Chapter 10

Estimating subjective probability distributions Thomas L. Griffiths, Adam N. Sanborn, Raja Marjieh, Thomas Langlois, Jing Xu, & Nori Jacoby

It should be clear from the previous chapters that the predictions of Bayesian models of cognition depend intimately upon the choice of particular probability distributions – how people learn will reflect the prior probabilities of different hypotheses, and how people categorize objects will be determined by the distributions that represent different categories. Estimating these distributions is thus an important part of defining Bayesian models.

Sometimes, we can measure these distributions from the world. For example, in the "predicting the future" experiments of Griffiths and Tenenbaum (2006), discussed in Chapter 3, prior distributions for various everyday quantities could be estimated from online datasets. Likewise, the explanation of the perceptual magnet effect offered by Feldman et al. (2009) highlighted in Chapter 5 required representing phonetic categories using a mixture model, but clues about the parameters of the mixture components could be taken from the human speech signal. Using distributions that are derived from the world is attractive because it minimizes the assumptions we have to make about how our subjective probabilities – our internal degrees of belief – might differ from the objective, measurable probabilities of the world around us. For this reason, Anderson (1990) recommended this approach in his definition of rational analysis, and demonstrated how it could be used to explain phenomena such as power-law curves for forgetting in terms of mimicking the statistical structure of our environment.

In other cases, we are interested in seeing what the consequences are of assuming different distributions, comparing the resulting models with human behavior. Griffiths and Tenenbaum (2006) used this approach to infer the form of people's prior distributions for phenomena such as waiting on the telephone when trying to buy tickets, a phenomenon for which it is hard to obtain objective data. From people's judgments – in thie case, the fact that the longer people waited the longer they expected their additional wait to be – they were able to diagnose that people assume such wait times follow a power-law distribution. Under this approach, we might define a model and then try to find the distribution that results in the best fit between that model and human behavior. This is an effective strategy, but faces two kinds of risks. One risk is *underfitting* human behavior, because there are relatively few parametric families of distributions and the particular distribution that best captures human behavior may not be in one of these families. The other risk is *overfitting* human behavior, and ending up with a distribution that captures performance on the particular task being modeled, but doesn't generalize to other closely related tasks.

In this chapter we consider a different approach to estimating subjective probability distributions, focused on designing novel experimental methods for measuring those distributions directly. The key idea is to design experiments that let us *sample* from subjective probability distributions. In this way, we can form an estimate of the distribution from the samples. We begin by summarizing standard methods for eliciting subjective probability distributions that have been used in statistics, and then turn to a set of experimental methods based on the sampling algorithms introduced in Chapter 6.

10.1 Elicitation of probabilities

Statisticians, social scientists, and computer scientists often need to capture people's beliefs about a continuous quantity in the form of a probability distribution. To solve this problem, they have developed a variety of **elicitation methods** which use a combination of asking people quantitative questions and then inferring a distribution that corresponds as closely as possible to the answers (see Garthwaite, Kadane, & O'Hagan, 2005 and O'Hagan et al., 2006 for reviews).

A standard approach to elicitation is to ask people to provide quantiles for quantities, or quantities for quantiles. For example, if the goal were to estimate the probability distribution that somebody assigned to the grosses of movies, this could be done by asking people to name a dollar amount corresponding to the lowest 5% of grosses, the lowest 10%, etc. Alternatively, people could be asked to assign a percentile rank to various dollar amounts, indicating where they think those dollar amounts fall in the overall distribution. Either set of questions will provide a set of numbers that can be used to approximate a cumulative density function, from which an estimate of a probability density function can be recovered.

These traditional elicitation methods can be effective in settings where the goal is to estimate a distribution over a single quantity, and they do not have any limitations in terms of the form of the resulting distribution (although specific schemes for analyzing the data, such as finding the Gaussian distribution that best corresponds to people's estimates, can introduce additional constraints). However, they have two weaknesses as a general method for estimating probability distributions to be used in Bayesian models of cognition.

First, traditional elicitation methods are only feasible to use for simple, low-dimensional quantities. The grosses of movies all fall along a single dimension – dollar amounts – and the corresponding distribution can be captured by a univariate probability density function. Even generalizing to two dimensions creates challenges in terms of assessing appropriate quantiles and quantities, although it's possible to navigate these. Higher-dimensional distributions over more complex spaces with no natural ordering or representational format, such as people's prior distributions over categories, functions, or causal relationships, lie outside the scope of these methods.

Second, these methods rely on people having veridical access to their subjective probabilities. For a one-dimensional quantity this might be a reasonable assumption, although there is plenty of evidence that asking people for explicit probability judgments can be problematic (e.g., Tversky & Kahneman, 1974), which is one reason why the experiments presented in this book typically try to avoid asking people to state probabilities. However, people might not have the same kind of access to the distributions that Bayesian models use to characterize prior probabilities or category representations. Measuring people's distribution over the physiognomy of things that fit into the category of dogs, or the prior probability that they assign to deterministic causal relationships, might be a challenge.

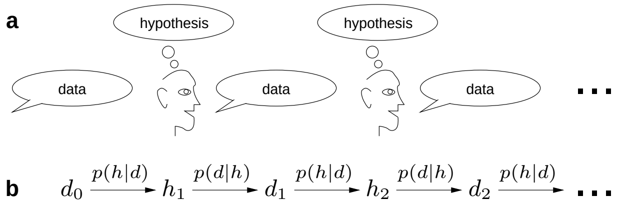
For this reason, the methods we present in the remainder of the chapter are designed to be effective for estimating subjective probability distributions over arbitrarily complex objects, using naturalistic judgments that do not require people to state subjective probabilities. They also make no assumption about the form of the underlying distribution. To do so, they leverage techniques that computer scientists and statisticians have developed for sampling from complex probability distributions. However, the inspiration for these methods came from neither of these disciplines – it came from linguistics.

10.2 Iterated learning

When a child learns a language, she learns it from speakers who in turn learned it from other speakers. Languages are transmitted via a process that has been called **iterated learning** (Kirby, 2001), being passed from learner to learner. Figure 10.1 (a) provides a schematic illustration of the simplest version of iterated learning, in which a language is passed along a single chain of learners. Each learner observes linguistic data generated by the previous learner, forms a hypothesis, and then generates data provided to the next learner based on that hypothesis.

A natural question to ask is how the process of transmission by iterated learning should be expected to influence the structure of languages. Figure 10.1 (b) shows that we can analyze this simple form of iterated learning as a Markov chain on data d and hypotheses h. If we assume that the learners are applying Bayesian inference, then the transition probabilities in this Markov chain result from sampling h from the posterior $p(h|d)$ and then d from the corresponding likelihood function $p(d|h)$.

Formulating iterated learning as a Markov chain allows us to ask the question of what the stationary distribution of this Markov chain might be. Recall that a Markov chain will converge to its stationary distribution, provided it satisfies the conditions for ergodicity (see Chapter 6). Griffiths and Kalish (2005; 2007) proved that if all the learners have the same prior, the stationary distribution on h is the prior



$$d_0 \xrightarrow{p(h|d)} h_1 \xrightarrow{p(d|h)} d_1 \xrightarrow{p(h|d)} h_2 \xrightarrow{p(d|h)} d_2 \xrightarrow{p(h|d)} \ldots$$

Figure 10.1: Iterated learning. (a) Information is passed down a chain of learners, with each learner forming a hypothesis based on data generated by the previous learner and then generating data in turn. (b) This process defines a Markov chain on data d and hypotheses h. With Bayesian learners, we sample h from $p(h|d)$ and then d from $p(d|h)$. Figure reproduced with permission from Griffiths and Kalish (2005).

distribution $p(h)$. In the context of language learning, this implies that over time we should expect languages to change to become easier to learn, conforming more closely with human inductive biases as reflected in this prior distribution.

This theoretical analysis potentially has interesting implications for understanding cultural transmission, but needs to be validated empirically. Kalish et al. (2007) conducted an experiment that provides a good test of the theory, using the function learning task discussed in Chapter 9. People were taught a relationship between two variables, represented with colored bars on a computer screen. They saw 50 pairs of values for these variables, and then were asked to generate 50 predictions of the value of one variable given the other. These 50 predictions were taken as the data for the next participant, creating an iterated learning chain.

Function learning provides a good test of the theory because it is a case where people's inductive biases are well understood. Decades of research on human function learning has shown that people find it easiest to learn positive linear functions (ie. functions that are linear with positive slope), followed by negative linear functions, followed by nonlinear functions. We can translate this information into a prior distribution. If a hypothesis has higher prior probability, then it should require less data consistent with that hypothesis in order to end up with a high posterior probability – it should be easier to learn. Thus, we should expect positive linear functions to have high probability under people's prior on functions. Consequently, the analysis of iterated learning given by Griffiths and Kalish (2007) predicts that we should expect positive linear functions to emerge with high probability from iterated function learning.

Figure 10.2 shows the results from Kalish et al. (2007). Regardless of how chains were initialized, they were dominated by positive linear functions after just nine iterations of transmission. These results provide strong support for the idea that iterated learning produces outcomes that are consistent with people's inductive biases. The positive feedback process that it establishes – where the initial data is repeatedly passed through a biased learning system – magnifies those biases significantly. It's possible to detect that people find it easier to learn positive linear functions by looking at the first iteration – there are fewer errors for the positive linear function, and the errors that people make on other functions tend towards positive linear – but it is much more obvious in the final iteration.

These results raise another possibility: that we could use iterated learning as an experimental paradigm for measuring people's prior distributions. There is no need to actually have information transmitted between participants – we can still construct a Markov chain within participants, with each person seeing a sequence of trials in which the stimuli seen on subsequent trials are determined by their responses on previous trials.

Lewandowsky et al. (2009) explored this possibility in an experiment designed to measure people's priors for everyday quantities using iterated learning. The task was the "predicting the future" problem

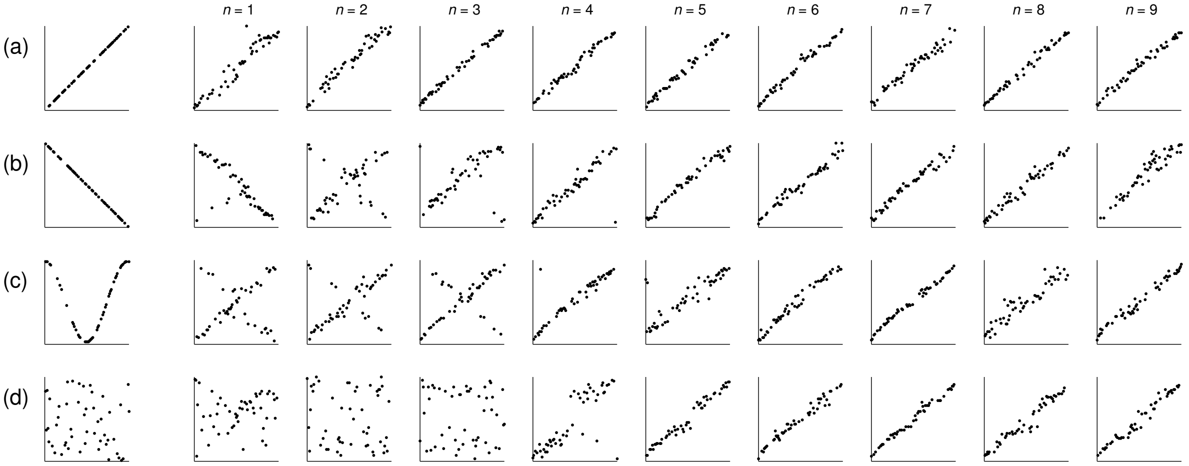


Figure 10.2: Iterated function learning. Each row shows the results produced by one chain of learners, initialized with different functions (shown in the first column). Each column is the predictions of a single learner, which become the training data for the next learner in the row. Regardless of initialization, after just nine iterations the functions have transformed into positive linear functions. Kalish et al. (2007) ran 32 such chains, of which 28 finished as positive linear functions and 4 finished as negative linear functions. Figure adapted from Kalish et al. (2007).

discussed in Chapter 3: given an observed quantity so far, t, such as the amount of money a movie has made, predict the total, t_{total}, such as the total gross for the movie. Those values of t_{total} could be used to generate stimuli on the next trial by using assumption of uniform sampling in the likelihood of the corresponding Bayesian model, $p(t|t_{\text{total}})$. In this case, that means sampling the next value of t uniformly from between 0 and t_{total}. If people's responses are samples from $p(t_{\text{total}}|t)$, then over time the resulting Markov chain will converge to the distribution $p(t_{\text{total}})$.

Figure 10.3 shows the estimated stationary distributions produced by applying iterated learning to the predicting the future task. The stationary distributions were estimated by aggregating the last half of all chains across participants. There is a close correspondence between the true distributions of these quantities and the estimated stationary distributions, consistent with the hypothesis that iterated learning can be used to estimate human prior distributions. Subsequent work has used the same approach to estimate prior distributions on concepts (Griffiths, Christian, & Kalish, 2008; Canini, Griffiths, Vanpaemel, & Kalish, 2014) and causal relationships (Yeung & Griffiths, 2015).

The priors inferred by iterated learning can be quite revealing, and can improve the predictions of Bayesian models. Figure 10.4 shows inferred priors on the parameters of the Noisy-OR and Noisy-AND-NOT functions used in the causal models for elemental causal induction (see Chapter 4). The Noisy-OR is used to capture people's assumptions about generative causes, where the cause increases the probability of the effect, and the Noisy-AND-NOT corresponds to preventive causes, where the cause decreases the probability of the effect. In both functions, w_0 is the background rate of the effect and w_1 is the strength of the cause. The joint distribution over these two parameters tells us what people's expectations are about the rate at which effects occur and the assumed strength of causes. Yeung and Griffiths (2015) identified this distribution using an iterated causal learning task, in which people saw contingency data, estimated w_0 and w_1, and then subsequently saw new contingency data generated using the probability distribution that resulted from their estimates. Figure 10.4 shows the stationary distribution of this process – our best estimate of people's prior on w_0 and w_1 – for generative and preventive causes. The results suggest that people are relatively indifferent about the background rates at which effects occur (the distribution on w_0 is roughly uniform), but expect causal relationships to be relatively strong (the

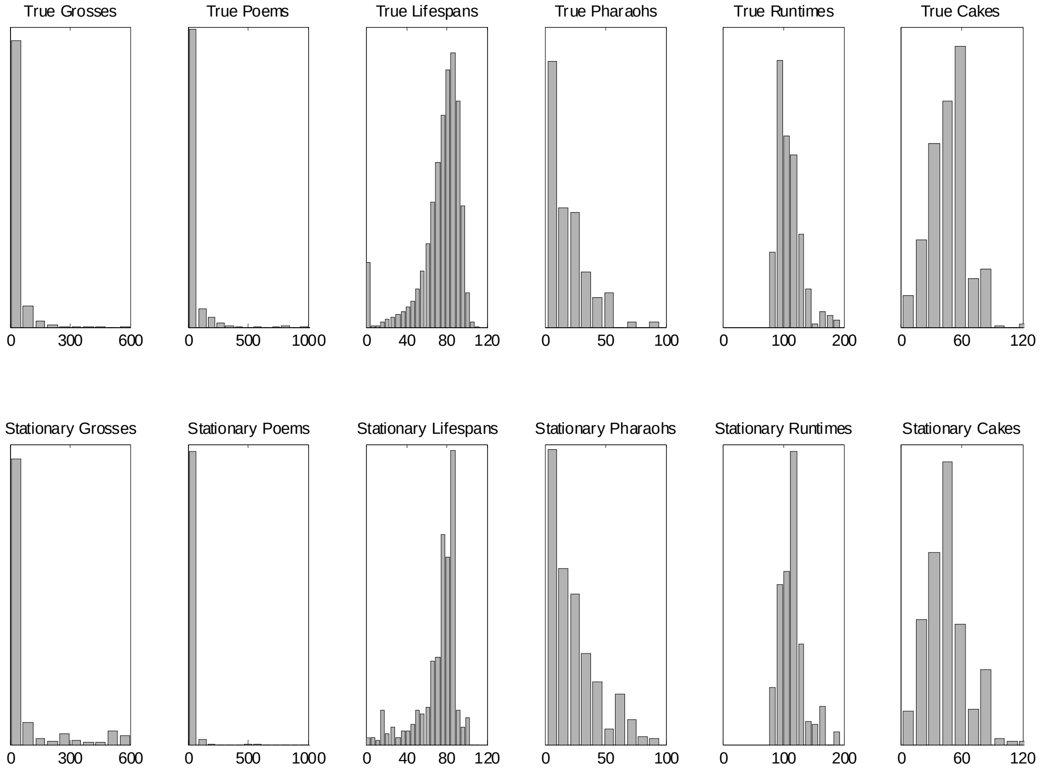


Figure 10.3: Predicting the future. The top row shows the actual distributions of a set of everyday quantities. The bottom row shows the stationary distributions over those quantities resulting from applying the iterated learning paradigm to the predicting the future task of Lewandowsky et al. (2009). Figure reproduced with permission from Lewandowsky et al. (2009).

distribution assigns higher probabilities to higher values of w_1). Taking this into account results in models that fit human judgments about causal relationships better than those that use other priors (Yeung & Griffiths, 2015).

While we have focused on inferring priors here, it is worth noting that iterated learning can be used to study the expected effects of human perception, learning, and memory on the cultural artifacts that are transmitted across generations. Xu, Dowman, and Griffiths (2013) studied the effect of cultural transmission on systems of color terms (see Figure 10.5) using an array of Munsell color chips (see Figure 10.5A) that were originally used to collect a large cross-cultural sample of systems of color terms from non-industrial societies in the World Color Survey (Kay et al., 2009). Participants were initially presented with a random subset of colors (Figure 10.5B top) which were classified into arbitrary categories labeled with pseudo-words, and were then asked to generalize what they had learned for the remaining colors (i.e., categorize new colors). The results of one generation of learners became the input for the next generation. Importantly, the number of color terms varied from one condition to another, simulating how the number of "basic" color terms varies across languages. Remarkably, a striking similarity emerged

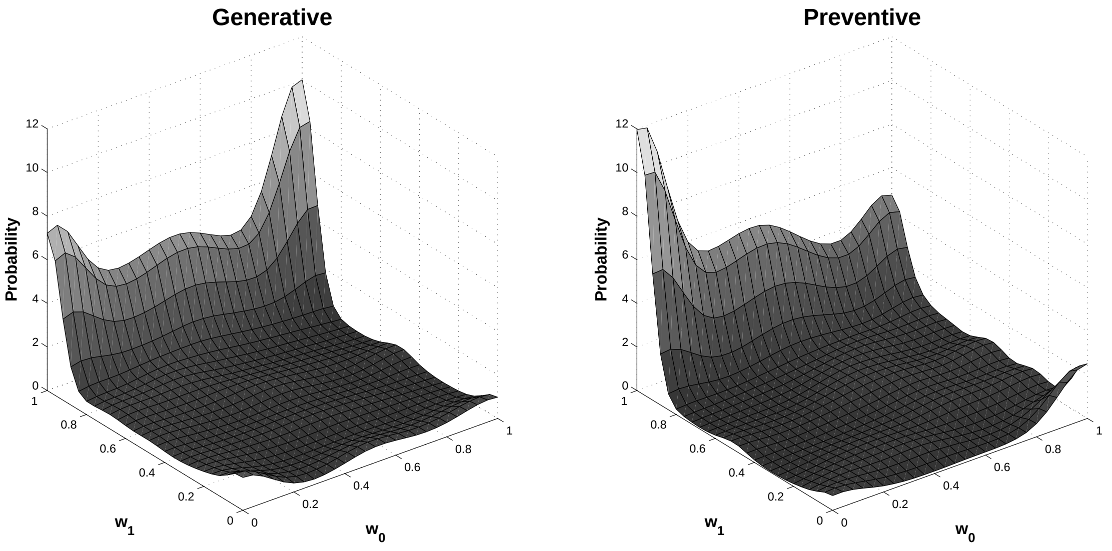


Figure 10.4: Smoothed empirical estimates of human priors on the parameters of Bayesian networks used for elementary causal induction – the background rate of the effect occurring, w_0, and strength of the cause, w_1 – produced by iterated learning. Figure reproduced with permission from Yeung and Griffiths (2015).

between the resulting artificial color systems and those found in different cultures around the world after as few as 13 iterations (Compare Figure 10.5B and Figure 10.5C).

One way to understand why iterated learning converges to the prior is to recognize that it is a form of Gibbs sampling (see Chapter 6). In Gibbs sampling, we construct a Markov chain that converges to a particular stationary distribution on a set of variables by iteratively sampling a value for each variable from its conditional distribution given the current values of all other variables. If the Bayesian learners have a prior $p(h)$ and a likelihood function $p(d|h)$, we can define the joint distribution $p(d, h) = p(d|h)p(h)$. It is then easy to recognize that sampling from the posterior $p(h|d)$ and then the likelihood $p(d|h)$ corresponds to iteratively sampling from the two conditional distributions of this joint distribution. As a result, the distribution on d and h will converge to $p(d, h)$ over time, and the marginal distribution on h will converge to $p(h)$.

This gives us another way to think about iterated learning as an experimental method: it's an implementation of a Gibbs sampling algorithm, in which samples are generated by people rather than the computer. This establishes a link between sampling algorithms and experimental paradigms that can potentially be used to convert methods that computer scientists use to generate samples from distributions represented by computers into methods that cognitive scientists can use to generate samples from subjective distributions inside people's heads.

10.3 Serial reproduction

Methods closely related to iterated learning have previously been used in psychology to study the effects of human cognition on the cultural transmission of information. The most famous of these methods is the **serial reproduction** paradigm introduced by Bartlett (1932). In this paradigm, a participant

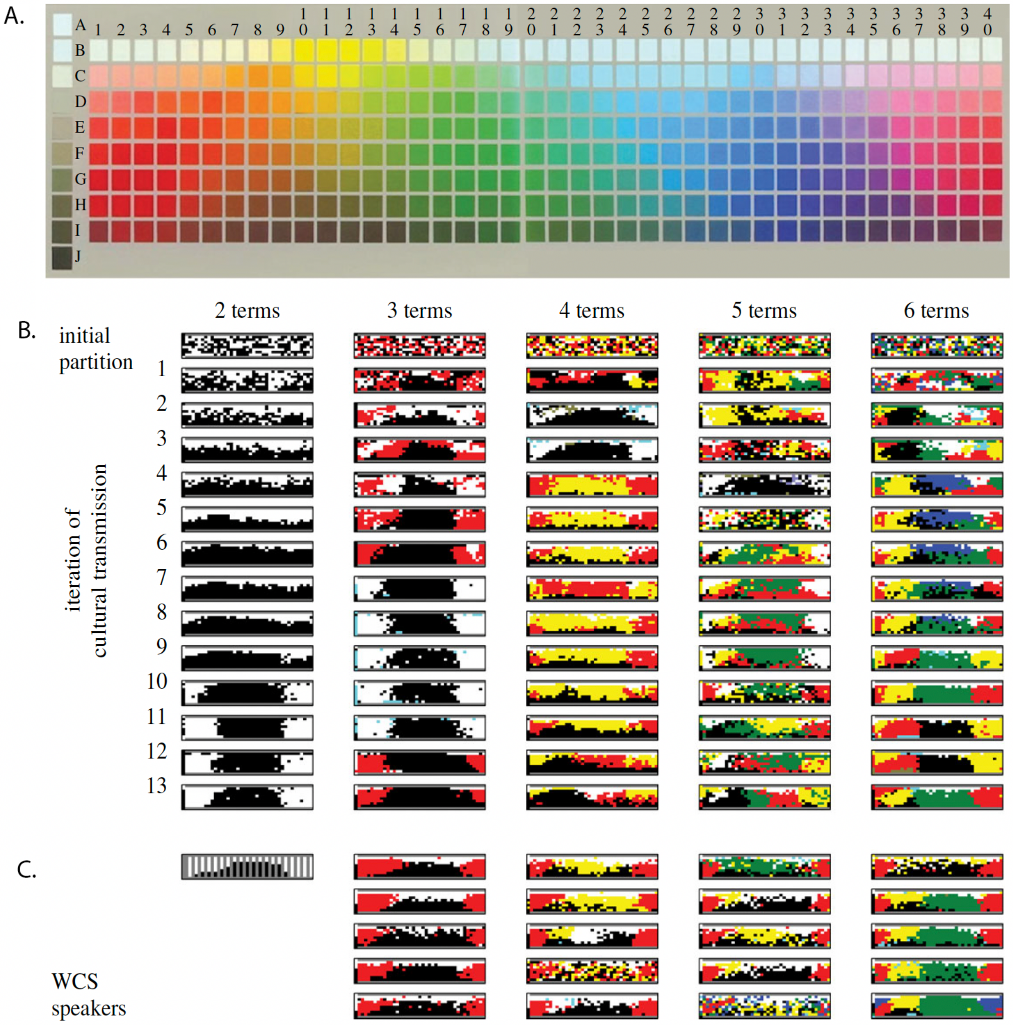


Figure 10.5: Simulating the cultural transmission of color term systems. A. The stimulus set is based on that used in the World Color Survey, which collected systems of color terms from the languages of 110 non-industrial societies (Kay et al., 2009) B. Iterated learning chains for the different term divisions. C. Sample speakers from World Color Survey (WCS) languages that are similar to the experimental results. The results for two terms show an estimated partition for the Dani language, as no two-term languages appear in the WCS. Figure reproduced with permission from Xu et al. (2013).

is shown a stimulus, such as a story or image, and reproduces it from memory after a delay. The reproduction produced by that first participant is then shown to a second participant, who produces their own reproduction. As this process is repeated the stimulus changes significantly.

Perception and memory are noisy processes; images and sounds are rarely transmitted without distortion, and first impressions of stories and images are far from permanent. To deal with such uncertainties, humans often rely on prior information to fill the gaps. The reliance on prior information will often lead to biases, where the average participant's response deviates from the real response. Such prior information may reflect the distribution of stimuli in the world, x (Jacoby & McDermott, 2017; Langlois, Jacoby, Suchow, & Griffiths, 2021), or the states people infer from those stimuli, μ (Xu & Griffiths, 2010). Serial reproduction capitalizes on this observation to construct a process that allows for the precise characterization of human prior information. By repeatedly observing and reproducing stimuli from memory, systematic biases due to internalized priors can build up and become manifest. In other words, serial reproduction amplifies biases of perception and memory in a way that reveals the shared priors that generate them.

Formally, serial reproduction implements a Gibbs sampler in the form of a Markov chain $... \to x_t \to \mu_t \to x_{t+1} \to ...$ over the space of (x, μ) pairs. Our goal is then to characterize the observed stationary distribution of this process $p(x)$ as a function of the prior probabilities. Two different analyses of the inference process exist in literature, leading to subtle differences in the interpretation of $p(x)$. We shall describe both of them for completeness and to avoid future confusion. First, the model of Xu and Griffiths (2010) posits that, when presented with a noisy stimulus x_t, participants attempt to infer the true state of the world μ_t. This can be modeled as a Bayesian inference of the form $p(\mu_t|x_t) \propto p(x_t|\mu_t)\pi(\mu_t)$ where $\pi(\mu_t)$ is the prior over the possible states of the world μ_t and $p(x_t|\mu_t)$ is the likelihood of observing x_t if the true state of the world is μ_t. If predictions are then generated by simply sampling from the posterior (see Griffiths & Kalish, 2007, for discussion of other possibilities), the resulting stationary distribution over stimuli x is the posterior predictive distribution $p(x) = \int p(x|\mu)\pi(\mu)d\mu$ (Xu & Griffiths, 2010). Second, Jacoby and McDermott (2017) and Langlois et al. (2021) interpreted serial reproduction as follows: given a *true* incoming stimulus x_t, a noisy percept μ_t is generated through the likelihood $p(\mu_t|x_t)$. At the reconstruction stage (and assuming no production noise for simplicity), participants attempt to infer the true underlying stimulus by incorporating prior information regarding the distribution of stimuli in the world $\pi(x)$. This can be modeled using the posterior $p(x_{t+1}|\mu_t) \propto p(\mu_t|x_{t+1})\pi(x_{t+1})$. Assuming as before that participants generate inferences by sampling from the posterior, it is possible to show that the stationary distribution over stimuli converges to the prior itself $p(x) = \pi(x)$ (Jacoby & McDermott, 2017; Langlois et al., 2021).

Serial reproduction has been used to study priors in a variety of domains. Xu and Griffiths (2010) demonstrated the practical soundness of the paradigm by applying it to simple one-dimensional domains. For example, in one task participants were trained to distinguish between two types of fish, namely, fish-farm fish and ocean fish. Fish stimuli were generated schematically, and varied only in terms of their width. The width of fish-farm fish was normally distributed with certain mean and variance whereas that of ocean fish was uniformly distributed. By training participants on different farm-fish distributions, and then running a serial reproduction task in which participants saw a fish and had to reproduce it knowing that it came from a farm (initial fish were not necessarily from the fish-farm distribution), the authors showed that the process gradually recovers the training distributions.

In a more complex application of serial reproduction, Langlois, Jacoby and colleagues (Langlois et al., 2021) revealed shared priors in spatial memory by iterating a task in which participants reproduced precise point locations within images (Figure 10.6). In the task, participants viewed a red point positioned at random on top of an image, such as a gray circle or triangle shape. Following a delay, the image reappeared on the screen without the red point, and participants were instructed to indicate the exact location of the red point shown during the stimulus phase (see Figure 10.6A for illustration of the task and serial reproduction procedure). Past work (Huttenlocher, Hedges, & Duncan, 1991; Wedell, Fitting, & Allen, 2007) highlighted consistent biases in spatial memory, such as a clear tendency for point reconstructions within an image of a triangle to be biased towards the triangle vertices. Serial reproduction revealed details that had eluded past experimental approaches (Figure 10.6C). In particular, it revealed that spatial memory for point locations over a circle image are biased towards quadrant edges, and not the quadrant centers in a departure from previous work (Huttenlocher et al., 1991; Wedell et al., 2007) (Figure 10.6A and C).

Priors can originate from short-term interactions with the stimuli, possibly within the duration of the experiment (Xu & Griffiths, 2010; Jazayeri & Shadlen, 2010) but can also correspond to life-long culturally-dependent learning, for example in the case of language and music. For example, Jacoby and

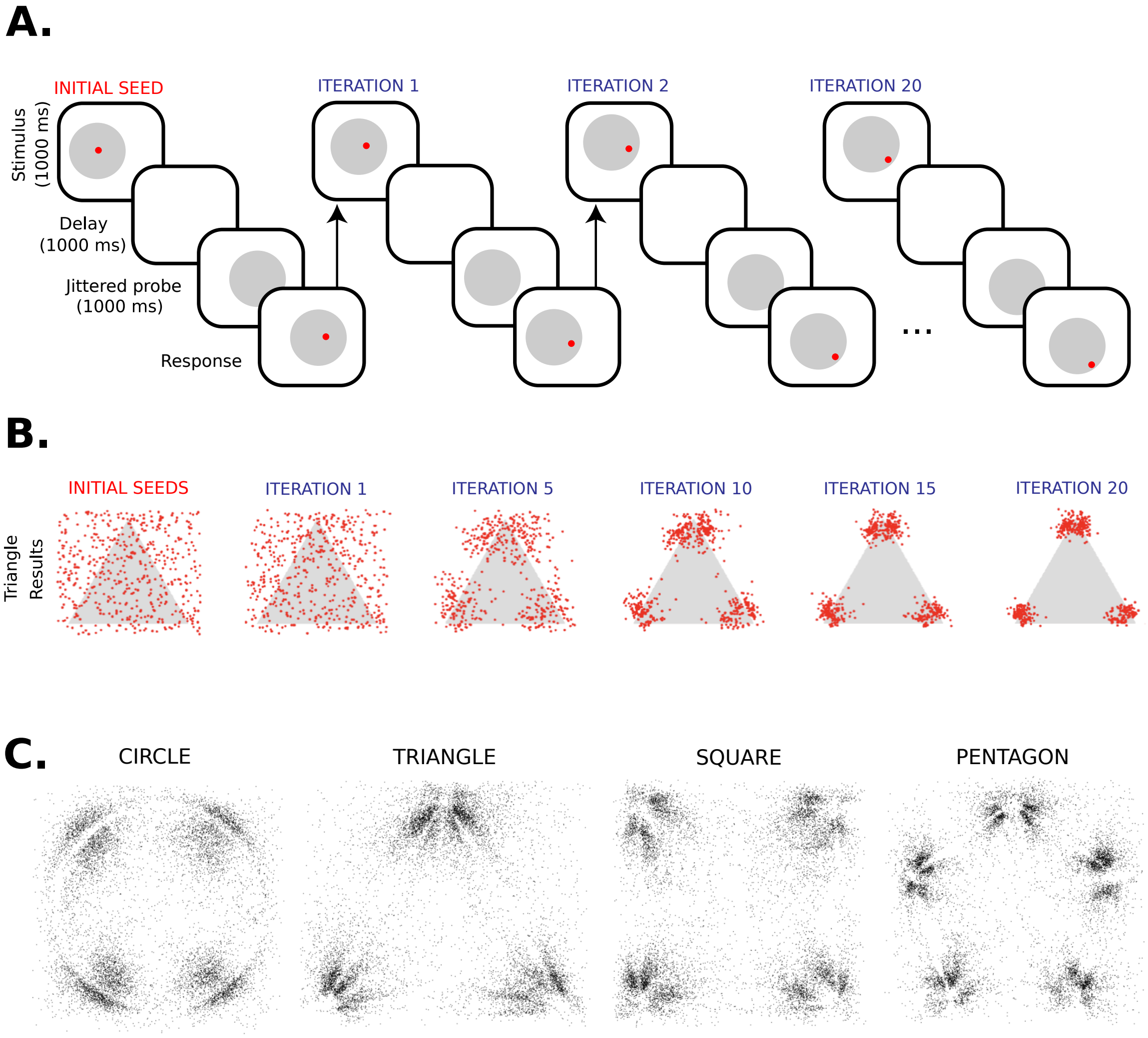


Figure 10.6: Estimating priors for reconstruction from visuospatial working memory. A. Illustration of the serial reproduction method for an image of a circle. The first participant views an image (a circle) with a point overlaid in a random position and is then asked to reproduce its location from memory. The next participant views the same image, but with the point located at the position reconstructed by the previous participant. The process is repeated for a total of 20 iterations. B. Serial reproduction results for the remembered position of point locations overlaid on an image of a triangle. The initial uniform distributions of 500 points are shown (far left) becomes increasingly structured with more iterations of serial reproduction. C. Experimental results showing superposition of responses across all iterations of the chains for images of simple geometric shapes: a circle, triangle, square, and pentagon, highlighting which regions of these shapes are estimated to have high prior probability. Figure adapted from Langlois et al. (2021).

McDermott (2017) used serial reproduction to reveal culturally-dependent priors in rhythm perception. In their task, participants were presented with a simple random rhythm and were asked to reproduce it using finger tapping. Western participants were shown to have very different rhythm representations compared with participants recruited from the Bolivian Amazon (see Figure 10.7). In an extension of

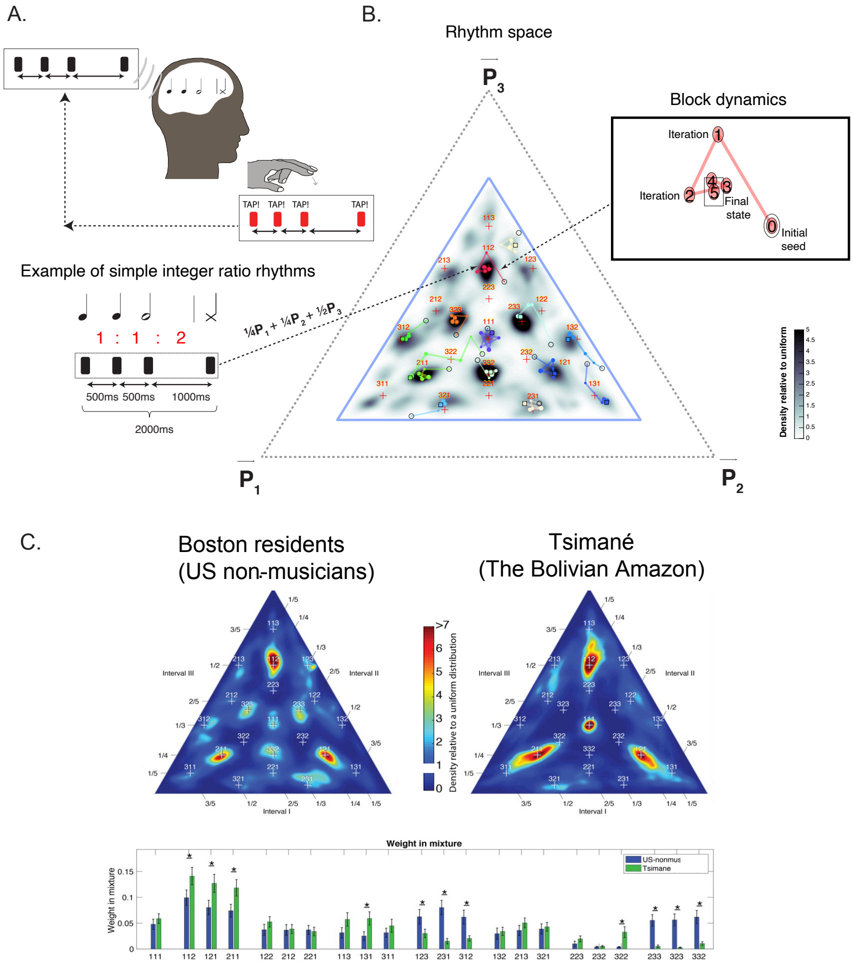


Figure 10.7: Serial reproduction experiment with musical rhythm from Jacoby et al. (2017). A) Schematic of experiment. Participants are presented with initially random seed rhythms (a repeating cycle of three clicks, defined by three inter-beat time intervals) and reproduce them by tapping. The reproduction becomes the stimulus on the following trial. This procedure is iterated 5 times. B) The rhythm space in which results are plotted. Each axis of the triangle represents one of the three intervals in a rhythm. Integer ratio rhythms, in which the time intervals are related by integer ratios, occupy a subset of points in the rhythm space. One example simple integer ratio rhythms (1:1:2) is displayed on the left. The colored dots connected by lines show trajectories from example experimental trials. Inset shows one example trial in more detail, converging in this case to the 1:1:2 rhythm. C. Experimental results showing large cross-cultural differences between Tsimane' participants from the Bolivian Amazon and US non-musicians from Boston. Despite the marked difference, there is overlap between the modes of the empirical results and integer-ratios. Results shows kernel density estimates of the responses in the last iterations. The panel below shows the relative importance ("weight") of different integer-ratio categories within a Gaussian mixture model fitted to the data. The results show categories with significant differences (such as 1:1:2/1:2:1/2:1:1 and 2:3:3/3:2:3/3:3:2) possibly reflecting different life-long exposure to music. Figure adapted from Jacoby et al. (2017).

that project, Jacoby et al. (2021) studied participants from 39 group from 15 countries. They found that priors depend on the nature of musical practices in each culture but also share universal features such as the existence of discrete rhythm categories at small integer ratios. Viewed together, these studies highlight the prospect of serial reproduction as a modern tool for studying perceptual priors in a wide range of contexts, and for creating meaningful comparisons of these priors across groups.

10.4 Markov chain Monte Carlo with People

Iterated learning and serial reproduction are effective methods for studying subjective probability distributions of specific kinds: iterated learning can reveal the prior distributions that inform learning, and serial reproduction can reveal the prior distributions that inform perception and memory. However, the Bayesian models presented in this book assume subjective probability distributions of many kinds that do not fall into these two classes. For example, models of categorization assume that categories are associated with probability distributions over stimuli. How could those distributions be estimated?

One way to engage with the broader problem of estimating subjective probability distributions is to take the key insight behind iterated learning and serial reproduction – that Markov chain Monte Carlo algorithms (such as Gibbs sampling) can be implemented with people – and generalize it. Fortunately, there are other types of Markov chain Monte Carlo algorithms that we can use with people, including the most famous such algorithm: the Metropolis-Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings, 1970).

Sanborn and colleagues (Sanborn & Griffiths, 2008; Sanborn, Griffiths, & Shiffrin, 2010) explored the potential of Metropolis-Hastings as a scheme for sampling from subjective probability distributions. This algorithm does not require participants to generate new examples as done in iterated learning. Instead they are given a choice between two items, where those items are selected in such a way that they implement a Markov chain Monte Carlo algorithm. In the Metropolis-Hastings algorithm, a Markov chain that converges to a particular stationary distribution is constructed by using a proposal distribution to propose a variation on the current state and an acceptance rule that depends on the target distribution to decide whether to accept that variation. Sanborn and colleagues realized that this kind of structure could naturally be translated into an experimental paradigm.

The experimental paradigm that Sanborn et al. developed makes it possible to sample from a probability distribution $p(x)$ proportional to any non-negative subjective quantity $f(x)$, such as the probability or utility assigned to an outcome x. The key is to design a task where people choose between two alternatives x^* and x such that the probability that they choose x^* is

$$P(\text{choose } x^*|x, x^*) = \frac{f(x^*)}{f(x) + f(x^*)}. \tag{10.1}$$

This is potentially straightforward to do, as Equation 10.1 is simply the Luce choice rule (Luce, 1959), which is widely used to model people's choices. If a task of this kind can be identified, a Markov chain can be constructed in exactly the same way as in the Metropolis algorithm, presenting the current value of the chain and a proposed alternative to people, asking them to choose between these options, and taking the result of the choice as the new current value. Equation 10.1 thus becomes the acceptance probability in this algorithm, resulting in a Markov chain that has $p(x) \propto f(x)$ as its stationary distribution. While Equation 10.1 is not identical to the acceptance rule used in the Metropolis-Hastings algorithm, it corresponds to another valid acceptance rule known as the Barker rule (Barker, 1965; Neal, 1993) and it is easy to check that it satisfies the detailed balance condition discussed in Chapter 6.

One example application of **Markov chain Monte Carlo with People (MCMCP)** is estimating the structure of natural categories. If a category c containing objects x is represented by a probability

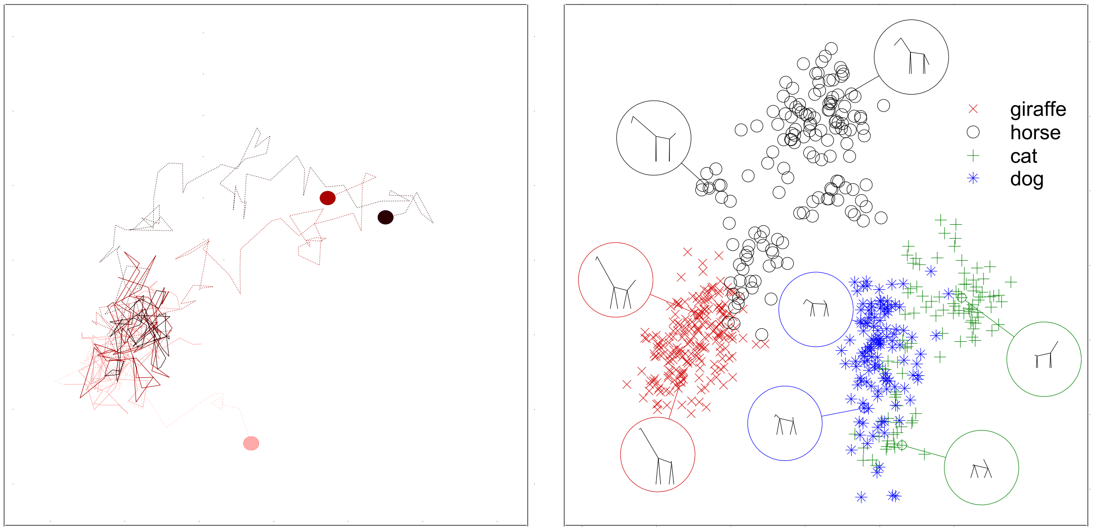


Figure 10.8: Markov chains and samples for one participant in an experiment exploring categories of animals. Both panels are a two-dimensional projection of a nine-dimensional space, where the original dimensions corresponded to the lengths and angles of lines in a stick-figure quadruped. The left panel shows three Markov chains for the "giraffe" category, started at disparate points in the space but converging on a fixed region relatively quickly. The right panel shows samples for all four categories, taken from the corresponding Markov chains after an appropriate burn-in. The bubbles show specific examples of stick-figures from these categories, illustrating that they seem to do a good job of capturing both the content and the variation associated with each category. Figure reproduced with permission from Sanborn and Griffiths (2008).

distribution $p(x|c)$, then we can construct a task that satisfies Equation 10.1 with $f(x) = p(x|c)$ by presenting people with two objects and asking them to indicate which is most likely to belong to the category (for details, see Sanborn et al., 2010). This provides a way to explore the structure of categories that people have learned through experience ("natural categories"). Sanborn and colleagues conducted an experiment where participants were shown stick-figure animals that varied along nine dimensions, such as the angle of their head and the length of their neck, and asked to make judgments about the membership of particular stick figures in four categories – cats, dogs, horses, and giraffes. The proposal distribution was a Gaussian distribution in this nine-dimensional space, with a small probability of a big jump to another more distant point. After a few hundred choices, the Markov chains produced by people's responses tended to converge on specific regions of the space, picking out different distributions for the different categories. The results from one participant are shown in Figure 10.8. Other examples include investigating mental representations in intuitive physics (Cohen & Ross, 2009), expressions in cartoon faces (McDuff, 2010), and even what people mean when they say they have had "a good night's sleep" (Ramlee, Sanborn, & Tang, 2017).

This method can be scaled up to spaces with many more dimensions than the nine used in the stick figure experiment. Martin, Griffiths, and Sanborn (2012) applied it to investigating facial expressions in a 175-dimensional space defined by the eigenvectors of images of faces ("eigenfaces"). Participants in this experiment saw pairs of faces and were asked to judge which face of the pair was happier or sadder. The Markov chains in this experiment were linked over participants: the face last chosen by one participant was the starting point for the next participant. The result of this procedure were realistic-looking facial expressions of the type that would be difficult to describe in words, as shown in the eigenface result in Figure 10.9e.

Of course even the Metropolis-Hastings algorithm will have trouble sampling from complex probability distributions in high dimensions. While the sampler will converge to the correct distribution eventually, it may take too many trials for the approach to be feasible for use with human participants. Fortunately, Metropolis-Hastings is an extremely flexible algorithm and researchers have developed many clever ways in which to increase its efficiency. One way is to introduce the idea of "momentum" to the sampler, so if the Markov chain is travelling along a ridge of high probability then it will tend to stay on that ridge and not waste time probing the low probability valleys. Interleaving trials in which participants can select the future direction the sampler will travel in can increase its efficiency (Blundell, Sanborn, & Griffiths, 2012).

A problem with trying to construct a feature space for images is that it is very difficult and often, as in the case of eigenfaces, points in a feature space do not correspond with any sensible image. An alternative to constructing a feature space is to use only real images. Hsu, Martin, Sanborn, and Griffiths (2019) extended this approach to let it be used in a setting where the goal is to estimate a distribution over a discrete set of stimuli. In this case, the proposal distribution is constructed by using a similarity measure on the stimuli to define a b-matching, which is a graph in which every node is connected to b other nodes. Each node in the graph corresponds to one stimulus, and the connections are made so that stimuli are linked to other stimuli to which they are similar. The proposal distribution for the Markov chain Monte Carlo algorithm is then a random walk on this graph – picking one of the b edges at each node uniformly at random. Since each node has the same number of edges, this distribution is symmetric.

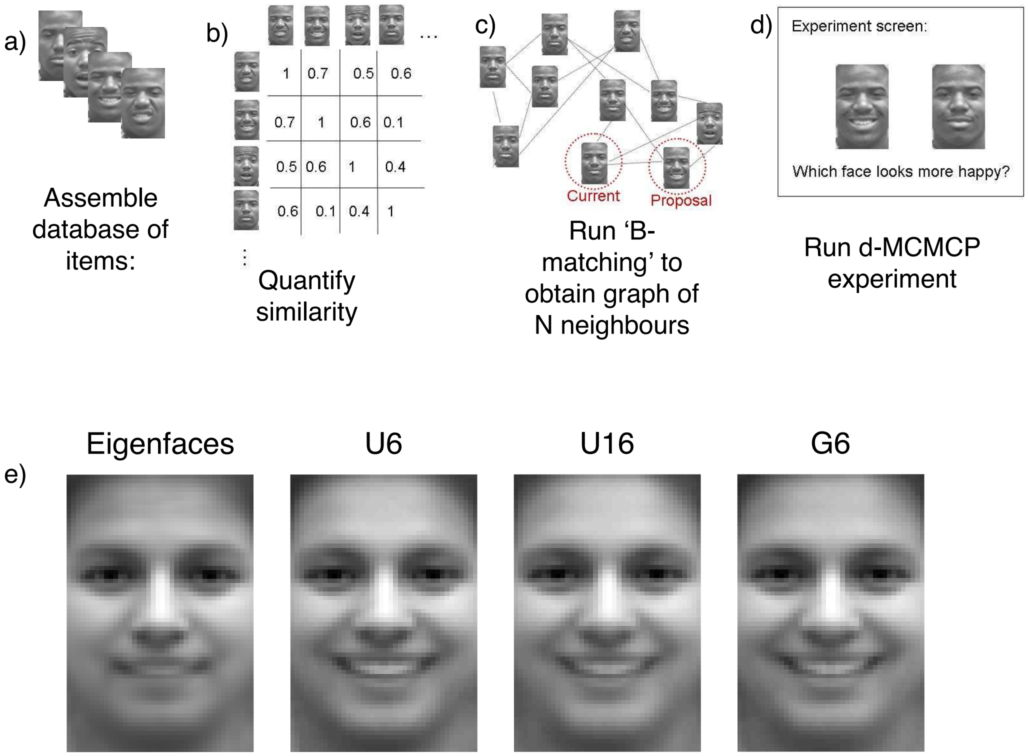
Figure 10.9 shows how this works for a set of stimuli corresponding to faces displaying different emotions. A computer vision algorithm is used to construct a similarity matrix, and the b-matching algorithm constructs the graph. People see pairs of faces that are neighbors in this graph, and are asked to choose which face of the pair is a better match for a category – in this case, happy faces. Hsu et al. found that this approach outperformed an MCMC algorithm in which the faces are first transformed into a continuous space and a Gaussian proposal distribution is applied in that space.

Hsu et al. found that this approach outperformed the eigenface feature space used in Martin et al. (2012), and this approach has also been applied to better understand how surgeons mentally represent fractures of the humerus (Jabbar et al., 2013).

10.5 Gibbs Sampling with People

Despite the flexibility of MCMCP, there are certain aspects of the paradigm that can make it hard to apply in certain domains. Specifically, the two-alternative forced choice interface of MCMCP provides a single bit of information per trial. This, in turn, can be quite time consuming in highly multimodal domains that require a fair bit of exploration. Similarly, the performance of MCMC algorithms critically depends on the choice of a proposal distribution; a distribution that is too narrow may not converge in practice, and a distribution that is too broad may miss important details in the subjective distribution. When computer simulations are involved, this may not be such a big problem as it is often relatively cheap to experiment with a variety of proposal widths and pick the best parameter (e.g., through cross validation). However, this is not the case when the sampler involves human recruitment in the loop which can be quite expensive.

To remedy this, Harrison, Marjieh and colleagues (Harrison et al., 2020) proposed an alternative paradigm called **Gibbs Sampling with People (GSP)** that instantiates another variation on the Gibbs sampling process discussed earlier. Unlike serial reproduction and iterated learning where the conditional sampling alternates between hypotheses and stimuli, here the process alternates directly over the stimulus



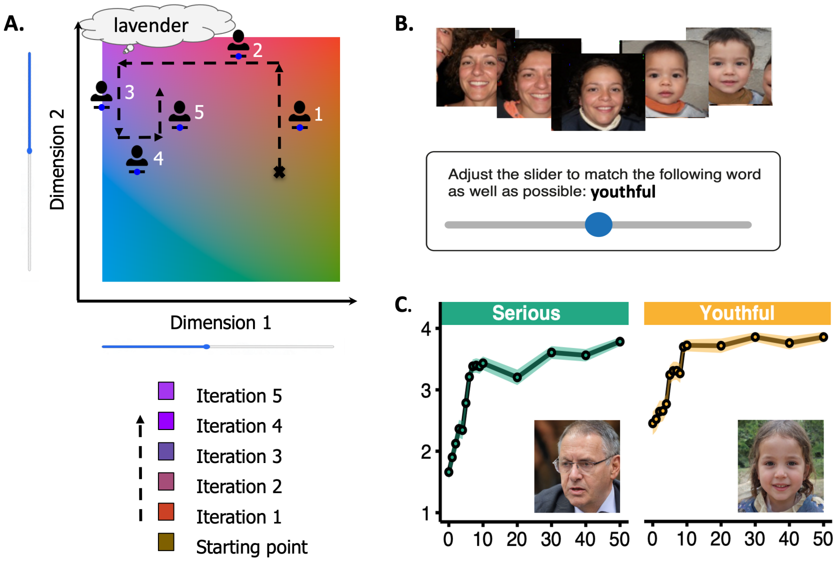
1	0.7	0.5	0.6
0.7	1	0.6	0.1
0.5	0.6	1	0.4
0.6	0.1	0.4	1

Experiment screen:

Which face looks more happy?

Figure 10.9: Using Markov chain Monte Carlo with people for a discrete set of stimuli. (a) Stimulus images. (b) A computer vision algorithm is used to build a similarity matrix. (c) The b-matching algorithm constructs a graph based on the similarity matrix, ensuring each node has the same number of neighbors. (d) The MCMC algorithm consists of people making decisions about which of a pair of neighboring images in the graph best captures the category. (e) This approach results in better estimates of the average happy face than running a MCMC algorithm in a continuous space where faces are represented by the eigenvectors of the set of images in the database. U6 is a uniform random walk on a graph with 6 neighbors per stimulus, U16 a uniform random walk with 16 neighbors, and G6 a geometric proposal distribution in which the number of steps is first chosen from a geometric distribution and those steps are then taken via a uniform random walk on a graph with 6 neighbors. Figure adapted from Hsu et al. (2019).

dimensions. Specifically, given a d-dimensional stimulus space (e.g., colors) and a parameterization $(x_1, ..., x_d)$ (e.g., RGB channels), in a GSP trial we fix $d-1$ parameters and have a participant explore and select the value of one free dimension (e.g., the R channel) using a slider (see Figure 10.10A), so that the generated stimulus best matches a certain target category (e.g., strawberry). The resulting stimulus then gets passed to a new participant where they get to explore a different dimension and so on. Each such iteration constitutes a sample from the conditional probability $p(x_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_d)$ and by circulating over the dimensions this process instantiates a Gibbs sampler from the subjective distribution $p(x_1, ..., x_d)$. A related paradigm with a similar interpretation was applied to the study of subjective randomness of discrete coin-flip sequences in Griffiths, Daniels, Austerweil, and Tenenbaum (2018).



Adjust the slider to match the following word as well as possible: **youthful**

Figure 10.10: Gibbs Sampling with People (GSP). A. Schematic procedure. At each iteration participants control one stimulus dimension by moving a slider to optimize for a certain category (e.g., lavender). The result of one generation of participants becomes the input for a new generation. B. GSP applied to faces. Participants interact with a slider and seek faces that best match a target categories. C. Example faces sampled from GSP chains for the categories "serious" and "youthful." Curves denote average sample-quality ratings elicited from a separate group of raters as a function of chain iterations. Samples quickly converge on high quality images with respect to the target categories. Figure adapted from Harrison et al. (2020).

Harrison et al. (2020) applied GSP for the study of subjective categories over a variety of domains ranging from simple colors and up to fully naturalistic faces. For example, by coupling GSP with the first 10 principal dimensions of the latent space of StyleGAN (Härkönen, Hertzmann, Lehtinen, & Paris, 2020), a modern neural architecture for naturalistic image synthesis, the authors showed that GSP can be effectively used for the study of human bias with respect to face categories (such as what constitutes a "serious" or a "youthful" face, Figure 10.10). GSP has been applied also for the study of emotional prototypes in prosody (Van Rijn et al., 2021), the study of the acoustic pleasantness of chords (Marjieh, Harrison, Lee, Deligiannaki, & Jacoby, 2022), and for the generation of structured task distributions in reinforcement learning (Kumar et al., 2022).

10.6 Future directions

Both Markov chain Monte Carlo with People and Gibbs Sampling with People illustrate how algorithms that were originally designed for computers can be reconstrued as new methods for studying human cognition. The critical step is designing a task for which it is reasonable to interpret human behavior as a sample from a particular probability distribution. Many other algorithms make use of random samples to solve particular problems, raising the possibility that there are other algorithms that can be equally useful as experimental methods.

For example, so far we discussed MCMCP chains as separate entities, however, modern samplers often use multiple simultaneous chains to adaptively optimize their proposal function (Goodman & Weare, 2010). It is conceivable, therefore, to come up with new paradigms in which information is shared across multiple sampler chains to achieve better convergence behavior. Likewise, Markov chain Monte Carlo and deterministic optimization can be viewed as two limits of a continuum of algorithms where one manipulates the stochasticity (or **temperature**) of the process: MCMC is the "high temperature" case where the goal is to bounce around the distribution and it is okay to visit regions with lower probability, while optimization is the "low temperature" case in which changes to the state should increase its probability. As shown in Harrison et al. (2020), one way to control this level of stochasticity in MCMC algorithms with people is to aggregate over multiple human judgments per iteration, driving the sampler toward more deterministic behavior. This can be particularly useful if we're interested in characterizing the modes of distributions rather than their general shape, or optimizing over subjective losses defined by a population of participants. There are probably as many optimizers out there as samplers, and bringing them into the toolbox of the modern psychologist can be of great value.

More generally, the methods we use in cognitive science are beginning to undergo an important change. Studying the mind became a science in the twentieth century, and for most of that century used a specific methodology: people would come into a lab, do a task for around an hour, perhaps being assigned to one of a few different conditions. This methodology was in part a consequence of the constraints of running a physical laboratory: people had to travel to get there, so it made sense to have them stay for a while, and experiments would be administered by research assistants and consequently couldn't manipulate too many variables.

Twenty-first century cognitive science is in a very different situation. Experiments are increasingly run using online crowdsourcing services. Often, these experiments are just scaled-up versions of lab experiments. However, crowdsourcing offers a completely different profile from a physical lab: people can be paid to make as little as a single decision, and the tasks that people are presented with are selected by a computer that has access to all previous decisions. This sets up an environment where there is far greater freedom to explore innovative experimental designs, and where experiments look much more like algorithms that are run with people. Thinking intelligently about how to design those algorithms, drawing on ideas from computer science and statistics, has the potential to shed a lot more light on the nature of human cognition (for further discussion of this point, see Suchow & Griffiths, 2016).

10.7 Conclusion

Bayesian models of cognition assume probability distributions over complex, high-dimensional objects that would be difficult to estimate using traditional elicitation methods. However, a variety of approaches exist for estimating these distributions, drawing on algorithms for estimating probability distributions that are used in computer science and statistics. Iterated learning can be used to estimate prior distributions that inform learning. Serial reproduction can be used to infer prior distributions that result in biases in perception and memory. Markov Chain Monte Carlo with People and Gibbs Sampling with People offer more general algorithms that can be used to reveal the structure of psychological representations in a remarkable variety of settings. Just as computer scientists and statisticians constantly innovate on the methods they used to estimate probability distributions, we see these methods as providing a foundation on which further tools for gaining insight into human cognition can be built.

References

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

Barker, A. A. (1965). Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Australian Journal of Physics*, *18*, 119-133.

Bartlett, F. C. (1932). *Remembering: a study in experimental and social psychology.* Cambridge University Press.

Blundell, C., Sanborn, A. N., & Griffiths, T. L. (2012). Look-ahead Monte Carlo with people. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society* (p. 1356-1361).

Canini, K. R., Griffiths, T. L., Vanpaemel, W., & Kalish, M. L. (2014). Revealing human inductive biases for category learning by simulating cultural transmission. *Psychonomic Bulletin & Review*, *21*(3), 785–793.

Cohen, A., & Ross, M. (2009). Exploring mass perception with Markov chain Monte Carlo. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 1833-1844.

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, *100*(470), 680–701.

Goodman, J., & Weare, J. (2010). Ensemble samplers with affine invariance. *Communications in Applied Mathematics and Computational Science*, *5*(1), 65–80.

Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science*, *32*, 68-107.

Griffiths, T. L., Daniels, D., Austerweil, J. L., & Tenenbaum, J. B. (2018). Subjective randomness as statistical inference. *Cognitive Psychology*, *103*, 85–109.

Griffiths, T. L., & Kalish, M. L. (2005). A Bayesian view of language evolution by iterated learning. In *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society* (p. 827-832).

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*(3), 441–480.

Härkönen, E., Hertzmann, A., Lehtinen, J., & Paris, S. (2020). GANspace: Discovering interpretable GAN controls. *Advances in Neural Information Processing Systems 33*, 9841–9850.

Harrison, P., Marjieh, R., Adolfi, F., Rijn, P. van, Anglada-Tort, M., Tchernichovski, O., Larrouy-Maestri, P., & Jacoby, N. (2020). Gibbs sampling with people. *Advances in Neural Information Processing Systems*, *33*, 10659–10671.

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

Hsu, A. S., Martin, J. B., Sanborn, A. N., & Griffiths, T. L. (2019). Identifying category representations for complex stimuli using discrete Markov chain Monte Carlo with people. *Behavior research methods*, *51*(4), 1706–1716.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: prototype effects in estimating spatial location. *Psychological Review*, *98*(3), 352.

Jabbar, Y., Majed, A., Hsu, A., Fairhurst, P., Vlaev, I., Reilly, P., & Emery, R. J. (2013). Decision-making in proximal humeral fractures. *Shoulder & Elbow*, *5*(2), 78–83.

Jacoby, N., & McDermott, J. H. (2017). Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. *Current Biology*, *27*(3), 359–370.

Jacoby, N., Polak, R., Grahn, J., Cameron, D., Lee, K. M., Godoy, R., Undurraga, E. A., Huanca, T., Thalwitzer, T., Doumbia, N., et al.. (2021). Universality and cross-cultural variation in mental representations of music revealed by global comparison of rhythm priors.

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience*, *13*(8), 1020–1026.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin and Review*, *14*, 288-294.

Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2009). *The world color survey.* CSLI Publications.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation*, *5*, 102-110.

Kumar, S., Dasgupta, I., Marjieh, R., Daw, N. D., Cohen, J. D., & Griffiths, T. L. (2022). Disentangling abstraction from statistical pattern matching in human and machine learning. *arXiv preprint arXiv:2204.01437*.

Langlois, T. A., Jacoby, N., Suchow, J. W., & Griffiths, T. L. (2021). Serial reproduction reveals the geometry of visuospatial representations. *Proceedings of the National Academy of Sciences*, *118*(13), e2012938118.

Lewandowsky, S., Griffiths, T. L., & Kalish, M. L. (2009). The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive Science*, *33*(6), 969–998.

Luce, R. D. (1959). *Individual choice behavior.* Wiley.

Marjieh, R., Harrison, P. M., Lee, H., Deligiannaki, F., & Jacoby, N. (2022). Reshaping musical consonance with timbral manipulations and massive online experiments. *bioRxiv*.

Martin, J. B., Griffiths, T. L., & Sanborn, A. N. (2012). Testing the efficiency of Markov chain Monte Carlo with people using facial affect categories. *Cognitive Science*, *36*(1), 150–162.

McDuff, D. (2010). A human-Markov chain Monte Carlo method for investigating facial expression categorization. In *Proceedings of the 10th International Conference on Cognitive Modeling* (pp. 151–156).

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.

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

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., & Rakow, T. (2006). *Uncertain judgements: eliciting experts' probabilities.* Wiley.

Ramlee, F., Sanborn, A., & Tang, N. (2017). What sways people?s judgement of sleep quality? a quantitative choice-making study with good and poor sleepers. *Sleep*, *40*(7), zsx091.

Sanborn, A. N., & Griffiths, T. L. (2008). Markov chain Monte Carlo with people. In *Advances in Neural Information Processing Systems 20.*

Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. M. (2010). Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology*, *60*(2), 63–106.

Suchow, J. W., & Griffiths, T. L. (2016). Rethinking experiment design as algorithm design. In *CrowdML: Workshop on Crowdsourcing and Machine Learning.*

Van Rijn, P., Mertes, S., Schiller, D., Harrison, P., Larrouy-Maestri, P., André, E., & Jacoby, N. (2021). Exploring emotional prototypes in a high dimensional TTS latent space. *arXiv preprint arXiv:2105.01891.*

Wedell, D. H., Fitting, S., & Allen, G. L. (2007). Shape effects on memory for location. *Psychonomic Bulletin & Review*, *14*(4), 681–686.

Xu, J., Dowman, M., & Griffiths, T. L. (2013). Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1758), 20123073.

Xu, J., & Griffiths, T. L. (2010). A rational analysis of the effects of memory biases on serial reproduction. *Cognitive Psychology*, *60*(2), 107–126.

Yeung, S., & Griffiths, T. L. (2015). Identifying expectations about the strength of causal relationships. *Cognitive Psychology*, *76*, 1–29.