Sampling as the human approximation to probabilistic inference

Adam N. Sanborn[1], Jian-Qiao Zhu[1], Jake Spicer[1], Joakim Sundh[1], Pablo León-Villagrá[1], and Nick Chater[2]
[1]University of Warwick
[2]Warwick Business School

We live in an uncertain world, which makes it difficult to know what we should believe. In the absence of certainty, the Bayesian approach provides a formal framework that results in assigning each possible state of the world a probability, and using the laws of probability to calculate what to believe and what to do. This theoretical framework was developed in the 1940s and 1950s to provide prescriptions for human behaviour, and has been particularly influential in developing theories of how people behave in economic situations (Edwards, 1961; Peterson & Beach, 1967; Savage, 1954; von Neumann & Morgenstern, 1947).

Advances in methodology and computational power in the past two decades has seen researchers begin to compare human performance to Bayesian models in complex domains, such as vision, motor control, language, categorisation or common-sense reasoning. In these domains, people's performance has been found to be similar to that of highly complex probabilistic models, if these models assume the same sensory limitations that people have (Anderson, 1991; Chater & Manning, 2006; Goodman et al., 2008; Griffiths et al., 2007; Griffiths & Tenenbaum, 2011; Houlsby et al., 2013; Oaksford & Chater, 2007; Pantelis et al., 2014; Petzschner et al., 2015; Sanborn, Griffiths, & Navarro, 2010; Wolpert, 2007; Yuille & Kersten, 2006). Some of the most compelling demonstrations have been in the domain of intuitive physics, in which participants are asked to make judgments about various physical quantities, such as blocks in motion or liquids. Despite the complexities of these predictions, complex probabilistic models often explain people's judgments better than competing frameworks like heuristics (Battaglia et al., 2013; Sanborn et al., 2013).

These results are surprising as they run counter to an extensive literature showing that people make systematic errors when reasoning about probabilities (Gigerenzer & Gaissmaier, 2011; Kahneman, 2011; Tversky & Kahneman, 1974). First, there are many demonstrations of how asking a question in different ways will alter the probability judgment that a person makes. For example, *unpacking effects* show that people's estimate of the

---

Manuscript is the author accepted copy of a chapter in *Human-Like Machine Intelligence* published by Oxford University Press. This manuscript is not the copy of record and may not exactly replicate the authoritative document. Please do not copy or cite without author's permission. Sanborn, Zhu, Spicer and Chater were partially supported by a grant from the ESRC Rebuilding Macroeconomics program. Sanborn, Zhu, Spicer, Sundh and León-Villagrá were supported by a European Research Council consolidator grant [817492-SAMPLING]. Chater was partially supported by the ESRC Network for Integrated Behavioural Science [ES/P008976/1].

probability that someone "can buy a gun in a hardware store" is more than the probability that they "can buy an antique gun or some other type of gun in a hardware store", but less than the probability that they "can buy a staple gun or some other type of gun in a hardware store", despite all three questions asking about the same set of events and therefore having the same probabilities (Dasgupta et al., 2017; Sloman et al., 2004). Perhaps more damning, however, is the observation that people's probability estimates are not consistent with one another – that combinations of estimates do not follow the rules of probability theory as they should (e.g., Costello and Watts, 2014). One salient demonstration of this is the conjunction fallacy made famous by Tversky and Kahneman (1983): people will more often than not judge the probability that a highly-educated, liberal-seeming person is a bank teller to be lower than the probability that this person is both a feminist and a bank teller, despite the fact that the group of bank tellers includes all feminists who are also bank tellers. As probability theory is at the heart of complex probabilistic models, it appears a paradox that people's judgments in complex tasks match those of probabilistic models, yet their probability judgments disagree with probability theory.

How can we explain this apparent paradox? First, we note that the idealised way of implementing complex probabilistic models, representing all possible probabilities and making exact calculations with these probabilities, is implausible for any physical system, including brains (Aragones et al., 2005; Sanborn & Chater, 2016). An example of why this is the case is to consider the problem of categorising objects in the world into different natural kinds, and then making a decision in the light of that categorisation. A common Bayesian approach to this problem is to represent all possible ways of dividing the observed objects into different categories, and then summing over all these possible partitions to make a decision (Anderson, 1991; Sanborn, Griffiths, & Navarro, 2010). This calculation becomes intractable long before we reach a number that could realistically correspond to a lifetime of experience: even for just 100 objects there are over 4.7×10^{115} ways to divide them into categories, which is far greater than the number of atoms in the observable universe. Therefore, explicitly using probabilities in categorisation or other complex domains where Bayesian models have been successful such as vision, intuitive physics, and language, is thus clearly impossible.

But how can a Bayesian model of categorisation, vision, intuitive physics, or language possibly work without explicitly representing probabilities? A key insight is that it is not necessary to explicitly represent probabilities in order to implement complex probabilistic models. Instead, these models can be approximated, and a straightforward way in which to do so is to draw samples from the probability distribution rather than representing it explicitly. Using sampling as an approximation to complex probabilistic models has a long history beginning in the 1940s and '50s (Metropolis et al., 1953); and as the computational resources available to researchers have increased, it has become a common way in which to approximate these models in both cognitive science and artificial intelligence (Griffiths et al., 2007; Susskind et al., 2008). The major attraction of sampling is that it comes with a theoretical guarantee: an infinite number of samples will provide the same answer as exactly calculating with explicit probabilities. Additionally, using a finite and achievable number of samples provides useful approximations, though this can lead to erroneous answers in some situations. Cognitive science researchers have recently been intrigued by the possibility that the mind implements a sampling algorithm, both for their positives and negatives: the pos-

itives could explain how people with finite brains could approximate complex probabilistic models, while the negatives could explain the systematic errors people show when reasoning about probabilities.

One particularly revealing bias is *probability matching*. For example, on a multiple-choice test if a student believes that Option A has a 90% chance of being the right answer, instead of always choosing Option A he or she will still choose an alternative 10% of the time (Mosteller & Nogee, 1951; Vulkan, 2000). This behaviour contradicts a key supposition of rationality — that people always choose the option they consider best — and instead shows that human decision making is stochastic. This puzzling bias can though be explained, at least as a first approximation, by sampling: drawing a single independent sample from the probability distribution of which option is correct will result in Option A being sampled 90% of the time (Vul et al., 2014). A host of other reasoning fallacies have also begun to be explained by sampling, including the unpacking effect described above (Dasgupta et al., 2017; Lieder et al., 2018a; Sanborn & Chater, 2016). However, research in this area is only beginning, and the current state of the art is that different algorithms are used to explain different effects, as the tasks investigated thus far make it very difficult to distinguish between sampling algorithms. Ignorance of the sampling process makes it then difficult to arrive at a coherent explanation of how sampling can produce biases, and also prevents precise quantitative predictions from being made.

A sense of location in the human sampling algorithm

The best-known and often most efficient method for drawing samples is to draw them independently from the probability distribution of interest – we term this *direct sampling*. In statistics, there are a variety of methods for drawing samples independently. Computer algorithms have been developed to generate samples from simple distributions such as Gaussian or uniform distributions, and with more complex distributions there are other methods that can generate independent samples, such as rejection or importance sampling. However, to take advantage of these efficient sampling methods requires knowing a fair amount about the distribution of interest – either characterizing it exactly or, as is the case for rejection or importance sampling, knowing it well enough to be able to identify another distribution that is very similar (Bishop, 2006).

However, it does not seem likely the mind or the brain directly samples from probability distributions. To develop an intuition for why this is the case, consider the task of unscrambling a jumbled-up string of letters to make a word, knowing that each string can only be unscrambled so as to make a single word. In this example, the three strings of letters are “CIBRPAMOLET”, “NNLNRIEITOAAT”, and “AABRMSTENMESR”. We can think of this problem as implying a probability distribution where the different hypotheses are each of the possible orderings of a letter string. This means that there are 11 factorial or 39,916,800 orderings for “CIBRPAMOLET” and 13 factorial or 6,227,020,800 orderings each for “NNLNRIEITOAAT” and “AABRMSTENMESR”. The probability of each ordering is then simply the probability that that ordering of the letters is a word. As there is only one way that each string can be unscrambled to be a word, this probability distribution must be concentrated on the single correct ordering, with the small remainder divided amongst the huge number of non-word orderings.

If it was possible to directly and efficiently sample from the probability distribution over hypotheses, it would be easy to unscramble each of the letter strings – as samples are generated according to their probabilities, almost all of the generated samples would be of the correct ordering. However, as will be obvious, we cannot immediately generate the correct answers. This might be because samples are just generated very slowly in this task, so another observation is useful: changing the task to be to unscramble the mildly scrambled strings "PROBELMATIC", "INTERNATOINAL", and "EMABRRASSMENT" makes it a lot easier, even though these are the same sets of letters. Therefore, taken from the perspective of sampling, sampling the correct answer is much easier when starting from a mildly scrambled string; but this cannot arise through direct sampling, which is independent of the starting point.

While direct sampling does not seem tenable as a result of these observations, there are sampling algorithms for which new samples do depend on previous samples. One very well-known algorithm that has this property is Markov Chain Monte Carlo (MCMC; Metropolis et al., 1953). MCMC works by constructing a Markov chain that is characterized by a set of transition probabilities between potential states of the chain. During any one iteration, the chain is in a specific state, and a nearby state is blindly proposed as the next potential state. The ratio of the probability of the new state to the probability of the current state is then calculated, and this ratio is used to stochastically decide whether the chain stays put or transitions to the proposed state.

An MCMC chain generates a series of states through many iterations of this procedure, and under mild assumptions this series of states can be treated as samples from the probability distribution. Of course, the order in which samples arise from this algorithm is not independent, as in direct sampling. This is because the proposed state is selected from those states that are nearby the current state, and so the states of the Markov chain change more slowly than do the states in direct sampling. The greater chance MCMC has of transitioning to nearby hypotheses, that is, a "sense of location," helps explain our observations in the example above, where it is much easier to unscramble the letter strings "PROBELMATIC", "INTERNATOINAL", and "EMABRRASSMENT" than it is to unscramble the strings we initially presented.

MCMC's sense of location has led to this algorithm being used to explain a variety of cognitive biases, such as the anchoring effect. In anchoring experiments, participants are first asked to make a decision about whether a quantity is higher or lower than an irrelevant number. For example, participants are asked to add 400 to the last three digits of their phone number, to think of the resulting number as a date, and to decide whether Atilla the Hun was defeated before or after that date. Finally, participants were asked to provide the specific year in which Atilla the Hun was defeated. Despite the fact that the numbers generated from the participants' phone numbers were transparently irrelevant, participants' estimates were pulled toward these values (Russo & Schoemaker, 1989).

MCMC has been used to explain these results by assuming that the decision about whether a quantity is higher or lower than an irrelevant number sets the initial state of the Markov chain. Once this initial state is set, the algorithm samples from the probability distribution (e.g., of possible dates when Atilla the Hun was defeated), and the last sample generated is taken as the estimated date of defeat. If the number of iterations is great enough, then distribution of estimates will be unbiased. However, a limited number of

iterations will result in an estimate distribution that is biased by the algorithm's starting point, producing an anchoring effect. MCMC can also explain how various manipulations affect the strength of the anchoring effect, including whether the anchor is provided or self-generated, the level of participant expertise, cognitive load, and financial incentives (Lieder et al., 2012; Lieder et al., 2018a, 2018b).

The final example of explaining cognitive biases with MCMC we will discuss here is the unpacking effect. In experiments by Dasgupta et al. (2017), participants were told that their friend sees a table in a visual scene that they themselves cannot see, and in the first condition were asked to judge the probability that "any object starting with a C" is also in the scene. In the second condition, participants were asked to judge the probability that a "chair, computer, curtain, or any other object starting with a C" shares the scene with the table. Finally, in the third condition, participants were asked to judge the probability that a "cannon, cow, canoe, or any other object starting with a C" shares the scene with the table. These three questions are formally identical: the two unpacked versions of the questions just list kinds of objects that are implicit in the packed question 'any object starting with a C". Despite this, average estimates are highest when the question is unpacked as "chair, computer, curtain, or any other object starting with a C", intermediate for the simple question "any object starting with a C", and lowest for unpacking "cannon, cow, canoe, or any other object starting with a C" .

As with anchoring, MCMC has been used to explain this effect as the result of its starting point. First, it is assumed that object names are arranged in a semantic space and that asking participants which objects share a scene with a table induces a probability distribution over objects. Then the question that is asked helps position the starting point of the sampler: towards a region in which objects that begin with C are likely as in "chair, computer, curtain, or any other object starting with a C", or towards a region in which objects that begin with C are unlikely as in "cannon, cow, canoe, or any other object starting with a C". This starting point bias can thus explain how this unpacking effect depends on the probability of the unpacked hypotheses (Dasgupta et al., 2017).

The list above is illustrative, but certainly not complete. A variety of other biases have also been explained by MCMC, including the base-rate fallacy, conjunction fallacy, the weak evidence effect, the dud alternative effect, the self-generation effect, and wisdom of the crowd effects (Dasgupta et al., 2017; Sanborn & Chater, 2016). Even perceptual effects, such as switching times in bistable perception, have also been explained by MCMC (Gershman et al., 2012).

Key properties of cognitive times series

While MCMC has been used to explain a variety of judgment biases, it is certainly not the only sampling algorithm with a sense of location. MCMC is often slow to converge, particularly for multi-modal probability distributions, properties that have been exploited for explaining cognitive biases. These weaknesses have resulted in computer scientists and statisticians developing a variety of algorithms based on simple MCMC that mitigate these problems. Various proposals include methods that learn to adapt state proposals to the problem at hand, methods that involve running multiple chains, and methods that use the gradient of the probability distribution (Robert et al., 2018). This list only considers

elaborations of MCMC algorithms, and additionally there are alternative algorithms such as particle filtering that allow for changing posterior distributions (Doucet et al., 2001).

For the most part, cognitive scientists comparing sampling algorithms to human data have evaluated the qualitative properties of these algorithms, though a few researchers have quantitatively fit individual sampling algorithms to human data (Abbott & Griffiths, 2011; Lieder et al., 2018a). What has been lacking are quantitative comparisons between sampling algorithms to determine which algorithm best matches human behavior amongst the many possible candidates.

Part of the problem stems from the fact that the types of data that researchers have been using sampling algorithms to explain are not very diagnostic. First, the qualitative finding that the starting point has an influence (e.g., the algorithm has a sense of location) can be produced by a number of algorithms. Second, cognitive biases are generally biases in decision making, and decisions are commonly understood to be the result of the aggregation of a number of samples (Bogacz et al., 2006). As sampling algorithms all converge to the correct distribution in the limit, the sample aggregates produced by various sampling algorithms will often be similar, and this problem is compounded by the fact that sampling algorithms can closely mimic one another, given suitable choices of parameters (Lieder et al., 2018a).

Intuitively, there should be more power to discriminate between sampling algorithms if individual samples are observed, rather than only observing a decision based on an aggregation of samples. In particular, different sampling algorithms will generate different proposals, and will show different dependencies on previously generated hypotheses, often for a wide range of the settings of their parameters. Characterizing the properties of the time series of candidate sampling algorithms and comparing them against the properties of "cognitive time series" is thus a promising avenue for distinguishing between algorithms.

Fortunately, there exists a body of work by psychologists investigating such cognitive time series, which we can re-purpose to compare and contrast sampling algorithms. Classic work by Bousfield and Sedgewick (1944) asked participants to generate responses in a task in which the number of potential responses was large but limited, e.g., asking participants to generate the names of quadruped mammals. While the focus of this work was quantifying the rate at which quadruped animal names and other responses were produced, it was noted in passing that responses tended to be clustered. For example, participants would first produce a set of animals that could be found on a farm, and then produce a set of animals that could be found on a safari.

More recent work in this paradigm by Rhodes and Turvey (2007) looked more closely at the time intervals between successive recalls of animal names. While retrieval intervals lengthened as the pool of unreported animal names shrank, there were also bursts of short retrieval intervals interleaved with long waits, which were perhaps due to participants slowly searching for a new cluster of animal names to report and then quickly reporting the names in that cluster. Qualitatively, there were many more short retrieval times than long retrieval times. Quantitatively, the retrieval intervals were examined in the raw data and in data that were de-trended to remove the effect of slowing retrieval intervals. In both data sets, the retrieval intervals l were well characterized by Lévy probability density distributions

$$P(l) \sim l^{-u} \tag{1}$$

showing a power-law relationship between the length of retrieval times and their probabilities. In particular, the best-fitting values of the exponent were $u \approx 2$ for most individual participants.

Lévy distributions with $u \approx 2$ suggest an interesting correspondence with the animal foraging literature. This same distribution (or a truncated version of it) has been used to characterize the mobility patterns of a wide array of species, including Albatrosses, marine predators, monkeys, and people (González et al., 2008; Ramos-Fernández et al., 2004; Sims et al., 2008; Viswanathan et al., 1996). The theoretical justification for Lévy distributions of mobility patterns is that when resources are patchy (i.e., clustered), steps that follow this distribution are more likely to result in successful foraging than Gaussian-distributed steps are. In particular, in environments with patchy resources, $u = 2$ has been analytically shown to be the exponent that produces the most effective foraging (Viswanathan et al., 1999). As a result of this correspondence, Rhodes and Turvey (2007) suggested that human memory retrieval is essentially a foraging task within a mental representation, with response times equated with the distances between samples.

Aside from the distances between successive responses, researchers have also found long-range dependencies in cognitive time series. These dependencies have been investigated in a separate line of work from that on step sizes, at least in the literature on sampling from internal representations, though the two have been found to co-occur in investigations of eye movements, which is a process of sampling information from the external world (Rhodes et al., 2011). Gilden et al. (1995) first gave participants one minute's worth of training with a metronome that was set to produce a target temporal interval, such as 1 second. Following this short period of training, participants were asked to repeatedly press the spacebar on a computer keyboard every time they believed the target interval had elapsed. Participants then continued to "drum" the keyboard at the target interval 1,000 times in a row, which generated enough responses to characterize how a new response depended on previous responses.

There are various ways in which responses can depend on one another. Most cognitive models assume that responses are independent of one another. For example, standard drift-diffusion models of response times assume that people make independent responses on each trial (Ratcliff, 1978)[1]. Standard models of categorization assume that responses are independent given what has been learned (Nosofsky, 1986). Alternatively, the next response may depend solely on the most recent response, as would result from a model that produces a random walk over the space of possibilities (e.g., Abbott et al., 2015).

However, the temporal production task of Gilden et al. (1995) showed neither independence nor short-range dependencies, but instead showed long-range dependencies, termed $1/f$ noise. This name comes from the process of quantifying long-range dependencies: performing a Fourier transform and examining how the spectral power $S(f)$ depends on frequency f. For independent responses, $S(f) = 0$, for random walk responses, $S(f) = 1/f^2$, and for long-range dependencies $S(f) = 1/f$. These long-range $1/f$ dependencies are much more difficult to generate than independent responses or the dependencies found in random walks. As such they are often considered the hallmarks of complex processes, and have been found in the dynamics of leaky faucets, heart rates, turbulence, and stock markets (Bak,

[1]However, these models can be augmented to produce long-range dependencies (Wagenmakers et al., 2004).

1996).

$1/f$ noise is also not unique to temporal production tasks, and it has been found in a variety of similar cognitive tasks, including reproducing complex drumming patterns (Hennig et al., 2011). It has also been found in the estimation of spatial intervals, the time taken for mentally rotating objects, the time taken for lexical decision, the time required for either serial or parallel visual search, and in measures of implicit bias (Correll, 2008; Gilden, 1997; Gilden et al., 1995). Interestingly, these long-range dependencies disappear if a different task is interleaved with the task of interest (Gilden, 2001), or if the task is both very simple and unpredictable (Gilden et al., 1995).

Sampling algorithms to explain cognitive times series

These two properties of human samples, Lévy distributed distances between samples, and $1/f$ noise, are diagnostic for any theory of human inference via sampling. In probabilistic terms, a patchy representation is one that is multimodal: it has regions of high probability separated by troughs of low probability. These types of distributions are a difficult challenge for sampling algorithms with a sense of location, and distances between samples that follow a Lévy probability density distribution are a sign that the sampling algorithm used is successfully navigating this challenge.

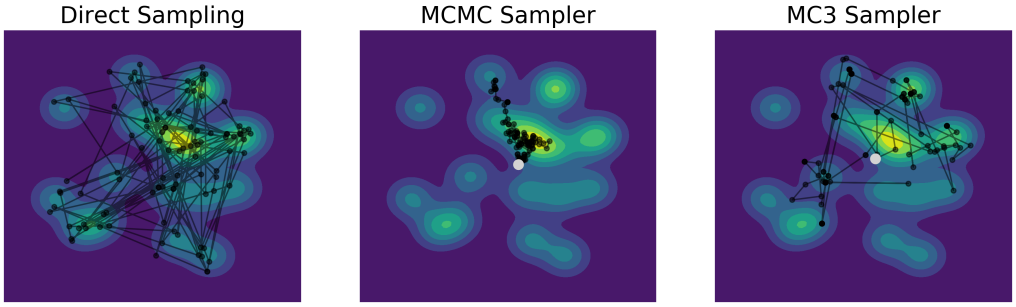
However for a sampling algorithm, $1/f$ noise is not at all desirable. Direct sampling, as described above, would be the most efficient in terms of sample size: N independent samples contain more information than the same number of dependent samples. For a set of dependent samples, we can estimate the number of independent samples that they would be equivalent to in terms of the information they contain, which is termed the effective sample size (ESS)

$$ESS = \frac{N}{1 + 2\sum_{k}^{\infty} C(k)} \tag{2}$$

where $C(k)$ is the degree of autocorrelation in the sample sequence at lag k. Thus, for an equivalent level of autocorrelation at the first lag, $1/f$ noise in the samples is also less efficient in terms of sample size than a random walk.

What algorithm could produce both of these properties, and why? As we have noted, direct sampling is the most efficient in uncovering the underlying distribution. Directly drawing independent samples from the underlying distribution will result in a posterior distribution resembling the true distribution, as long as sufficiently many samples are used. Thus, direct sampling will explore even far-apart modes of the true distribution. However, because these samples are drawn independently, direct sampling will not produce characteristic human autocorrelation patterns. Furthermore, while direct sampling allows exploration of far-apart modes, the rate at which distant and close regions of the landscape are visited does not resemble Lévy distributions, but instead will resemble uncorrelated, white noise (Zhu et al., 2018). For an example of direct sampling, see Figure 1a.

In contrast to direct sampling, MCMC does not require knowledge of the true underlying distribution and samples are not drawn independently. Instead, MCMC samplers are initialized at some random location and sequentially explore the probability landscape by performing a random walk, moving to nearby locations proportionally to the probability of the underlying space. For an illustration of the sampling behavior of MCMC in



(a) *Samples obtained from direct sampling.* (b) *Samples from MCMC.* (c) *Samples from MC^3.*

Figure 1

The behavior of the three sampling procedures in a patchy environment. The underlying distribution that the samplers are exploring corresponds to a mixture of 20 Gaussian distributions. Direct sampling does not require a starting position and subsequent samples are drawn independently from the underlying distribution. As a result, successive samples cover the whole distribution. In contrast, MCMC requires a starting state (solid gray point). Successive samples are proposed by performing a random walk, biased towards regions of high probability of the underlying distribution. However, these proposals do not allow the rapid exploration of multiple modes. Instead, the sampler will slowly traverse the space, and in many cases, never explore far-away modes. Finally, Metropolis-coupled Markov chain Monte Carlo (MC^3, explained below) also rests on a starting state and iteratively explores the underlying distribution. However, since it does so in multiple parallel chains at higher temperatures, it will occasionally jump into far-away modes.

multi modal environments, see Figure 1b. However, while these samples are correlated, the correlations do not exhibit long-range dependencies. Instead, MCMC samplers produce random-walk (Brownian) noise. Furthermore, distances between subsequent MCMC states do not resemble Lévy distributions. Importantly, this cannot be alleviated by replacing the common Gaussian proposal with a heavy-tailed distribution, as the resulting far-ranging proposals are very unlikely to ever be accepted, since these proposals tend to correspond to regions of very low probability[2].

Other algorithms can produce both long-range autocorrelations and Lévy distributed distances. We have previously suggested Metropolis-coupled Markov chain Monte Carlo (MC^3), a type of MCMC sampler[3]. As in MCMC, MC^3 is started at a random location and sequentially traverses the underlying probability landscape, producing a chain of locations, that, given enough samples, will be proportional to the true distribution. However, to allow the sampler to explore far-away areas of the distribution, MC^3 maintains several of these chains, each chain corresponding to an MCMC random walk.

The key idea underlying MC^3 is that of annealing Kirkpatrick et al., 1983 – to

[2]Heavy-tailed proposal distributions in a uniform space do however produce Lévy-distributed distances, as *every* proposal is equally likely to be accepted.

[3]This algorithm is also sometimes called parallel-tempering or replica-exchange MCMC.

allow the sampler to explore far-away modes each of its parallel chains explores an increasingly flatter version of the underlying distribution by applying different temperatures and thereby "melting down" modes of the underlying space. MC^3 generates chains in parallel at increasing temperatures and occasionally swaps the states of these chains, therefore allowing the sampler to jump to far-off modes of the distributions. For an example of MC^3 samples, see Figure 1c. The resulting posterior distribution is then commonly obtained from the first chain, for which no temperature is applied. As we have shown previously, this kind of sampler will sometimes produce long-range jumps, but commonly stay close to the previous location, thus producing Lévy distributed distances. Furthermore, samples obtained by MC^3 will produce slowly decaying autocorrelations resembling those of human data. Essentially, MC^3 pays the price of $1/f$ noise in order to generate the Lévy distributed distances that signify successful jumps between modes.

As, MC^3 is a type of MCMC procedure, it can account for the cognitive biases outlined above, by manipulating its starting point. Furthermore, inferences based on MC^3 will be strongly biased when the number of samples is reduced, for example due to the temporal constraints or cognitive load.

Going beyond individuals to markets

Interestingly, the properties arising in the structure of human behavior, including Lévy distributed distances and $1/f$ noise, also can arise in complex real-world tasks. In particular, many (although not all) financial time series, such as asset prices and currency exchange rates show these properties. It is therefore interesting to see if MC^3 can explain some of the excess variability seen in these prices. To do so, though, requires finding a bridge between internal samples (which might be an individual trader's estimate of the price at the next time step) and the asset prices. It turns out that, using a classic model from behavioral finance (De Long et al., 1990), it is possible to map samples (from traders) to prices in a straightforward way, such that it turns out that prices have the same statistical properties as the samples themselves[4] (Sanborn et al., 2019).

Before doing so, though, we consider the surprising empirical parallels between cognitive and financial time series. For example, the log price changes of cotton and stocks traded in the New York Stock Exchange has been modeled as a stochastic process with Lévy stable non-Gaussian increments (Fama, 1965; Mandelbrot, 1997). This indicates that large price changes in speculative markets happen far more frequently than a simple random-walk market would predict. That is, a person trading in a hypothetical random-walk market would expect a financial crisis of magnitudes greater than four standard deviations to occur merely once every 126 years. The random-walk assumption cannot, though, be correct, as that same person trading with a portfolio of the largest 100 UK companies listed in the London Stock Exchange would have experienced such losses 11 times just between 22 Oct 1987 and 21 Jan 2008, even excluding the 2008 financial crisis (Frain, 2009).

Another well-studied property of financial markets is volatility clustering (Granger & Ding, 1995; Mandelbrot, 1997). Qualitatively, this describes how large changes are more likely to be followed by large changes of both positive and negative changes, and

[4]The mapping between sampled expected future prices and actual future prices is linear, at least in the simplest case.

similarly for small changes (Mandelbrot, 1997). That is, markets do not allocate volatile time periods randomly across economic periods but the volatility of price changes is serially correlated. The long-range correlations in volatility has also been examined by the power spectrum analysis, and the absolute value of price changes of the Standard & Poor 500 index measured in one-hour intervals can be characterized as $1/f$ noise with the power-law exponent estimated equal to 0.7 (Liu et al., 1999; Mantegna & Stanley, 1997).

The heavy-tailed distributions of price changes and long-range dependence in the magnitudes of price changes resemble the Lévy distributed distances and $1/f$ noise that psychologists have observed in time estimation and animal naming tasks, where people's change in hypothesis space is measured (Bousfield & Sedgewick, 1944; Gilden, 1997, 2001). We have begun to do a more careful parallel analysis of price dynamic and cognitive time series is much needed in order to establish appropriate analogies and differences between price changes in the market and hypothesis changes in the mind (Sanborn et al., 2019). Here, we envision a hypothesis that a large part of variability in price changes can be attributed to the variability in opinion changes among market participants. As searching behaviours in mental space of hypotheses can be understood through sampling in a mental space, the price dynamics could reflect the stochastic behaviour of a sampler searching in a mental space regarding the future prospects of a commodity, a stock, or a financial portfolio.

Making the sampling algorithm more Bayesian

While sampling algorithms are commonly employed to approximate the answer that a complex probabilistic model would produce under uncertainty, they themselves are not Bayesian: the algorithms have no sense of the uncertainty in the answers that they produce. The algorithms can however be augmented to have probabilistic models over their outputs and therefore giving the sampling algorithms a way to incorporate uncertainty. In statistics and machine learning, this has been called Bayesian Monte Carlo (Rasmussen & Ghahramani, 2003), but for consistency with the above work we term it the *Bayesian sampler*.

Although the aforementioned sampling processes can explain a range of biases in human judgment as the consequence of dependent samples drawn from a large and unevenly distributed hypothesis space, this does not explain why biases also arise when the hypothesis space is small and easy to explore, such as outcomes of six-sided dice (Wedell & Moro, 2008). Human probability judgments in particular tend to exhibit a conservatism bias, in the sense that people's probability estimates tend to be less extreme than one would expect (Costello & Watts, 2014, 2017; Erev et al., 1994; Fiedler, 1991; Hilbert, 2012; Peterson & Beach, 1967). This effect cannot be explained by sampling in itself, but it can be shown that such conservatism is a natural consequence of reasoning with samples of limited size.

Imagine an urn with an unknown proportion of red and/or blue balls. If we draw one ball that turns out to be blue, then presumably we would not on that basis alone conclude that the urn contained only blue balls. Assuming that we lack any prior information regarding the proportion of red and blue balls (i.e., assuming a uniform prior distribution), the optimal Bayesian estimate is that the urn has a proportion of .67 blue balls, that is, that the probability of drawing a blue ball is .67. More generally, for a prior defined by the Beta distribution Beta(α, β) the optimal Bayesian probability estimate $\hat{P}$ based on S outcomes in a sample of size N is

$$\hat{P} = \frac{S + \alpha}{N + \alpha + \beta} \tag{3}$$

From this equation, it is easy to see that for any prior distribution where $\alpha = \beta > 0$ (i.e., any prior that is both symmetric and continuous) a Bayesian estimate must necessarily be moderated towards the middle of the distribution; even if we observe ten blue balls in a row, and assuming a uniform prior, the optimal Bayesian estimate of the probability of drawing a blue ball is approximately .92 rather than 1. Thus, if we presume that people generally make judgments based on a relatively small number of samples, and there is evidence suggesting this is the case (Goodman et al., 2008; Mozer et al., 2008; Vul et al., 2014), then, in order to minimize average error, conservatism is not a bias but a necessity.

This adjustment to the sampled proportions will sometimes result in incoherence in the sense that estimates for mutually exclusive events will not necessarily sum to one (De Finetti, 1937), and it has been shown that such adjustments will produce the same quantitative conservatism biases as have been observed empirically (Zhu et al., 2020). Indeed, if one assumes that conjunctions require more effort and time to sample than singular events, in turn resulting in relatively fewer samples, then conjunctions will quite naturally be subject to greater Bayesian adjustment, producing conjunction fallacies. For example, sampling the proportions of liberal-seeming persons who are bank tellers is arguably more straightforward than sampling the proportion of liberal-seeming persons who are both feminists and bank tellers and, as a consequence, the latter proportion is likely to be based on a smaller sample and therefore subject to greater adjustment.

Efficient accumulation of samples explains perceptual biases

This Bayesian sampler can also be extended to provide potential explanations for previously observed perceptual biases. When making estimates of perceptual features such as stimulus motion or numerosity, an initial decision regarding this feature can bias subsequent estimates: for example, deciding whether the linear direction of motion of a set of dots is clockwise or counter-clockwise of some boundary line pushes direct estimates of the direction of that motion further from the considered boundary compared with estimates made without such preceding decisions (Jazayeri & Movshon, 2007; Luu & Stocker, 2018; Zamboni et al., 2016). These results then contrast with the anchoring effects described in the previous sections: both tasks observe an impact of a decision on subsequent estimates, but in the cognitive domain, estimates move towards the queried boundary, while in the perceptual domain, estimates move away from the queried boundary.

While other explanations have been offered for these effects, we suggest that one explanation could be the reuse of samples between decisions and estimates, known as amortisation (Gershman & Goodman, 2014): learners may draw samples from a sensory representation to make their initial decision regarding the boundary, then reuse those samples in their estimates rather than expending further cognitive resources on additional sampling. This then creates a consistency between the two responses, as both the decision and estimate are based on the same set of observations, and so will both reflect any pattern contained in the sample. Simple amortised sampling with a fixed number of samples will not however produce such repulsion effects as there is no bias in this sample: the samples taken for a decision and reused for an estimate would be roughly equivalent to those used

for an estimate alone, leading to no systematic difference between these two cases. There is however the possibility that the number of samples is not fixed, but adapts to the strength of evidence collected to that point: if a set of samples provides compelling evidence towards a particular decision, the cost of further sampling may outweigh any potential gain in information, encouraging the early termination of sampling. This would then produce a bias in the sample, as sampling is more likely to stop where a high number of samples favor one decision: in the direction of motion example, if several successive samples are clockwise of the decision boundary, we may conclude that the true direction of motion is on this side, and stop sampling.

This then raises the question of how the threshold for the termination of sampling is decided. In keeping with the above sections, this could use a Bayesian updating process in which prior beliefs are updated with each sample to provide a posterior probability for each potential decision. This then allows for the comparison of the cost of terminating sampling with the cost of continuing sampling. Thus, the cost of termination is the expected probability of making an error based on the currently collected evidence, given by the minimum posterior at that point. The cost of continuation, meanwhile, is the sum of the inferred costs of the outcomes of future samples, plus a fixed cost for the generation of the sample itself. As with the probability estimates described above, we assume a Beta prior across the two potential sampling outcomes, here being the two sides of the decision boundary. The sampler therefore begins in a position of ambiguity, and updates this belief with each piece of evidence until the value of further information is outweighed by the cost of its generation. We term this system the Bayesian Amortised Sequential Sampler, or BASS. In comparisons with empirical data, we find BASS provides a better match to behaviour than previously offered candidate models: while other methods are able to predict the decision bias, BASS also explains the strong consistency between decisions and estimates shown by real learners, and more closely matches belief distributions collected from participants regarding their estimates (Zhu et al., 2019).

A question remaining to be answered regarding this process however is how these samples are drawn, as discussed previously in this chapter; amortisation describes the reuse of samples in decision making, but makes no assumptions about the mechanism by which these samples are originally generated. As noted previously, one possibility is direct sampling from the sensory representation; indeed, the results described above were based on direct sampling, and show that such a mechanism is able to predict previously observed perceptual biases. If however the BASS system were to use a sampler with a sense of location such as the MCMC algorithm noted in the previous sections, the resulting estimation system could then capture both these perceptual biases as well as the more traditional anchoring effects found in the cognitive domain (e.g. Russo and Schoemaker, 1989). Specifically, as noted above, MCMC can account for anchoring effects under the suggestion that the anchor provides a starting point for the sampler which, under a limited number of samples, the chain is unable to move far from (Lieder et al., 2012; Lieder et al., 2018a, 2018b). Combining an MCMC sampling algorithm with an adaptive stopping rule such as BASS could then provide a single estimation system able to produce both the attraction and repulsion effects observed in existing research. Future work may then wish to examine whether both effects can appear in the same task as a test of such a system, including the potential cross-over in these effects between the cognitive and perceptual domains.

Conclusions

In this chapter, we have explored the idea that the brain carries out approximate probabilistic reasoning through local sampling, rather than through intractable Bayesian calculations. This approach has many of the virtues of a Bayesian analysis of cognition, because it explains why the cognitive system will reason successfully, if the number of samples is sufficiently large. In practice, though, the probability distributions that the brain must deal with will be enormously complex and cannot possibly be sampled in their entirety. If the Bayesian sampling perspective is correct, we might hope that the biases observed in human cognition, including those directly involved the probabilistic estimation, might be those that would be expected from limited Bayesian sampling, where there will be excessive influence of the starting point (as in the anchoring effect in probability judgment). Moreover, a concrete sampling account requires choosing a specific sampling algorithm. We have argued that the characteristic statistics of successive samples (e.g., $1/f$ autocorrelation between durations in rhythmic tapping; and a Lévy distribution on the sizes of jumps between successive durations) provide powerful empirical constraints on the sampling process. We suggest that a specific sampling algorithm, Metropolis-coupled Markov chain Monte Carlo (MC^3), designed to deal with complex multimodal distributions, may be a good candidate sampling mechanism, able to capture patterns in both human judgements, and financial time series, which presumably arise from the aggregation of many judgements. We note that the brain should not simply read off the relative frequencies from any sample that it generates. Instead, the correction of such a sample based on prior knowledge is likely to be appropriate, leading to what appears to be conservatism in some cognitive tasks. Moreover, given that sampling is likely to be cognitively slow and costly, an intelligent sampler will actively continue or terminate the sampling process, depending on how results are accumulating. As we have seen, this can lead to estimation biases that push away from a decision boundary, in some ways yielding the opposite pattern to that observed in anchoring.

From the perspective creating human-like computation, we suggest that sampling algorithms provide an attractive research direction. Such algorithms provide a mechanism for approximating complex calculations required to deal with a rich and highly uncertain world, a challenge as relevant for artificial intelligence as for the human brain. Even for designers of machine intelligence that only aspire to effectively interact with people rather than imitate them, samples can be a common framework for collaboration (e.g., Sanborn, Griffiths, and Shiffrin, 2010). Finally, an interesting commonality in our work is that people seem to utilise only a handful of samples – far fewer than what is considered the minimum in statistical applications (e.g., Gelman and Rubin, 1992) – but use them effectively. As people of course operate effectively in the world despite theses restrictions, this may offer a broad lesson for designers of machine intelligence that needs to operate in real-time: careful use of a few samples can provide a rough but effective characterisation of uncertainty.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. *Psychological Review*, *122*(3), 558–569.

Abbott, J. T., & Griffiths, T. L. (2011). Exploring the influence of particle filter parameters on order effects in causal learning. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Aragones, E., Gilboa, I., Postlewaite, A., & Schmeidler, D. (2005). Fact-free learning. *American Economic Review*, *95*, 1355–1368.

Bak, P. (1996). *How nature works: The science of self-organised criticality*. Springer-Verlag.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

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

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765.

Bousfield, W. A., & Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. *The Journal of General Psychology*, *30*(2), 149–165.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*(7), 335–344.

Correll, J. (2008). 1/f noise and effort on implicit measures of bias. *Journal of Personality and Social Psychology*, *94*(1), 48–59.

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, *121*(3), 463–480.

Costello, F., & Watts, P. (2017). Explaining high conjunction fallacy rates: The probability theory plus noise account. *Journal of Behavioral Decision Making*, *30*(2), 304–321.

Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive Psychology*, *96*, 1–25.

De Finetti, B. (1937). La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré*, *7*(1), 1–68.

De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy*, *98*(4), 703–738.

Doucet, A., de Freitas, N., & Gordon, N. (2001). *Sequential Monte Carlo methods in practice*. Springer.

Edwards, W. (1961). Behavioral decision theory. *Annual Review of Psychology*, *12*(1), 473–498.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over-and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*(3), 519–527.

Fama, E. F. (1965). The behavior of stock-market prices. *The Journal of Business*, *38*(1), 34–105.

Fiedler, K. (1991). Heuristics and biases in theory formation: On the cognitive processes of those concerned with cognitive processes. *Theory & Psychology*, *1*(4), 407–430.

Frain, J. C. (2009). *Studies on the application of the alpha-stable distribution in economics*. Trinity College.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Gershman, S. J., & Goodman, N. D. (2014). Amortized inference in probabilistic reasoning. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society* (pp. 517–522). Cognitive Science Society.

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation*, *24*, 1–24.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Gilden, D. L. (1997). Fluctuations in the time required for elementary decisions. *Psychological Science*, *8*(4), 296–301.

Gilden, D. L. (2001). Cognitive emissions of 1/f noise. *Psychological Review*, *108*(1), 33–56.

Gilden, D. L., Thornton, T., & Mallon, M. W. (1995). 1/f Noise in Human Cognition. *Science*, *267*, 1837–1839.

González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779–782.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*, 108–154.

Granger, C. W. J., & Ding, Z. (1995). Some properties of absolute return: An alternative measure of risk. *Annales d'Economie et de Statistique*, 67–91.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

Griffiths, T. L., & Tenenbaum, J. B. (2011). Predicting the future as Bayesian inference: People combine prior knowledge with observations when estimating duration and extent. *Journal of Experimental Psychology: General*, *140*(4), 725–743.

Hennig, H., Fleischmann, R., Fredebohm, A., Hagmayer, Y., Nagler, J., Witt, A., Theis, F. J., & Geisel, T. (2011). The nature and perception of fluctuations in human musical rhythms. *PloS One*, *6*(10), e26457.

Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin*, *138*(2), 211.

Houlsby, N. M. T., Huszár, F., Ghassemi, M. M., Orbán, G., Wolpert, D. M., & Lengyel, M. (2013). Cognitive tomography reveals complex, task-independent mental representations. *Current Biology*, *23*(21), 2169–2175.

Jazayeri, M., & Movshon, J. A. (2007). A new perceptual illusion reveals mechanisms of sensory decoding. *Nature*, *446*(7138), 912–915.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kirkpatrick, S., Gelatt Jr, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 671–680.

Lieder, F., Griffiths, T. L., & Goodman, N. D. (2012). Burn-in, bias, and the rationality of anchoring. *Advances in Neural Information Processing Systems*, 2690–2798.

Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2018a). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, *25*(1), 322–349. https://doi.org/10.3758/s13423-017-1286-8

Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2018b). Empirical evidence for resource-rational anchoring and adjustment. *Psychonomic Bulletin & Review*, *25*(2), 775–784.

Liu, Y., Gopikrishnan, P., Cizeau, P., Meyer, M., Peng, C.-K., & Stanley, H. E. (1999). Statistical properties of the volatility of price fluctuations. *Physical Review E*, *60*(2), 1390–1400.

Luu, L., & Stocker, A. A. (2018). Post-decision biases reveal a self-consistency principle in perceptual inference. *eLife*, *7*, e33334. https://doi.org/10.7554/eLife.33334

Mandelbrot, B. B. (1997). The variation of certain speculative prices. *Fractals and scaling in finance* (pp. 371–418). Springer.

Mantegna, R. N., & Stanley, H. E. (1997). Physics investigation of financial markets. *Proceedings of the International School of Physics "Enrico Fermi", Course CXXXIV, IOS Press, Amsterdam*.

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.

Mosteller, F., & Nogee, P. (1951). An experimental measurement of utility. *Journal of Political Economy*, *59*(5), 371–404.

Mozer, M. C., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science*, *32*(7), 1133–1147.

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

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Pantelis, P. C., Baker, C. L., Cholewiak, S. A., Sanik, K., Weinstein, A., Wu, C.-C., Tenenbaum, J. B., & Feldman, J. (2014). Inferring the intentional states of autonomous virtual agents. *Cognition*, *130*(3), 360–379.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*, 29–46.

Petzschner, F. H., Glasauer, S., & Stephan, K. E. (2015). A Bayesian perspective on magnitude estimation. *Trends in Cognitive Sciences*, *19*(5), 285–293.

Ramos-Fernández, G., Mateos, J. L., Miramontes, O., Cocho, G., Larralde, H., & Ayala-Orozco, B. (2004). Lévy walk patterns in the foraging movements of spider monkeys (ateles geoffroyi). *Behavioral Ecology and Sociobiology*, *55*(3), 223–230.

Rasmussen, C. E., & Ghahramani, Z. (2003). Bayesian Monte Carlo. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 505–512). MIT Press. http://papers.nips.cc/paper/2150-bayesian-monte-carlo.pdf

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Rhodes, T., Kello, C., & Kerster, B. (2011). Distributional and temporal properties of eye movement trajectories in scene perception. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*(33), 178–184.

Rhodes, T., & Turvey, M. T. (2007). Human memory retrieval as Lévy foraging. *Physica A: Statistical Mechanics and its Applications*, *385*(1), 255–260.

Robert, C. P., Elvira, V., Tawn, N., & Wu, C. (2018). Accelerating MCMC algorithms. *Wiley Interdisciplinary Reviews: Computational Statistics*, *10*(5), e1435.

Russo, J. E., & Schoemaker, P. J. (1989). *Decision traps: Ten barriers to brilliant decision-making and how to overcome them*. Simon; Schuster.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893.

Sanborn, A. N., Chater, N., Zhu, J.-Q., & Spicer, J. (2019). *Macroeconomics implications of the sampling brain* (tech. rep.). National Institute of Economics and Social Research.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to the rational model of categorization. *Psychological Review*, *117*, 1144–1167.

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

Sanborn, A. N., Mansinghka, V., & Griffiths, T. L. (2013). Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychological Review*, *120*, 411–437.

Savage, L. J. (1954). *Foundations of statistics*. John Wiley & Sons.

Sims, D. W., Southall, E. J., Humphries, N. E., Hays, G. C., Bradshaw, C. J., Pitchford, J. W., James, A., Ahmed, M. Z., Brierley, A. S., Hindell, M. A., et al. (2008). Scaling laws of marine predator search behaviour. *Nature*, *451*(7182), 1098.

Sloman, S., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., & Fox, C. R. (2004). Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(3), 573–582.

Susskind, J. M., Hinton, G. E., Movellan, J. R., & Anderson, A. K. (2008). Generating facial expressions with deep belief nets. In V. Kordic (Ed.), *Affective computing, focus on emotion expression, synthesis and recognition* (pp. 421–440). ARS Publishers.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Tversky, A., & Kahneman, D. (1983). Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Viswanathan, G. M., Afanasyev, V., Buldyrev, S. V., Murphy, E. J., Prince, P. A., & Stanley, H. E. (1996). Lévy flight search patterns of wandering albatrosses. *Nature*, *381*(6581), 413.

Viswanathan, G. M., Buldyrev, S. V., Havlin, S., da Luz, M. G. E., Raposo, E. P., & Stanley, H. E. (1999). Optimizing the success of random searches. *Nature*, *401*(6756), 911–914.

von Neumann, L. J., & Morgenstern, O. (1947). *Theory of games and economic behavior* (Vol. 60). Princeton University Press.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*, 599–637.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*, 101–118.

Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (2004). Estimation and interpretation of $1/f^{\alpha}$ noise in human cognition. *Psychonomic Bulletin & Review*, *11*(4), 579–615.

Wedell, D. H., & Moro, R. (2008). Testing boundary conditions for the conjunction fallacy: Effects of response mode, conceptual focus, and problem type. *Cognition*, *107*(1), 105–136.

Wolpert, D. M. (2007). Probabilistic models in human sensorimotor control. *Human Movement Science*, *26*(4), 511–524.

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

Zamboni, E., Ledgeway, T., McGraw, P. V., & Schluppeck, D. (2016). Do perceptual biases emerge early or late in visual processing? Decision-biases in motion perception. *Proceedings of the Royal Society B*, *283*, 20160263. https://doi.org/10.1098/rspb.2016.0263

Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2018). Mental sampling in multimodal representations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 5748–5759).

Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2019). Why decisions bias perception: An amortised sequential sampling account. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st annual conference of the cognitive science society* (pp. 3220–3226). Cognitive Science Society.

Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). The Bayesian sampler: Generic Bayesian inference causes incoherence in human probability judgments. *Psychological Review*.