Complex Probabilistic Inference: From Cognition to Neural Computation

Samuel J. Gershman[1] and Jeffrey M. Beck[2]
[1]Department of Psychology and Center for Brain Science, Harvard University
[3]Department of Neurobiology, Duke University Medical School

March 2, 2017

Abstract

Our understanding of probabilistic inference in the brain has progressed rapidly. However, there remains a big gap between the relatively simple probabilistic inference problems facing low-level sensory systems and the intractably complex problems facing high-level cognitive systems. Psychologists have begun exploring cognitively plausible algorithms for approximately solving complex inference problems. We review recent attempts to connect these algorithmic accounts to neural circuit mechanisms, and argue that neural mechanisms for solving low-level sensory inference problems can be extended to tackle complex inference problems.

Keywords: Bayesian inference; computational neuroscience; cognitive science
Word count: 6548

1 Introduction

Sensory receptors collect a limited amount of noisy data, from which the brain must reconstruct the external world. This problem is fundamentally ambiguous; for example, the image of an object projected on the retina is equally consistent with a small object close to the eye and a large object far away from the eye. To resolve such ambiguities, an ideal observer should combine sensory data with prior knowledge (e.g., the typical sizes of objects) through the application of Bayes' rule. However, these constraints still do not fully resolve all ambiguities—uncertainty is an irreducible facet of information processing. The brain's reconstruction of the external world explicitly represents its uncertainty in the form of probability distributions over internal models. Understanding the nature of these representations and how they are computed is the goal of a vigorous program of research [1].

Much of the theoretical neuroscience research on "simple" (low-dimensional and analytically tractable) probabilistic inference has focused on low-level perceptual domains such as multi-sensory cue integration and motion perception (e.g., [2, 3, 4]). While the same principles apply, at least in theory, to higher-level cognitive domains, the increase in complexity of the internal models poses daunting computational challenges. The intractability of probabilistic inference in even modestly complex models necessitates approximations, which means that the kinds of mechanisms previously proposed for probabilistic inference in low-level neural systems (mostly based on exact inference schemes) may not be appropriate for high-level cognition. It is not clear how these neural mechanisms can be "scaled up" to the kinds of domains that cognitive psychologists study. Nonetheless, we know that complex inference pervades these domains [5], and is also an inherent part of basic sensory processing in visual and auditory cortex, as we illustrate below. A variety of neural schemes for complex inference have been proposed [6, 7, 8, 9, 10], but so far these have made relatively little contact with the rich literature on psychological mechanisms.

We attempt to bridge the gap between neural mechanisms for simple inference and psychological mechanisms for complex inference. We begin by briefly reviewing evidence for complex inference, using examples of both low-level sensory processing and high-level cognition. We then describe the algorithmic challenges facing complex probabilistic inference. These challenges have been tackled in the machine learning literature by using two families of techniques: Monte Carlo approximations [11], which replace the exact posterior with a set of stochastically generated samples, and variational approximations [12], which replace the exact posterior with a tractable surrogate distribution optimized to be as close as possible to the exact posterior. Both families have been explored as psychologically plausible mechanistic models of probabilistic inference [13]. Finally, we discuss attempts to implement these techniques in neural circuits, and the experimental evidence supporting different implementation schemes.

2 Tractable algorithmic approaches to complex inference

Given data D, Bayes' rule stipulates how to convert *prior* beliefs $P(Z)$ about latent variable Z into *posterior* beliefs $P(Z|D)$:

$$P(Z|D) = \frac{P(D|Z)P(Z)}{P(D)}, \tag{1}$$

where $P(D|Z)$ is the *likelihood* of data D conditional on latent variable Z. For example, D might be an image patch (corrupted by sensory noise) and Z is the orientation of an edge in the patch. When the image contrast is higher (lower sensory noise), or the display is viewed for longer (evidence accumulation), the likelihood of the data under the true orientation increases. The prior encodes the distribution of oriented edges in natural images (e.g., cardinal orientations are more common than oblique orientations). Taken together, the prior and likelihood can be understood as constituting a *generative model*—a recipe for generating observed data from latent variables. Bayes' rule inverts this generative model to produce a belief about the latent variables after observing data.

Bayesian models have been applied to a wide variety of perceptual phenomena [14, 15], and form the cornerstone of signal detection theory [16]. The same principles, applied to different generative models, have been used to explain more complex cognitive phenomena, such as causal reasoning [17], semantic memory [18], language processing [19], and concept learning [20, 21, 22]. One hallmark of these models is that they are *complex*: the latent variables are high dimensional and often combinatorial. As a consequence, exactly computing Bayes' rule is intractable. Below, we summarize several tractable algorithmic approximations.

2.1 Monte Carlo methods

By drawing a set of samples $\{Z^n\}_{n=1}^N$ from the posterior distribution, the posterior probability density can then be approximated as an empirical point-mass function:

$$P(Z|D) \approx \frac{1}{N}\sum_{n=1}^{N}\delta[Z^n, Z], \tag{2}$$

where $\delta[\cdot,\cdot] = 1$ if its arguments are equal, and 0 otherwise (for simplicity our exposition uses discrete distributions, but applies with minor modifications to the continuous case). As the number of samples N approaches infinity, the posterior is approximated to arbitrary accuracy.

The key challenge in applying Monte Carlo methods to Bayesian inference is generating the samples, since the posterior cannot be sampled directly. Most approaches involve sampling from an alternative distribution from which an approximate posterior can be constructed. We will focus on the two most widely used approaches: *importance sampling* and *Markov chain Monte Carlo* (MCMC).

The idea behind importance sampling is to sample from a proposal distribution $\phi(Z)$ and then weight the samples according to their "importance":

$$P(Z|D) \approx \frac{1}{N}\sum_{n=1}^{N} w^n\delta[Z^n, Z] \tag{3}$$

$$w_t^n \propto \frac{P(D|Z^n)P(Z^n)}{\phi(Z^n)}. \tag{4}$$

When the proposal is equal to the prior, $\phi(Z) = P(Z)$, importance sampling reduces to likelihood weighting: $w^n \propto P(D|Z^n)$. *Particle filtering* is a form of importance sampling applied to sequentially structured models. For example, in a hidden Markov model, the latent variable at time t depends on its state at time $t-1$ through the transition distribution $P(Z_t|Z_{t-1}$, and the observations at time t are generated conditional on Z_t through an observation distribution $P(D_t|Z_t)$. Particle filtering samples $Z_t^n \sim \phi(\cdot)$ and applies the importance sampling equation recursively:

$$P(Z_t|D_{1:t}) \approx \frac{1}{N}\sum_{n=1}^{N} w_t^n\delta[Z_t^n, Z_t] \tag{5}$$

$$w_t^n \propto w_{t-1}^n \frac{P(D_t|Z_t^n)P(Z_t^n|Z_{t-1}^n)}{\phi(Z_t^n)}, \tag{6}$$

where $D_{1:t}$ denotes the set of observations $\{D_1, \ldots, D_t\}$. Analogously to the general importance sampling method, sampling from the transition distribution yields likelihood weighting: $w_t^n \propto w_{t-1}^n P(D_t|Z_t^n)$.The success of importance sampling of particle filtering depends crucially on the proposal distribution; the prior or transition distribution is not in general the optimal choice. A common pitfall is degeneracy, where most weights go to zero and the effective sample size shrinks accordingly. This occurs when the proposal distribution focuses on a region of the hypothesis space that has low posterior probability, such that few samples land in regions of high posterior probability, and these samples end up dominating the Monte Carlo approximation.

Particle filters have been successfully applied to problems with dynamical structure like object tracking [23] and robot navigation [24]. However, for problems with complex static structure they are less widely applied, and static importance sampling methods will often fail on these problems due to the difficulty in specifying a good proposal distribution. MCMC methods can overcome this limitation to some extent by making local stochastic updates to hypothesis samples. The basic idea is to construct a Markov chain whose stationary distribution is the posterior. One generic way to do this, known as the Metropolis-Hastings algorithm [25], is to draw samples from a proposal distribution $\phi(Z^n|Z^{n-1})$ and accept the proposal with probability

$$A = \min\left[1, \frac{P(D|Z^n)P(Z^n)\phi(Z^{n-1}|Z^n)}{P(D|Z^{n-1})P(Z^{n-1})\phi(Z^n|Z^{n-1})}\right]. \tag{7}$$

If the proposal is rejected, $Z^n = Z^{n-1}$. Importantly, the proposal distribution can make local modifications to Z^{n-1}. When the proposal distribution is symmetric, $\phi(Z^{n-1}|Z^n) = \phi(Z^n|Z^{n-1})$, the acceptance function simplifies to:

$$A = \min\left[1, \frac{P(D|Z^n)P(Z^n)}{P(D|Z^{n-1})P(Z^{n-1})}\right]. \tag{8}$$

Intuitively, this equation says that proposals that increase the joint probability will be deterministically accepted, but proposals that decrease the joint probability can also be accepted with some probability. Writing the acceptance function in this way allows us to draw a connection between Metropolis-Hastings and an important stochastic optimization algorithm known as *simulated annealing* [26], which raises the joint probability to a power $1/T$, where T is a "temperature" parameter emulating the temperature in a thermodynamic system. When $T > 1$, the posterior is overdispersed, and when $T < 1$, the posterior is underdispersed. By decreasing T as a function of n (according to an annealing schedule), the equilibrium distribution will collapse onto the mode of the posterior. High initial temperatures serve the purpose of facilitating exploration of the hypothesis space without getting stuck in local optima.

A special case of Metropolis-Hastings, known as *Gibbs sampling* [27], draws iteratively from the conditional distribution $P(Z_i^n|Z_{\mathcal{C}(i)}^{n-1})$ where i indexes one variable (or a collection of variables) in Z and $\mathcal{C}(i)$ denotes the set of other variables upon which Z_i depends probabilistically (formally, the "Markov blanket" of Z_i). Gibbs sampling is one of the most widely used MCMC methods, and will appear again in our discussion of psychological and neural mechanisms.

Note that while importance sampling and particle filtering represent multiple hypotheses *simultaneously*, MCMC methods typically represent hypotheses sequentially. This sequential structure is dictated by the algorithmic dynamics, rather than the structure of the probabilistic model as in particle filtering (although the model structure will also have an influence on the dynamics of MCMC). Recent work has explored ways to meld these approaches, by considering an ensemble of samples that can evolve according to a Markov chain [28].

2.2 Variational methods

Monte Carlo methods can be viewed as "nonparametric" in the sense that the posterior approximation does not have a fixed structure: the "complexity" of the approximation grows with the number of samples. This flexibility comes with asymptotically vanishing approximation error, but at possibly great computational expense. An alternative approach is to consider approximations belonging to some parametric family, and choose the parameters that make the approximation as similar as possible to the true posterior. If the posterior does not belong to the parametric family, then approximation error will never vanish, but the optimal parametric approximation may be sufficiently good and computationally cheaper than sampling.

Variational methods [12] provide a principled framework for choosing a parametric approximation, by formulating inference as an optimization problem. Let $Q(Z)$ be a parametrized distribution belonging to family $\mathcal{Q}$. The most widely used variational method chooses $Q(Z)$ to minimize the Kullback-Leibler (KL) divergence between $Q(Z)$ and $P(Z|D)$:

$$\mathrm{KL}[Q(Z)||P(Z|D)] = \sum_Z Q(Z) \log \frac{Q(Z)}{P(Z|D)}. \tag{9}$$

When $Q(Z)$ is chosen to factorize over variables (or groups of variables), $Q(Z) = \prod_i Q_i(Z_i)$, this optimization problem is known as *mean-field variational inference*. Another approach is to optimize the opposite KL divergence, $\mathrm{KL}[P(Z|D)||Q(Z)]$; this leads to *expectation propagation* [29].

Optimizing the KL divergence is not itself tractable, since it is a function of the true posterior. However, minimizing $\mathrm{KL}[Q(Z)||P(Z|D)]$ is equivalent to maximizing a lower bound $\mathcal{L}[Q]$ on the log marginal likelihood (or "evidence"), $\log P(Z)$, using the following relation:

$$\log Z = \mathcal{L}[Q] + \mathrm{KL}[Q(Z)||P(Z|D)] \tag{10}$$

$$\mathcal{L}[Q] = \sum_Z Q(Z) \log \frac{P(D|Z)P(Z)}{Q(Z)}. \tag{11}$$

Notice that the evidence lower bound $\mathcal{L}[Q]$ depends only on the joint probability of Z and D, and hence is tractable to compute. Moreover, when the factors of $Q(Z)$ are in the same conjugate-exponential family as $P(D, Z)$, then $\mathcal{L}[Q]$ can be optimized via closed-form coordinate ascent updates (we present an example below).

3 Psychological mechanisms

Both Monte Carlo and variational algorithms have been proposed as psychologically plausible mechanisms for probabilistic inference [13], although Monte Carlo algorithms have received much more attention and thus will be our focus in this section. Broadly speaking, the psychological evidence for Monte Carlo algorithms falls into 3 categories: (1) stochasticity, (2) dynamics, and (3) resource constraints. It should be noted at the outset, however, that these sources of evidence may not decisively discriminate between Monte Carlo and variational algorithms. While the evidence for variational algorithms mostly comes from studies implicating particular parametric approximations, variational algorithms can also exhibit stochasticity [30, 31], as well as dynamics and resource constraints resembling Monte Carlo methods. Because these different approaches have rarely been directly compared to each other as models of psychological phenomena, discriminating them empirically remains an open challenge.

3.1 Stochasticity

Monte Carlo methods are inherently stochastic. One implication of this property is that mental representations, and possibly also behavioral responses, should be stochastic. Bayesian sampling specifically predicts that the stochasticity should reflect the posterior distribution: high probability hypotheses should be sampled more often than low probability hypotheses. This is reminiscent of "probability matching" in instrumental choice, the observation that humans and animals choose actions with probability proportional to their payoffs [32]. Indeed, evidence suggests that the visual system also uses a probability matching strategy. Wozny et al. [33] studied location estimation in an auditory-visual cue combination experiment, where probability matching predicts that the distribution of location estimates should be bimodal when auditory and visual information conflict, but importantly there will be some probability mass in between the two modes due to their overlapping distribution. Most participants' estimates were consistent with this probability matching strategy (see also [34, 35]). However, this assertion has been controversial, with some arguing, in accordance with classical signal detection theory, that humans make Bayes-optimal perceptual decisions [36]. Other evidence suggests that the stochastic representation of belief is a power function of the posterior, such that the response rule is somewhere between probability matching and selecting the posterior mode [37]. Probability matching has also been found in higher-level cognition. The variability of children's causal inferences matches the posterior distribution [38], and some evidence from adult concept learning is also consistent with the probability matching hypothesis [20].

One important subtlety in considering probability matching is that the Monte Carlo methods do not require that the *decision rule* is stochastic; it may be a deterministic function of the posterior approximation. If the approximation is stochastic, then the decision rule will be a stochastic function of the data. If enough samples are drawn, variability due to the Monte Carlo approximation will eventually disappear, and decisions will appear deterministic as a function of data. It has been argued that because sampling is costly and good decisions often do not require a high fidelity approximation, only a small number of samples will typically be drawn, and therefore probability matching will arise naturally even with a deterministic decision rule [39, 40].

A particularly interesting form of stochasticity arises in multistable perception, where conflicting interpretations of sensory data alternately dominate the percept. The stochastic dynamics underlying multistable perception have been the subject of extensive study, and are characterized by a richly varied phenomenology [41]. The most prominent example is binocular rivalry, where different images are presented to each eye, resulting in one image dominating the percept at a time [42]. Gershman et al. [43] proposed a probabilistic model of binocular rivalry that used Gibbs sampling to approximate the posterior. They showed that this model could explain not only switching behavior, but also traveling waves [44], the contrast-dependence of dominance durations [45], and the conditions under which fused percepts will be observed [46, 47]. In related work, Moreno-Bote et al. [48] showed how an attractor neural network implementing another form of MCMC (Langevin Monte Carlo) could account for multistable perception of drifting gratings.

While stochasticity is a hallmark of Monte Carlo methods, it can also arise from other algorithms. For example, stochastic optimization uses noise to explore the hypothesis space, but is not forming an approximation of the posterior. Randomness in the initialization of otherwise deterministic algorithms can also produce stochasticity that is not meaningfully related to approximate inference.

In some cases, apparent stochasticity may even be an illusion; Beck et al. [49] have argued that behavioral variability may be explained by suboptimal, deterministic inference algorithms. Thus, interpretations of noise in terms of Monte Carlo sampling must be made with caution, an issue we explore further below.

3.2 Dynamics and resource constraints

Particle filtering and MCMC exhibit conceptually different dynamics. Whereas particle filtering involves multiple samples evolving as new data are collected, MCMC involves an individual sample evolving over time given a fixed data set. Both forms of dynamics are constrained by the structure of the probabilistic model. For example, Gershman et al. [43] showed how variations of the underlying image model shaped the time course of binocular rivalry: altering spatial coupling of neighboring nodes in the image increased the propagation time of traveling waves, consistent with the data of Wilson et al. [44]. Similarly, the dynamics of particle filtering reflect the transition structure of the probabilistic model. In multiple object tracking, for example, the set of represented hypotheses (object identities) evolve in accordance with assumptions about object motion. When these assumptions are violated, memory is impaired [50].

The number of samples in particle filtering can be used as a proxy for cognitive resource availability: more resources translate to more samples. This form of explanation has been invoked to explain failures of change detection [51], object tracking [50], category learning [52], and word segmentation [53]. Resource constraints can interact with across-trial dynamics; for example, the correct hypotheses may not be represented in the ensemble if it is disfavored by data early in the sequence and is therefore killed off by resampling. This gives rise to “garden path” effects in linguistics, where sentences like “the horse raced past the barn fell” are difficult to comprehend because the correct parse is disfavored by the early data [54].

4 Neural implementations of probabilistic inference

We now turn to the question of how the brain might implement the approximate inference schemes described above. We begin with a generic treatment of neural probability coding, and then consider how sampling and variational algorithms could produce such codes in a biologically plausible manner.

4.1 Coding and computation

There currently exist two (not necessarily mutually exclusive) hypotheses about the neural implementation of probabilistic inference. The core distinctions between them concern how neurons represent probability distributions and how cortical circuits approximate probabilistic inference. Curiously, most (if not all) of the proposed neural implementations of probabilistic inference share a common overall network structure. This is because networks used to implement inference mimic the structure of the associated generative model. Figure 1 depicts this relationship. On the left

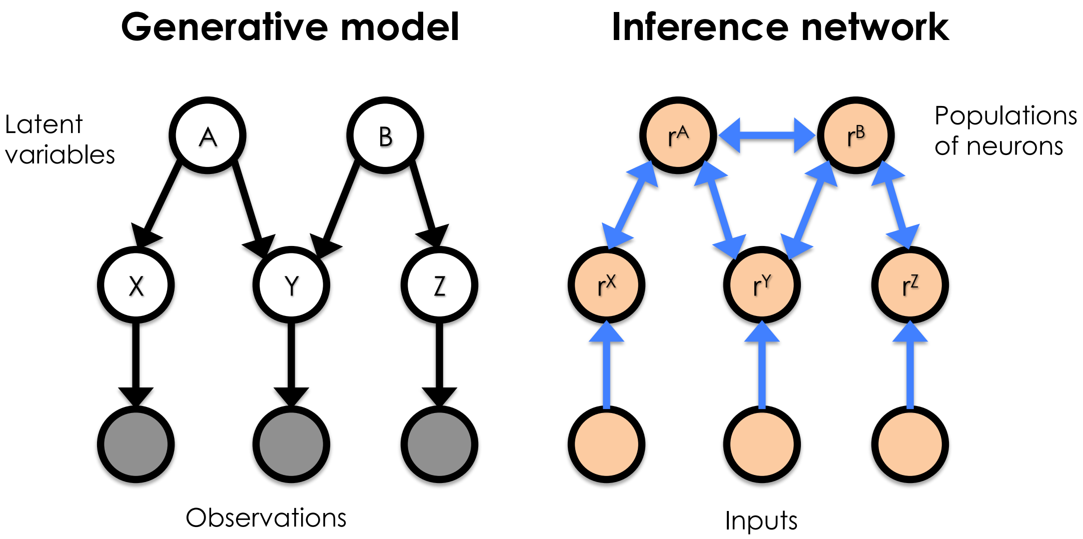

Figure 1: **Illustration of probabilistic inference with neurons**. (Left) Generative model: unshaded nodes represent latent variables, shaded nodes represent observed variables, and arrows represent probabilistic dependencies. (Right) Inference network: nodes represent neural populations and arrows represent pathways between populations.

is a generic directed graphical model. Here, nodes represent variables and arrows encode conditional independence relationships. For example, the arrows from A and B to Y indicate that Y is conditionally independent of all other variables given A and B. On the right is the associated network for probabilistic inference. Here, a node is to be thought of as containing a population neurons that represent marginal posterior distributions over the associated latent variable in the generative model. Arrows going into a particular node tell us that in order to update the beliefs about the associated latent variable we need information from the population of neurons that is at the source of those the arrows. This relationship between generative models and inference networks is most strongly associated with message passing algorithms for probabilistic inference on directed graphical models, but is also a generic property of the vast majority of the approximate methods used for performing probabilistic inference.

So while the structure of approximate probabilistic inference algorithms remains the same, what differs between competing hypotheses for neural implementations of probabilistic inference is (1) the means by which probability distributions are represented and (2) the specific mathematical details of the computations performed by the neural circuity. For example, consider a simple cue combination or evidence integration task depicted in Figure 2. Here, S is the position of an object while A and V are noisy representations of that position given either auditory or visual information. In the neural network on the right, node r^S represents a population of neurons used to represent a probability distribution over position, and nodes r^A and r^V represent populations of neurons encoding auditory and visual information (respectively) about position. For simplicity we will assume that the sensory neurons (r^A and r^V) encode a Gaussian likelihood over position so that

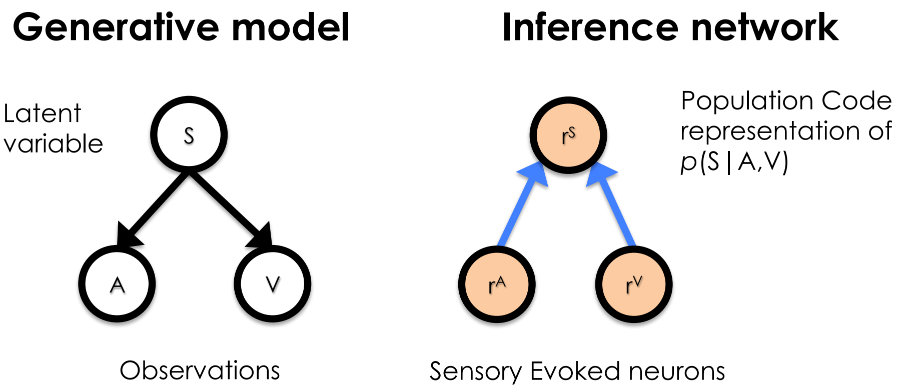

Figure 2: **Multisensory cue combination**. (Left) Generative model. S represents object position, A represents auditory information about object position, and V represents visual information about object position. (Right) Inference network.

as a function of the position S of the object we have

$$A|S \sim \mathcal{N}(\mu_A, \sigma_A^2), \tag{12}$$

$$V|S \sim \mathcal{N}(\mu_V, \sigma_V^2). \tag{13}$$

When this is the case, Bayes rule implies that the posterior over S given visual and auditory information is also normally distributed and can be obtained by multiplying prior and likelihoods:

$$P(S|A, V) \propto P(S)P(A|S)P(V|S). \tag{14}$$

Now consider two neural representations of probability distributions: a labeled line probability code [55] and a labeled line log probability code [2]. For the probability code, neural activity in neuron i in a population r^S is assumed to be proportional to the probability that the latent variable encoded by that population takes on value S_i. For the observed variables, the corresponding neuron encodes the likelihood of the observed variables given $S = S_i$. In the multisensory cue combination example, this means:

$$P(S = S_i|A, V) \propto r_i^S \tag{15}$$

$$P(A|S = S_i) \propto r_i^A \tag{16}$$

$$P(V|S = S_i) \propto r_i^V \tag{17}$$

$$P(S = S_i) \propto r_i^{\text{prior}}. \tag{18}$$

Since Bayes' rule stipulates multiplication of probabilities, the network implementation of the probability code implies that the operation performed by neural circuits must also be a multiplication:

$$r_i^S = r_i^A r_i^V r_i^{\text{prior}}. \tag{19}$$

In contrast, for a log probability code, neural activity in neuron i is proportional to the log likelihood or log probability:

$$\log P(S = S_i|A, V) \propto r_i^S \tag{20}$$

$$\log P(A|S = S_i) \propto r_i^A \tag{21}$$

$$\log P(V|S = S_i) \propto r_i^V \tag{22}$$

$$\log P(S = S_i) \propto r_i^{\text{prior}}. \tag{23}$$

As previously stated, the choice of code does not change the structure of the network: Populations representing the likelihoods in nodes A and B will drive the population pattern of activity in node S. However, the choice of code does affect the neural operations that these circuits must perform. In the case of the log probability code, the circuit must perform a sum instead of a multiplication:

$$r_i^S = r_i^A + r_i^V + r_i^{\text{prior}}. \tag{24}$$

It is worth nothing that this process could have been inverted: We could have started out by assuming that the evidence integration or cue combination operation is implemented by neurons that linearly combine their inputs, as has been observed in multisensory tasks such as [56] and most famously in sequential evidence integration tasks [57, 4], and then asked what neural code for probability distributions is consistent with that empirical observation. We would then have concluded that neurons use a log probability code.

4.2 Sampling vs. parametric codes

The two neural representations probability distributions described above are simplified versions of the two competing hypothesis for neural mechanisms of probabilistic inference—namely, sampling-based (Monte Carlo) and parametric-based (variational) inference. Sampling schemes typically assume that individual neurons can be labeled by the latent variable that that neuron represents. For binary random variables, the spikes are often assumed to represent that random variable taking on a value of 1, as in a Boltzmann machine [58, 59]. When dealing with continuous random variables, it has been proposed that fluctuations in the underlying firing rate or membrane potential represent samples [60, 61]. In much of the sampling literature, the specific details of the mechanisms by which samples are generated are not investigated, and authors simply assume that the mechanism exists and compare predictions from a particular sampling algorithm with observed neural responses.

There are, however, two notable exceptions. Buesing et al. [7] proposed a mapping between spikes and samples that allows for the discrete nature of MCMC sampling to be implemented by continuous time spiking dynamics of neurons. This was accomplished by setting a spike to be an indicator that a particular binary latent variable took on the value of 1 at time t only if it occurred in the time window $[t - \tau, t]$. By introducing an additional latent variable (time since the last spike) for each neuron they were able to show that this continuous time stochastic dynamical system is capable of implementing MCMC sampling. This approach was generalized to multinomial latent variables by Pecevski et al. [10]. Similarly, Savin and Deneve [62] mapped the naturally continuous time dynamics of Langevin sampling onto a network of spiking neurons, using their previously published method for reliably instantiating a continuous-time dynamical system with spiking neurons [63]. Both or these approaches are quite appealing in their generality; they can be used to approximate complex multivariate posteriors without assuming a parametric form of the posterior (see below). Moreover, sampling-based schemes offer a natural explanation for neural variability. However, there is currently no concrete proposal for the source of the precisely tuned noise which must be added to neural dynamics in order to generate samples.

In contrast to sampling-based methods, parametric methods treat neural noise as a nuisance that is effectively eliminated by averaging over large populations of neurons jointly representing posterior marginals. For example, Rao [64] proposed a neural implementation of the sum-product message passing algorithm implemented in the log probability domain. He used an approximate expression for the resulting log of a sum of exponentials to generate linear rate equations for approximate inference. Beck et al. [65] proposed that neural activity is linearly related to the natural parameters of posterior distributions. This is a generalization of the log probability code discussed above, as it assumes that posterior marginals have an exponential family form:

$$P(S|\eta) = \exp\{\eta \cdot T(s) - A(\eta)\}, \tag{25}$$

where $T(s)$ are the sufficient statistics of the distribution and $A(\eta)$ is a normalizing constant. The vector of natural parameters η is assumed to be linearly related to the firing rates of the neurons that represent the posterior over S. In addition to being consistent with neural recordings, it was shown by Beck et al. [65] that simple probabilistic computations such as those involved coordinate transformation, auditory localization, object tracking (Kalman filtering), and cue combination can all be implemented using physiologically observed circuit level operation such as linear combination, coincidence detection, and divisive normalization. These circuit-level operations are typically derived by determining update rules for the natural parameters of a particular Bayesian computation.

In the same vein, Beck et al. [6] showed that when the posterior over multiple variables is approximated in a factorized form (i.e., a mean-field approximation), where each factor is in the exponential family, then variational inference algorithms can be implemented by similar circuit-level mechanisms. When applied to the problem of demixing odors, Beck et al. [6] demonstrated that the update equations for learning the synaptic connections specifying each learned odor have a simple Hebbian form. For the purposes of this chapter, the important insight offered by Beck et al. lies in the fact that complex multivariate posteriors can be approximated using the same machinery as simple univariate posteriors by constructing a network whose structure mirrors the factorization of the approximate posterior.

5 Conclusions and open questions

While Bayesian ideas have a long history in cognitive science, theoretical accounts are only beginning to grapple with the computational complexities of their implementation [39]. Nonetheless, some progress has been made, drawing heavily on advances in statistical machine learning. In particular, we have shown how two influential ideas about approximate inference (sampling and variational methods) have furnished plausible psychological hypotheses. Computational neuroscientists have been following a parallel path, exploring the biological implementation of sampling and variational methods, but so far making relatively little contact with the psychological literature. We see this is as the major frontier in the next generation of models.

Several open questions loom large. First, can the psychological manifestations of approximate inference (e.g., multistability, response variability, order effects) be connected to the neural manifestations (e.g., spiking stochasticity, membrane fluctuations, network dynamics)? For example, it is currently unknown whether variability in spiking activity is causally related to posterior probability matching [33, 38]. Second, does the brain contain a menagerie of approximation schemes, or is there a "master algorithm" (e.g., a canonical microcircuit; [66]) that applies universally? If the former, do different brain areas implement different approximations, or does the same area implement different approximations under different circumstances? One possibility is that the brain is designed to flexibly exploit the strengths and weaknesses of different approximations. For example, online approximations like particle filtering are well-suited to dynamical problems like object tracking, which is why some authors have proposed that such algorithms are used to make inferences about dynamic stimuli [67, 68], whereas algorithms with internal dynamics like belief propagation [64, 8, 9] and MCMC [43, 7, 62, 63] are better-suited to inference problems with complex static structure, like parsing a visual image. Finally, there has been renewed interest in "amortized inference" schemes that use a single inference network to approximate multiple posteriors [69, 70, 71]; while there is some psychological evidence for this kind of approximation scheme [72], it is currently unknown how amortization might be realized in a biologically plausible neural circuit (see [73] for some clues).

Acknowledgments

This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

References

[1] Pouget A, Beck JM, Ma WJ, Latham PE. Probabilistic brains: knowns and unknowns. Nature Neuroscience. 2013;16:1170–1178.

[2] Jazayeri M, Movshon JA. Optimal representation of sensory information by neural populations. Nature Neuroscience. 2006;9:690–696.

[3] Ma WJ, Beck JM, Latham PE, Pouget A. Bayesian inference with probabilistic population codes. Nature Neuroscience. 2006;9:1432–1438.

[4] Yang T, Shadlen MN. Probabilistic reasoning by neurons. Nature. 2007;447:1075–1080.

[5] Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND. How to grow a mind: Statistics, structure, and abstraction. Science. 2011;331:1279–1285.

[6] Beck JM, Pouget A, Heller KA. Complex inference in neural circuits with probabilistic population codes and topic models. In: Advances in Neural Information Processing Systems; 2012. p. 3059–3067.

[7] Buesing L, Bill J, Nessler B, Maass W. Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. PLoS Computational Biology. 2011;7:e1002211.

[8] George D, Hawkins J. Towards a mathematical theory of cortical micro-circuits. PLoS Computational Biology. 2009;5:e1000532.

[9] Litvak S, Ullman S. Cortical circuitry implementing graphical models. Neural computation. 2009;21:3010–3056.

[10] Pecevski D, Buesing L, Maass W. Probabilistic inference in general graphical models through sampling in stochastic networks of spiking neurons. PLoS Computational Biology. 2011;7:e1002294.

[11] Robert C, Casella G. Monte Carlo Statistical Methods. Springer Science & Business Media; 2010.

[12] Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. An introduction to variational methods for graphical models. Machine Learning. 1999;37:183–233.

[13] Sanborn AN. Types of approximation for probabilistic cognition: Sampling and variational. Brain and Cognition. 2015;.

[14] Kersten D, Mamassian P, Yuille A. Object perception as Bayesian inference. Annual Review of Psychology. 2004;55:271–304.

[15] Knill DC, Richards W. Perception as Bayesian Inference. Cambridge University Press; 1996.

[16] Macmillan NA, Creelman CD. Detection Theory: A User's Guide. Psychology Press; 2004.

[17] Holyoak KJ, Cheng PW. Causal learning and inference as a rational process: The new synthesis. Annual Review of Psychology. 2011;62:135–163.

[18] Griffiths TL, Steyvers M, Tenenbaum JB. Topics in semantic representation. Psychological Review. 2007;114:211–244.

[19] Chater N, Manning CD. Probabilistic models of language processing and acquisition. Trends in Cognitive Sciences. 2006;10:335–344.

[20] Goodman ND, Tenenbaum JB, Feldman J, Griffiths TL. A Rational Analysis of Rule-Based Concept Learning. Cognitive Science. 2008;32:108–154.

[21] Shepard RN. Toward a universal law of generalization for psychological science. Science. 1987;237:1317–1323.

[22] Tenenbaum JB, Griffiths TL. Generalization, similarity, and Bayesian inference. Behavioral and Brain sciences. 2001;24:629–640.

[23] Isard M, Blake A. Condensationconditional density propagation for visual tracking. International Journal of Computer Vision. 1998;29:5–28.

[24] Thrun S, Fox D, Burgard W, Dellaert F. Robust Monte Carlo localization for mobile robots. Artificial Intelligence. 2001;128:99–141.

[25] Chib S, Greenberg E. Understanding the Metropolis-Hastings algorithm. The American Statistician. 1995;49:327–335.

[26] Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science. 1983;220:671–680.

[27] Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 1984;6:721–741.

[28] Andrieu C, Doucet A, Holenstein R. Particle Markov chain Monte Carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2010;72:269–342.

[29] Minka TP. Expectation propagation for approximate Bayesian inference. In: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc.; 2001. p. 362–369.

[30] Hoffman MD, Blei DM, Wang C, Paisley J. Stochastic variational inference. The Journal of Machine Learning Research. 2013;14:1303–1347.

[31] Salimans T, Kingma D, Welling M. Markov chain Monte Carlo and variational inference: Bridging the gap. Proceedings of the 32nd International Conference on Machine Learning. 2015;.

[32] Vulkan N. An economists perspective on probability matching. Journal of Economic Surveys. 2000;14:101–118.

[33] Wozny DR, Beierholm UR, Shams L. Probability matching as a computational strategy used in perception. PLoS Computational Biology. 2010;6:e1000871.

[34] Gifford AM, Cohen YE, Stocker AA. Characterizing the Impact of Category Uncertainty on Human Auditory Categorization Behavior. PLoS Computational Biology. 2014;10.

[35] Mamassian P, Landy MS. Observer biases in the 3D interpretation of line drawings. Vision Research. 1998;38:2817–2832.

[36] Murray RF, Patel K, Yee A. Posterior Probability Matching and Human Perceptual Decision Making. PLoS Computational Biology. 2015;11:e1004342.

[37] Acerbi L, Vijayakumar S, Wolpert D. On the Origins of Suboptimality in Human Probabilistic Inference. PLoS Computational Biology. 2014;10:e1003661.

[38] Denison S, Bonawitz E, Gopnik A, Griffiths TL. Rational variability in childrens causal inferences: The sampling hypothesis. Cognition. 2013;126:285–300.

[39] Gershman SJ, Horvitz EJ, Tenenbaum JB. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. Science. 2015;349:273–278.

[40] Vul E, Goodman N, Griffiths TL, Tenenbaum JB. One and done? Optimal decisions from very few samples. Cognitive Science. 2014;38:599–637.

[41] Leopold DA, Logothetis NK. Multistable phenomena: changing views in perception. Trends in Cognitive Sciences. 1999;3:254–264.

[42] Blake R. A primer on binocular rivalry, including current controversies. Brain and Mind. 2001;2:5–38.

[43] Gershman SJ, Vul E, Tenenbaum JB. Multistability and perceptual inference. Neural Computation. 2012;24:1–24.

[44] Wilson HR, Blake R, Lee SH. Dynamics of travelling waves in visual perception. Nature. 2001;412:907–910.

[45] Levelt WJ. On Binocular Rivalry. Institute for Percep- tion Rvo-Tno; 1965.

[46] Burke D, Alais D, Wenderoth P. Determinants of fusion of dichoptically presented orthogonal gratings. Perception. 1999;28:73–88.

[47] Knapen T, Kanai R, Brascamp J, van Boxtel J, van Ee R. Distance in feature space determines exclusivity in visual rivalry. Vision Research. 2007;47:3269–3275.

[48] Moreno-Bote R, Knill DC, Pouget A. Bayesian sampling in visual perception. Proceedings of the National Academy of Sciences. 2011;108:12491–12496.

[49] Beck JM, Ma WJ, Pitkow X, Latham PE, Pouget A. Not noisy, just wrong: the role of suboptimal inference in behavioral variability. Neuron. 2012;74:30–39.

[50] Vul E, Frank MC, Alvarez G, Tenenbaum JB. Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. In: Advances in Neural Information Processing Systems; 2009. p. 1955–1963.

[51] Brown SD, Steyvers M. Detecting and predicting changes. Cognitive Psychology. 2009;58:49–67.

[52] Sanborn AN, Griffiths TL, Navarro DJ. Rational approximations to rational models: alternative algorithms for category learning. Psychological Review. 2010;117:1144–1167.

[53] Frank MC, Goldwater S, Griffiths TL, Tenenbaum JB. Modeling human performance in statistical word segmentation. Cognition. 2010;117:107–125.

[54] Levy RP, Reali F, Griffiths TL. Modeling the effects of memory on human online sentence processing with particle filters. In: Advances in Neural Information Processing Systems; 2009. p. 937–944.

[55] Anderson CH. Basic elements of biological computational systems. International Journal of Modern Physics C. 1994;5:313–315.

[56] Gu Y, Angelaki DE, DeAngelis GC. Neural correlates of multisensory cue integration in macaque MSTd. Nature Neuroscience. 2008;11:1201–1210.

[57] Gold JI, Shadlen MN. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. Neuron. 2002;36:299–308.

[58] Ackley DH, Hinton GE, Sejnowski TJ. A learning algorithm for Boltzmann machines. Cognitive Science. 1985;9:147–169.

[59] Savin C, Dayan P, Lengyel M. Optimal recall from bounded metaplastic synapses: predicting functional adaptations in hippocampal area CA3. PLoS Computational Biology. 2014;10:e1003489.

[60] Haefner RM, Berkes P, Fiser J. Perceptual decision-making as probabilistic inference by neural sampling. arXiv preprint arXiv:14090257. 2014;.

[61] Hennequin G, Aitchison L, Lengyel M. Fast sampling-based inference in balanced neuronal networks. In: Advances in Neural Information Processing Systems; 2014. p. 2240–2248.

[62] Savin C, Deneve S. Spatio-temporal representations of uncertainty in spiking neural networks. In: Advances in Neural Information Processing Systems; 2014. p. 2024–2032.

[63] Boerlin M, Machens CK, Denève S. Predictive coding of dynamical variables in balanced spiking networks. PLoS Computational Biology. 2013;9:e1003258.

[64] Rao RP. Bayesian computation in recurrent neural circuits. Neural Computation. 2004;16:1–38.

[65] Beck JM, Latham PE, Pouget A. Marginalization in neural circuits with divisive normalization. The Journal of Neuroscience. 2011;31:15310–15319.

[66] Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ. Canonical microcircuits for predictive coding. Neuron. 2012;76:695–711.

[67] Huang Y, Rao RP. Neurons as Monte Carlo Samplers: Bayesian? Inference and Learning in Spiking Networks. In: Advances in Neural Information Processing Systems; 2014. p. 1943–1951.

[68] Legenstein R, Maass W. Ensembles of spiking neurons with noise support optimal probabilistic inference in a dynamically changing environment. PLoS Computational Biology. 2014;10:e1003859.

[69] Dayan P, Hinton GE, Neal RM, Zemel RS. The Helmholtz machine. Neural Computation. 1995;7:889–904.

[70] Rezende D, Mohamed S, Wierstra D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In: Proceedings of The 31st International Conference on Machine Learning; 2014. p. 1278–1286.

[71] Rezende D, Mohamed S. Variational Inference with Normalizing Flows. In: Proceedings of The 32nd International Conference on Machine Learning; 2015. p. 1530–1538.

[72] Gershman SJ, Goodman ND. Amortized inference in probabilistic reasoning. In: Proceedings of the 36th Annual Conference of the Cognitive Science Society; 2014. .

[73] Yildirim I, Kulkarni TD, Freiwald WA, Tenenbaum JB. Efficient analysis-by-synthesis in vision: A computational framework, behavioral tests, and comparison with neural representations. In: Proceedings of the Thirty-Seventh Annual Conference of the Cognitive Science Society; 2015. .