (in press). Learning overhypotheses with hierarchical bayesian models. *Developmental Science.*

Kemp, C., & Tenenbaum, J. B. (2003). Theory-based induction. In *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society.*

Korb, K., & Nicholson, A. (2003). *Bayesian artificial intelligence.* Boca Raton, FL: Chapman and Hall/CRC.

Kording, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, *10*, 319-326.

Lagnado, D., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 856-876.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211-240.

Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science*, *30*, 555–580.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, *28*, 203-208.

Mackay, D. J. C. (2003). *Information theory, inference, and learning algorithms.* Cambridge: Cambridge University Press.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

Mansinghka, V. K., Kemp, C., Tenenbaum, J. B., & Griffiths, T. L. (2006). Structured priors for structure learning. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI).*

Marr, D. (1982). *Vision.* San Francisco, CA: W. H. Freeman.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207-238.

Metropolis, A. W., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087-1092.

Minka, T., & Lafferty, J. (2002). Expectation-Propagation for the generative aspect model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI).* San Francisco, CA: Morgan Kaufmann.

Myung, I. J., Forster, M. R., & Browne, M. W. (2000). Model selection [special issue]. *Journal of Mathematical Psychology*, *44*.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin and Review*, *4*, 79-95.

Neal, R. M. (1992). Connectionist learning of belief networks. *Artificial Intelligence*, *56*, 71-113.

Neal, R. M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods* (Tech. Rep. No. CRG-TR-93-1). University of Toronto.

Neal, R. M. (1998). *Markov chain sampling methods for Dirichlet process mixture models* (Tech. Rep. No. 9815). Department of Statistics, University of Toronto.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The university of south florida word association, rhyme, and word fragment norms.* (http://w3.usf.edu/FreeAssociation/)

Newman, M. E. J., & Barkema, G. T. (1999). *Monte carlo methods in statistical physics.* Oxford: Clarendon Press.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, *90*(4), 339–363.

Norris, J. R. (1997). *Markov chains.* Cambridge, UK: Cambridge University Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39-57.

Nosofsky, R. M. (1998). Optimal performance and exemplar models of classification. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (p. 218-247). Oxford: Oxford University Press.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, *5*, 349-357.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*(2), 185-200.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems.* San Francisco, CA: Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, reasoning and inference.* Cambridge, UK: Cambridge University Press.

Pitman, J. (1993). *Probability.* New York: Springer-Verlag.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 393-407.

Rice, J. A. (1995). *Mathematical statistics and data analysis* (2nd ed.). Belmont, CA: Duxbury.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior*, *14*, 665-681.

Russell, S. J., & Norvig, P. (2002). *Artificial intelligence: A modern approach* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives.* Oxford: Oxford University Press.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19.

Spirtes, P., Glymour, C., & Schienes, R. (1993). *Causation prediction and search.* New York: Springer-Verlag.

Steyvers, M., Griffiths, T. L., & Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences*, *10*, 327-334.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*, 453-489.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In T. Leen, T. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13* (p. 59-65). Cambridge, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal induction. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15* (p. 35-42). Cambridge, MA: MIT Press.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Science*, *10*, 309-318.

Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (2007). Intuitive theories as grammars for causal inference. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation.* Oxford: Oxford University Press.

Tenenbaum, J. B., & Niyogi, S. (2003). Learning causal laws. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th annual meeting of the cognitive science society.* Hillsdale, NJ: Erlbaum.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*, 337-375.

Xu, F., & Tenenbaum, J. B. (in press). Word learning as bayesian inference. *Psychological Review.*

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, *10*, 301-308.