9

Reverse Engineering Bayesian Computations from Spike Trains

The preceding chapters have developed a range of probabilistic models for neural spike trains that leverage our intuitions about neural types, features, and states to inform structured prior distributions over the dynamics of neural activity. In this chapter, we take a first step toward reconciling these intuitive models with the host of theoretical models of neural computation. From a Bayesian perspective, theoretical models can be seen as prior distributions on activity — albeit highly sophisticated ones. By connecting theory to observation in a hierarchical probabilistic model, we provide the link necessary to test, evaluate, and revise our theories in a data-driven fashion. As an exercise in meta-reasoning, the theory we test with this Bayesian approach is that neural populations are themselves performing Bayesian inference.

This chapter is organized as follows. First, in Section 9.1 we briefly review the "Bayesian brain" hypothesis, the various lines of evidence supporting this hypothesis. While not strictly required by the Bayesian brain hypothesis, we review some of the hypothesized means by which neurons could represent probability distributions and carry out Bayesian calculations. Then, in Section 9.2 we present a distributed representation scheme, and in Section 9.3 we provide a novel analysis of its complexity in the spirit of Valiant (1994). This provides important constraints on biological plausibility of this scheme and the number of neurons we must observe in order to test this theory. Section 9.4 adapts existing theories to show how neural circuits could perform mean-field variational inference in a restricted class of graphical models given our representation scheme. A simple example of inference in a mixture model is illustrated in Section 9.5. Finally, in Section 9.6 we show that the dynamics of inference in this model are equivalent to a nonlinear autoregressive model with weights drawn from a stochastic block model, thus providing a "top-down," theoretical justification for the intuitive models developed in Chapter 5. With this insight, we show how simple probabilistic models can be reverse engineered from spike trains using Bayesian inference and a Bayesian theory of neural computation. Section 9.7 considers some important open problems and directions for future work.

The "Bayesian Brain" Hypothesis

Bayesian theories of neural computation address a fundamental question: how do organisms reason, act, and make decisions given only limited, noisy information about the world around them? Bayes' rule tells us how an optimal agent should combine noisy information with prior knowledge to make posterior inferences. That the brain may employ or approximate Bayesian methods is an idea that dates back as far as von Helmholtz and Southall (1925). At the cognitive level, Bayesian models have proven extraordinarily useful for understanding and explaining human and animal behavior (Tenenbaum et al., 2011; Griffiths et al., 2008). These cognitive models span a variety of domains, from lower level systems like visual perception (Knill and Richards, 1996; Brainard and Freeman, 1997; Weiss et al., 2002; Yuille and Kersten, 2006; Stocker and Simoncelli, 2006; Simoncelli, 2009) and motor control (Körding and Wolpert, 2004) to higher level systems of sensory integration (Ernst and Banks, 2002), time interval estimation (Jazayeri and Shadlen, 2010), language processing (Chater and Manning, 2006), attention (Whiteley and Sahani, 2012; Chikkerur et al., 2010; Dayan and Solomon, 2010), and learning (Tenenbaum et al., 2006; Courville et al., 2006) The success of these models in explaining behavior suggests that the brain may be performing, or at least approximating, Bayesian computations.

Further buttressing the Bayesian brain hypothesis, some experiments have shown Bayes-optimal behavior along with simultaneous neural responses that are strongly correlated with relevant probabilistic quantities. For example, Yang and Shadlen (2007) trained monkeys to make an eye movement to either the left or the right based on an observed set of shapes. Each shape contributed an additive “weight” to the log probability that reward would be given for leftward movements rather than rightward, so the optimal strategy (once the weights were learned) was to sum the weights, compute the log probability of left versus right, and choose the direction most likely to yield a reward. In effect, the monkeys had to perform inference in a simple mixture model. The monkeys learned to perform this task optimally, and Yang and Shadlen (2007) found that the firing rates of neurons in parietal cortex were proportional to the log likelihood ratio of left versus right. Subsequent work showed that when the paradigm was extended to allow the monkey to opt-out of making a decision and obtain a smaller but guaranteed reward, the monkey chose to opt out only when the probability of left versus right was below a threshold. This impliest that the brain has access to not only the most likely direction, but also its uncertainty (Kiani and Shadlen, 2009).

Further evidence of neural probabilistic inference has been found in other simple tasks. In a time interval reproduction task, non-human primates exhibited behavior consistent with a Bayesian model, and simultaneous recordings in parietal cortex found that some neurons encoded interval estimates that could support this behavior (Jazayeri and Shadlen, 2015). In another line of work, neural correlates of multisensory cue integration were found in macaque monkeys performing a heading discrimination task. These neurons combined both visual and vestibular inputs in a manner consistent with Bayesian theory (Gu et al., 2008; Morgan et al., 2008; Fetsch et al., 2009; 2012). While these experiments provide some compelling evidence in favor of a simple probabilistic computations in neural circuits, there is a large gap between these experiments and the rich array of cognitive phenomena surveyed above. To bridge this gap, we need a broader theory of Bayesian inference in neural circuits, and more powerful tools to link these theories to neural activity.

The past decade has witnessed a surge of interest in theoretical models of Bayesian inference with spiking neurons, and this work has been the subject of a number of recent surveys (Simoncelli, 2009; Fiser et al., 2010; Pouget et al., 2013; Ma and Jazayeri, 2014). These theories can be broadly characterized by their answers to three successive questions:

1. How are probabilities are represented, and how are the conditional probability distributions that constitute the probabilistic model encoded? Are these distributions represented in a parametric manner? How are the parameters instantiated in a neural system?

2. Given a representation, how do neural dynamics compute the desired posterior distribution? In other words, what is the algorithm of probabilistic inference, and how is it reified in a population of neurons? These dynamics must respect the natural constraints of neural systems, for example, that neural connectivity is sparse and that neurons have limited computational power.

3. Finally, how are the parameters of the probabilistic model learned, and how are new variables of interest incorporated into an existing model?

The simplest and most common answer to the first question is that neural firing rates are proportional to probability (Hinton and Sejnowski, 1983; Hinton, 1992; Anderson and Essen, 1994; Barber et al., 2003; Buesing et al., 2011; Berkes et al., 2011; Nessler et al., 2013; Legenstein and Maass, 2014) or some function of the probability, like its log (Rao, 2004; Beck and Pouget, 2007; Rao, 2007; Litvak and Ullman, 2009). Others have suggested that probability distributions are encoded implicitly by the stochasticity of neurons (Zemel et al., 1998; Sahani and Dayan, 2003; Ma et al., 2006). Still others have contended that neurons employ a predictive coding scheme to convey probabilistic information (Rao and Ballard, 1999; Deneve, 2008; Huang and Rao, 2011). According to the predictive coding hypothesis, neurons only communicate spikes when their internal state cannot otherwise be inferred by their downstream neighbors. Finally, an interesting variant of rate coding suggests that distributions may be implicitly encoded in the number of neurons representing a particular value or, similarly, in the width of neural tuning curves (Shi and Griffiths, 2009; Ganguli and Simoncelli, 2010).

Along with this host of representational hypotheses has come an equally broad set of proposed inference algorithms. In the simplest models, like mixture models with a single latent variable, inference can be performed exactly in a single step. For more complicated models, some dynamic and often approximate inference algorithm is necessary. Of these, belief propagation and related message passing algorithms (Rao, 2007; Litvak and Ullman, 2009), variational inference (Friston, 2010; Nessler et al., 2013), and sampling based methods like Markov chain Monte Carlo (MCMC) (Hoyer and Hyvarinen, 2003; Buesing et al., 2011; Berkes et al., 2011; Gershman et al., 2012b; Legenstein and Maass, 2014), Hamiltonian Monte Carlo (Aitchison and Lengyel, 2014), importance sampling (Shi and Griffiths, 2009), and particle filtering (Lee and Mumford, 2003) have all been suggested. This amazing diversity speaks to both the computational power of neural populations as well as the enormous challenge in winnowing the field of contending theories.

The question of learning has received less attention; indeed, many theories have ignored it completely. Those that have addressed it tend to equate learning with synaptic plasticity. In some cases, synaptic plasticity rules like spike-timing dependent plasticity can be seen as maximizing a lower bound on the marginal log likelihood (Friston, 2010; Nessler et al., 2013; Rezende et al., 2011). In others, the stochasticity of synapses (e.g. stochastic vesicle release) is seen as sampling from a distribution over weights (Aitchison and Latham, 2015; Kappel et al., 2015b;a; Tully et al., 2014). These theories address unsupervised learning of model parameters, but the larger question of learning model structure, via either supervised or unsupervised means, remains largely a mystery.

Given the breadth and depth of existing theories of neural inference, our intention here is not to present a radically novel theory. Instead, our focus is on how we may assess the viability of a theory of neural inference. To that end, we provide a detailed description of a distributed representation of probability, analyze its complexity, show how inference could be performed with this representation, and provide a simple example of how this representation could be reverse engineered from neural spike train recordings.

A Direct Distributed Representation of Probability Distributions

As described above, the simplest representation of probability is a *direct* representation in which neural firing rates reflect instantaneous probabilities. Assume a population of neurons is responsible for representing the distribution over values that a set of random variables may take on. We denote this set of variables by, $\boldsymbol{z} = \{z_1, \dots, z_J\}$. For simplicity, assume for now that these variables can only assume a discrete set of values, $\{1, \dots, K\}$.

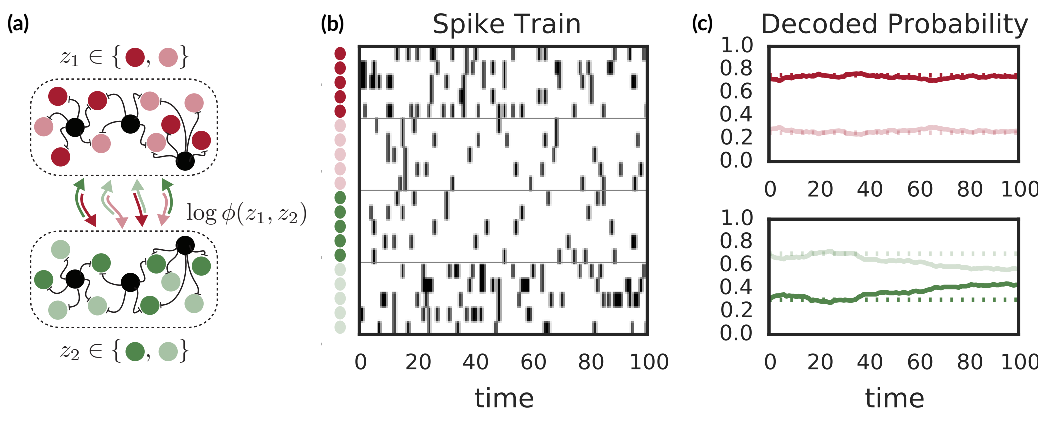

Figure 9.1: Example of a population of neurons encoding marginal probability distributions for two binary random variables, z_1 and z_2. Each variable is represented by a population of neurons, which is further divided into subpopulations for each value the variable can take on. In **(a)**, z_1 is represented by red neurons, with dark red neurons representing $z_1 = 1$ and light red representing $z_1 = 0$. Interneurons (black) provide normalization in the form of local inhibition. Excitatory connections between populations implement probabilistic inference. **(b)** The population spike train encodes the marginal distributions. For example, $\Pr(z_1 = 1)$ is proportional to the spike count of the subpopulation of dark red neurons. **(c)** These probabilities are decoded by integrating spike counts for each subpopulation over time and normalizing across subpopulations for each variable. Dotted lines: true probability.

Our neural population is thus tasked with representing probability vectors, $\boldsymbol{\pi}^{(j)}$, for each variable. The entries of these vectors are, $\pi_k^{(j)} = \Pr(z_j = k)$. In Section 9.7, we discuss how this representation could be extended to continuous probability density functions by assigning each neuron a basis function or tuning curve over the support of the distribution, as in Barber et al. (2003); Ma et al. (2006); Beck and Pouget (2007).

In a *distributed* representation, each variable-value pair, (z_j, k), is associated with a subpopulation of R neurons. This is inspired by Valiant (1994; 2005), and is similar to the ensemble based neural sampling code of Legenstein and Maass (2014). We further assume that these subpopulations are *non-overlapping* such that each neuron can be represented with at most one variable-value pair. Let $j_n \in \{1, \dots, J\}$ denote the index of the specific variable and $k_n \in \{1, \dots K\}$ denote the particular value that neuron n represents.

Now let $s_{t,n}$ denote the number of spikes fired by neuron n in the t-th time bin. The relative spike counts encode instantaneous probability distributions for each variable. If neurons representing a particular value fire twice as many spikes as neurons representing a competing value, then the first value is twice as likely. We introduce the notion of an *in-*

tegration time, T_I, over which spikes are counted to estimate the probability distribution. With this notation, the empirical distribution encoded by a population at a particular instant in time, $\widehat{\boldsymbol{\pi}}_t^{(j)}$, is defined by,

$$\widehat{\pi}_{t,k}^{(j)} = \frac{\sum_{\Delta=1}^{T_I} \sum_{n=1}^{N} \mathbb{I}[j_n = j, k_n = k]\, s_{t-\Delta,n}}{\sum_{\Delta=1}^{T_I} \sum_{n=1}^{N} \mathbb{I}[j_n = j]\, s_{t-\Delta,n}}.$$

The numerator counts spikes from neurons representing the particular value, $z_j = k$; the denominator counts all spikes from neurons representing z_j.

We assume these neurons are stochastic, each endowed with an instantaneous firing rate, $\lambda_{t,n}$, which gives rise to an instantaneous spike count, $s_{t,n}$ according to a Poisson distribution, $s_{t,n} \sim \mathrm{Poisson}(\lambda_{t,n})$. Thus, while the firing rate may encode one distribution, the distribution that is read out from a finite number of spikes will differ.

To summarize, the direct distributed representation entails the following assumption:

Assumption 1. *Neurons represent discrete probability distributions with a direct, distributed code. Each variable-value pair is allotted R neurons that emit spikes according to a Poisson distribution. Spikes are integrated over the subpopulation and over an integration time window, T_I, to obtain an unnormalized probability for the variable-value pair. By normalizing across subpopulations, they decode the probability distribution.*

Complexity of the Direct Distributed Representation

In pioneering work, Valiant (1994) suggested that the tools of computational learning theory, which provide rigorous computational limits on statistical learning, may apply equally well to learning in biological systems. All that is needed is a model of biological computation and a precisely stated learning algorithm.[*] However, if the resources required by this algorithm (e.g. number of neurons or spikes) grow exponentially quickly with the size of learning problem, then it is highly unlikely that the algorithm is employed by the brain. In other words, biology is bound by the same limits of computational tractability as our sili-

[*] Of course, the difficulty lies in specifying such a model and algorithm! While this is surely a challenge, there is no substitute for a formal declaration of assumptions.

con devices, but the natural measure of complexity may be spikes and synapses rather than FLOPs and bytes.

Valiant (1994) laid the groundwork for a model of neural computation that emphasizes a few key features: distributed representations, and sparse, random connectivity. In this and following work (Valiant, 2005; 2006), he showed that simple tasks, like learning an association between two variables and learning a linear classifier, could be performed efficiently within this model. Since the problem of approximate Bayesian inference is NP-hard in the worst case (Dagum and Luby, 1993; Roth, 1996), a provably tractable algorithm for neural inference is out of the question. Instead, we focus on the complexity of a prerequisite problem: simply representing a distribution with a population of stochastic neurons. If this is tractable, then we may consider the question of inference and the scenarios in which it may be possible.

Interestingly, the question of complexity has recently arisen in a somewhat different context. Gao and Ganguli (2015) have developed a theory of "task complexity" that predicts the dimensionality of neuronal dynamics as well as the number of neurons that must be measured in order to accurately recover the dynamics. In their case, the quantities of interest are the total number of neurons, number of observed neurons, number of stimuli, length of recording, and the unknown dimensionality of neural data. By fixing some of these parameters, they obtain bounds on the remaining parameters that govern when the dimensionality can be recovered. In our case, we derive similar results that govern the accuracy with which an encoded probability distribution can be recovered, either by a downstream population of neurons, or by an experimental observer.

In our case, we are concerned with the complexity, in terms of the number of neurons, time steps, or spikes, of stochastically encoding a distribution such that the decoded distribution is, with high confidence, within some tolerable error of the true distribution. The stochasticity of the spike counts implies that at any instant in time, the probability distribution that is represented by the population will be a random variable. First, we will show that this representation is unbiased. That is, if the firing rates, $\lambda_{t,n}$, are proportional to the probability, $\pi_{k_n}^{(j_n)}$, then the empirical probability distribution, $\widehat{\boldsymbol{\pi}}_t^{(j)}$, will equal $\boldsymbol{\pi}^{(j)}$ in expectation. Since we will be focusing on the representation of a single random variable z_j, we drop the superscripts $^{(j)}$ and $^{(j_n)}$ for the remainder of this section.

Lemma 2. *If the firing rates are proportional to a given probability distribution over the integration time window then the probability distribution represented by the population will have expectation equal to the given probability distribution. That is, if* $\lambda_{t,n} = \lambda_{\mathsf{max}} \pi_{k_n}$, *then* $\mathbb{E}[\widehat{\boldsymbol{\pi}}_t] = \boldsymbol{\pi}$.

Proof. Let,

$$S_{t,k} = \sum_{\Delta=1}^{T_I} \sum_{n=1}^{N} \mathbb{I}[j_n = j, k_n = k] \, s_{t-\Delta,n},$$

and

$$S_t = \sum_{\Delta=1}^{T_I} \sum_{n=1}^{N} \mathbb{I}[j_n = j] \, s_{t,n} = \sum_{k=1}^{K} S_{t,k}.$$

Iterating expectations, we have,

$$\mathbb{E}[\widehat{\boldsymbol{\pi}}_t] = \mathbb{E}\left[\frac{1}{S_t}\left(S_{t,1}, \dots, S_{t,K}\right)\right] = \mathbb{E}_{S_t}\left[\mathbb{E}_{\left(S_{t,1},\dots,S_{t,K}\right)}\left[\frac{1}{S_t}\left(S_{t,1}, \dots, S_{t,K}\right) \,\middle|\, S_t\right]\right].$$

Since $s_{t,n}$ are independent Poisson random variables, their partial sums are as well. Specifically,

$$\begin{aligned} S_{t,k} &\sim \text{Poisson}\left(\sum_{\Delta=1}^{T_I} \sum_{n} \mathbb{I}[j_n = j, k_n = k] \lambda_{t,n}\right) \\ &= \text{Poisson}\left(\lambda_{\mathsf{max}} T_I \sum_{n} \mathbb{I}[j_n = j, k_n = k] \, \pi_k\right) \\ &= \text{Poisson}(\lambda_{\mathsf{max}} \, T_I \, R \, \pi_k), \end{aligned} \tag{9.1}$$

which implies,

$$S_t = \sum_{k} S_{t,k} \sim \text{Poisson}(\lambda_{\mathsf{max}} \, T_I \, R).$$

Moreover, by the Poisson superposition principle, the vector $(S_{t,1}, \dots, S_{t,K})$ is multinomial distributed given S_t,

$$\begin{aligned}(S_{t,1}, \dots, S_{t,K}) \mid S_t &\sim \mathrm{Mult}\left(S_t, \left(\frac{\lambda_{\mathsf{max}}\, T_I\, R\, \pi_1}{\lambda_{\mathsf{max}}\, T_I\, R}, \dots, \frac{\lambda_{\mathsf{max}}\, T_I\, R\, \pi_K}{\lambda_{\mathsf{max}}\, T_I\, R}\right)\right) \\ &= \mathrm{Mult}(S_t, \boldsymbol{\pi}),\end{aligned}$$

with expectation $\boldsymbol{\pi} S_t$. Plugging this into the iterated expectation above,

$$\begin{aligned}\mathbb{E}[\widehat{\boldsymbol{\pi}}_t] &= \mathbb{E}_{S_t}\left[\mathbb{E}_{\left(S_{t,1},\dots,S_{t,K}\right)}\left[\frac{1}{S_t}\left(S_{t,1}, \dots, S_{t,K}\right) \,\middle|\, S_t\right]\right] \\ &= \mathbb{E}_{S_t}\left[\frac{\boldsymbol{\pi} S_t}{S_t}\right] \\ &= \boldsymbol{\pi}.\end{aligned}$$

Thus, this stochastic encoding is unbiased. □

While this stochastic representation may have the correct expectation, we would like to characterize the probability that it is "close" to its mean. As we hypothesized above, the difference between the true probability and that represented by the population should shrink as the number of spikes grows. We measure this difference with the ℓ_∞ norm of the difference between two probability vectors,

$$||\widehat{\boldsymbol{\pi}}_t - \boldsymbol{\pi}||_\infty = \max_k |\widehat{\pi}_{t,k} - \pi_k|.$$

While somewhat unorthodox, this metric is similar in spirit to the total variation distance. The following theorem provides an upper bound on the number of spikes required to guarantee that the represented probability differs from the true probability by more than ϵ.

Theorem 1. *Given a fixed probability vector $\boldsymbol{\pi}$, firing rates $\lambda_{t,n} = \lambda_{\mathsf{max}} \pi_{k_n}$ over the integration time window, a fixed error level $\epsilon < 1$, and a desired confidence $\delta < 1$, there exists a minimum number of spikes S^* such that if $S_t \geq S^*$, the conditional probability of error is bounded by $\Pr(||\widehat{\boldsymbol{\pi}}_t - \boldsymbol{\pi}||_\infty > \epsilon \mid S_t) < \delta$. Furthermore, this minimum number of spikes* *is at most,*

$$S^* \leq \frac{1}{2\epsilon^2} \ln \frac{2K}{\delta},$$

Proof. First, consider the probability that a particular entry differs from its mean by more than ϵ.

$$\begin{aligned}
&\Pr(|\widehat{\pi}_k - \pi_k| > \epsilon \,|\, S_t) \\
&\quad = \Pr(\widehat{\pi}_k - \pi_k > \epsilon \,|\, S_t) + \Pr(\widehat{\pi}_k - \pi_k < -\epsilon \,|\, S_t) \\
&\quad = \Pr\left(S_{t,k} > S_t \pi_k \left(1 + \frac{\epsilon}{\pi_k}\right) \,\middle|\, S_t \right) + \Pr\left(S_{t,k} < S_t \pi_k \left(1 - \frac{\epsilon}{\pi_k}\right) \,\middle|\, S_t \right) \\
&\quad = \Pr\left(S_{t,k} > \mathbb{E}[S_{t,k} \,|\, S_t] \left(1 + \frac{\epsilon}{\pi_k}\right) \right) + \Pr\left(S_{t,k} < \mathbb{E}[S_{t,k} \,|\, S_t] \left(1 - \frac{\epsilon}{\pi_k}\right) \right)
\end{aligned}$$

As in Lemma 2, we have used the fact that $S_{t,k} \,|\, S_t \sim \mathrm{Bin}(S_t, \pi_k)$ and hence has expectation $S_t \pi_k$. The probability of this binomial random variable exceeding its mean by a multiplicative constant is a decreasing function of the number of spikes, S_t. This implies that there exists a minimum number of trials S^* such that for $S_t \geq S^*$, this probability of error is bounded above by δ, hence proving the first part of the theorem.

Now suppose $S_t = S^*$. We use a Chernoff bound to upper bound the probability that the binomial random variable, $S_{t,k}$, deviates from its mean by more than a multiplicative factor. Leveraging the fact that $\pi_k \leq 1$, we have,

$$\begin{aligned}
\Pr\left(S_{t,k} > \mathbb{E}[S_{t,k} \,|\, S^*] \left(1 + \frac{\epsilon}{\pi_k}\right) \right) &\leq \exp\left\{-2S^* \epsilon^2\right\}, \\
\Pr\left(S_{t,k} < \mathbb{E}[S_{t,k} \,|\, S^*] \left(1 - \frac{\epsilon}{\pi_k}\right) \right) &\leq \exp\left\{-2S^* \epsilon^2\right\},
\end{aligned}$$

which together imply,

$$\Pr(|\widehat{\pi}_{t,k} - \pi_k| > \epsilon \,|\, S^*) \leq 2 \exp\left\{-2S^* \epsilon^2\right\}.$$

We bound the maximum deviation of any entry in $\widehat{\boldsymbol{\pi}}$ with a union bound,

$$\Pr(||\widehat{\boldsymbol{\pi}}_t - \boldsymbol{\pi}||_\infty > \epsilon \,|\, S^*) \leq 2K \exp\left\{-2S^*\epsilon^2\right\}.$$

Setting this probability equal to δ yields the desired bound on S^*,

$$S^* \leq \frac{1}{2\epsilon^2} \ln \frac{2K}{\delta}.$$

□

This theorem provides an upper bound on the minimum number of spikes necessary to guarantee that the ℓ_∞-distance between the true and estimated probability vectors is less than ϵ with probability $1 - \delta$. Notably, the relevant quantity is the number of spikes S_t, rather than the number of neurons. Thus, there is some flexibility in how the probability is estimated: a small population of neurons could be measured over many time bins, or a large population could be measured over a single time bin. Moreover, the population gain, λ_{max}, could be varied to adjust the number of spikes per time bin.

In practice, the number of spikes cannot be set directly. It, is a Poisson random variable whose mean, from Eq. 9.1, is $\mathbb{E}[S_t] = \lambda_{\mathsf{max}} T_I R$: the expected number of spikes per neuron times the number of neurons per outcome. This leads to the following theorem, which specifies a upper bound on the gain and number of neurons required to guarantee that the ℓ_∞-distance is less than ϵ with probability $1 - \delta$.

Theorem 2. *Given a fixed probability vector $\boldsymbol{\pi}$, firing rates $\lambda_{t,n} = \lambda_{\mathsf{max}} \pi_{k_n}$, a fixed error level $\epsilon < 1$, and a desired confidence $\delta < 1$, the probability of error is bounded by $\Pr(||\widehat{\boldsymbol{\pi}}_t - \boldsymbol{\pi}||_\infty > \epsilon) < \delta$ if $\lambda_{\mathsf{max}} T_I R \geq \mu^*$, where μ^* is at most,*

$$\mu^* \leq \frac{1}{1 - e^{-2\epsilon^2}} \ln \frac{2K}{\delta}.$$

Proof. We have,

$$\begin{aligned}
\Pr(||\widehat{\boldsymbol{\pi}}_t - \boldsymbol{\pi}||_\infty > \epsilon) &= \sum_{m=0}^{\infty} \Pr(S_t = m) \Pr(||\widehat{\boldsymbol{\pi}}_t - \boldsymbol{\pi}||_\infty > \epsilon \,|\, S_t = m) \\
&\leq \sum_{m=0}^{\infty} \Pr(S_t = m) \times 2K \exp\left\{-2m\epsilon^2\right\} \\
&= 2K\mathbb{E}_{S_t}\left[\exp\left\{-2S_t\epsilon^2\right\}\right] \\
&= 2K \exp\left\{\mu^*(e^{-2\epsilon^2} - 1)\right\},
\end{aligned}$$

where the last line follows from moment generating function of $S_t \sim \mathrm{Poisson}(\mu^*)$. Setting this equal to δ and solving for μ^* yields the stated bound. □

So far we have considered the estimated probability distribution obtained by "reading out" the entire population of neurons. What if we only observe a fraction of the population, as a neuron in a downstream population might? Assume each neuron in the population is "observed" with probability ρ. The expected number of observed neurons for a given variable-value pair is $R\rho$, and if we see exactly the expected number of neurons for each value (assume it is an integer), the estimated probability distribution will have the correct expectation. However, in practice we will incur some bias from seeing a different number of neurons for each value. Bounding the error theoretically is challenging due to this additional source of randomness, so we instead consider the simple case in which we see exactly $R\rho$ neurons for each value. Then, following the same logic as above, we have the following corollary.

Corollary 1. *Given a fixed probability vector $\boldsymbol{\pi}$, firing rates $\lambda_{t,n} = \lambda_{\mathsf{max}}\pi_{k_n}$ over the integration time window, a fixed error level $\epsilon < 1$, and a desired confidence $\delta < 1$, and ρR observed neurons for each of the K values, the probability of error is bounded by $\Pr(||\widehat{\boldsymbol{\pi}}_t - \boldsymbol{\pi}||_\infty > \epsilon) < \delta$ if*

$$(\lambda_{\mathsf{max}} T_I)(\rho R) \geq \mu^*,$$

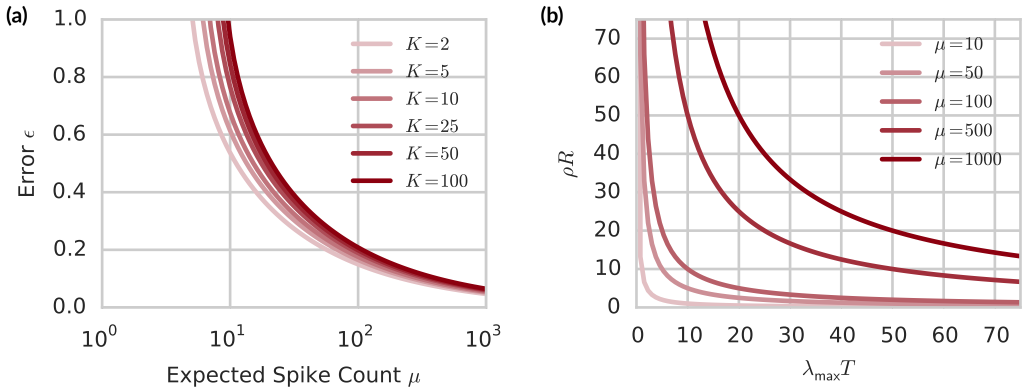

Figure 9.2: Theoretical relationship between between ℓ_∞-distance, expected spike count, and physiological parameters. **(a)** Theoretical upper bound on the 95th percentile of the ℓ_∞-distance, ϵ, as a function of the expected spike count, $\mu = (\rho R)(\lambda_{\text{max}} T_I)$, for increasing values of K. **(b)** The expected spike count is the product of the effective number of neurons, ρR, and the effective number of spikes per neuron, $\lambda_{\text{max}} T_I$. This shows how time and number of neurons can be balanced to obtain the desired expected spike count.

where μ^* is at most,

$$\mu^* \leq \frac{1}{1 - e^{-2\epsilon^2}} \ln \frac{2K}{\delta}.$$

Proof. This follows directly from Theorem 2 with ρR substituted for R. □

Corollary 1 provides theoretical connection between the fidelity of the representation, measured in terms of the error ϵ and confidence δ, for a given domain size K, maximum firing rate λ_{max}, integration time T_I, connection probability ρ, and representation size R. The expected spike count is the product of the effective number of neurons, ρR, and the expected number of spikes per neuron, $\lambda_{\text{max}} T_I$. Together, these allow us to deduce a manifold of trade-offs between population size and integration time that will achieve a desired error level and confidence.

Figure 9.2 plots these theoretical bounds. Fig. 9.2a shows the theoretical upper bound on the 95th percentile of the ℓ_∞-distance as a function of the expected spike count for a range of distribution sizes, K. Fig. 9.2b illustrates the trade-offs between effective number of neurons and expected number of spikes per neuron necessary to achieve a desired expected spike count.

Figure 9.3 shows the results of an empirical assessment of this theory under a vari-

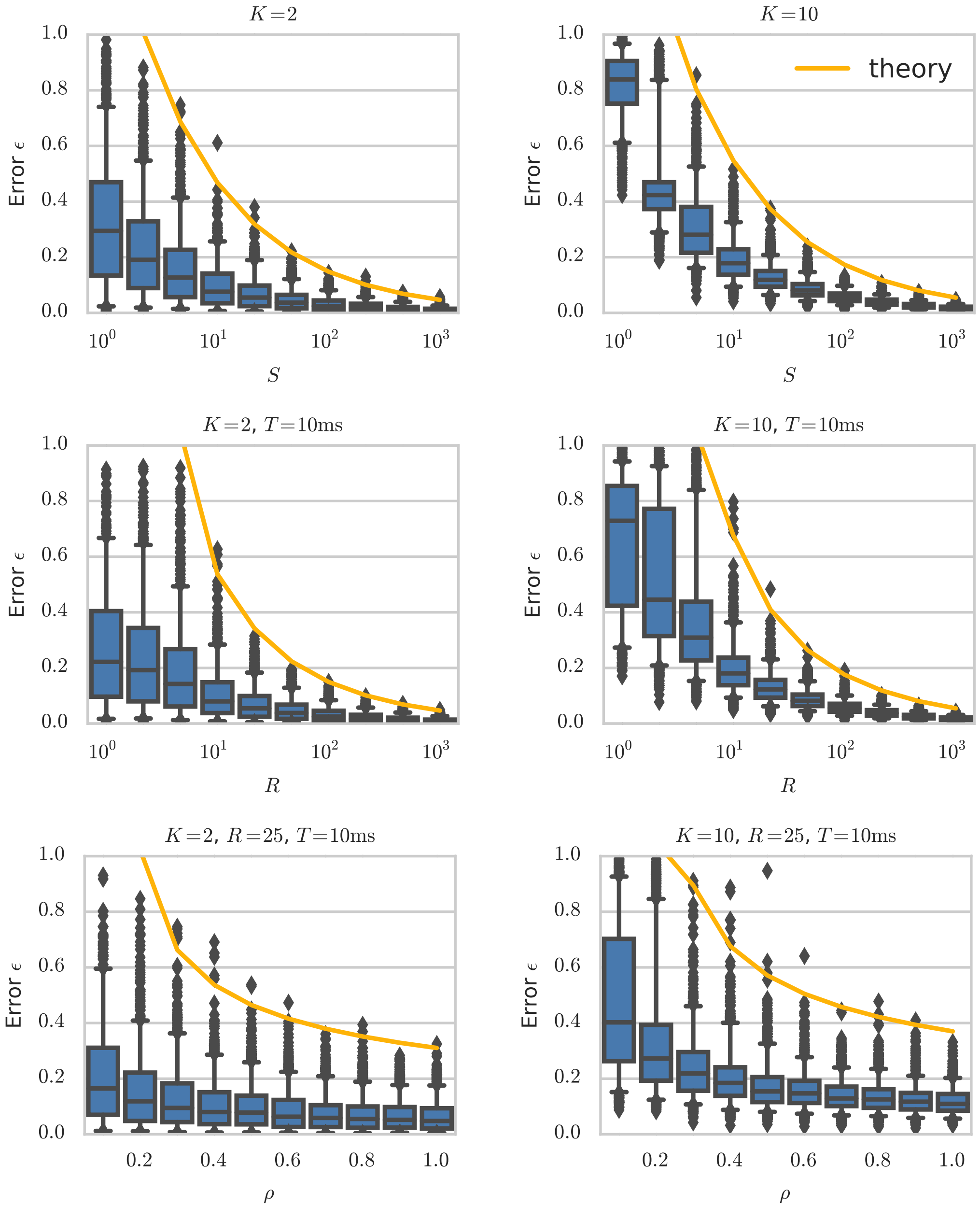

Figure 9.3: Empirical and theoretical ℓ_∞-distance under a variety of parameter regimes. Whiskers show the empirical 5th and 95th percentiles. Yellow line shows the theoretical upper bound on the 95th percentile. **Left** column: $K = 2$. **Right** column: $K = 10$. **Top** row: fixed population spike count, S. **Middle** row: varying the number of neurons per variable. **Bottom** row: varying the connection probability. See text for full description.

ety of parameter regimes. For each regime, we present results for discrete distributions over $K = 2$ values (left column) and $K = 10$ values (right column). We sample 1000 discrete distributions from a Dirichlet prior, $\boldsymbol{\pi} \sim \mathrm{Dir}(\mathbf{1}_K)$, and then we encode each true distribution with Poisson spiking neurons and measure the ℓ_∞-distance between the true and encoded distributions.

In the top row, we consider the case where the total population spike count is set explicitly, as in Theorem 1. In this case, the spikes are attributed to each value according to a multinomial distribution. We plot the theoretical bound from Theorem 1 in yellow.

In the middle row we measure the error as a function of the representation size, R, for $\rho = 1.0$, under the assumption that spikes are counted in one millisecond bins for $T_I = 10$ milliseconds, and a maximum firing rate of 100Hz (i.e. $\lambda_{\mathsf{max}} = 0.1$). Thus, in expectation the population will emit R spikes, allowing the top and middle rows to be directly compared. That is, in the top row, the population fires exactly the number of spikes expected in the middle row. We see that the stochastic population spike counts does indeed introduce extra variability in the error, as predicted by Theorem 2, though the median error is not substantially different from that of the fixed-S case.

Finally, the bottom row shows the empirical and predicted error as we vary the observation probability, ρ. Here, the representation size is fixed to $R = 25$, and the gain and integration time are set as in the middle row. While a strict upper bound is difficult to derive theoretically, the approximation from Corollary 1 provides a reasonable approximation for this parameter regime.

These complexity-theoretic bounds relate the number of spikes to the distance between the true and estimated distributions. From the number of spikes, we can deduce constraints on the representation size, integration time, and connection probability, for realistic gain levels. While the theoretical bounds do not exactly match the empirical error distributions, they appear to be roughly correct up to a multiplicative factor. Thus, if we can estimate some biophysical properties like connection probabilities, integration times, and firing rates, and we can constrain the error tolerances of the algorithms that consume these probability estimates, then we can estimate the number of neurons that must represent each variable-value pair. This will prove useful in guiding our bottom-up search, and serve as an important constraint for assessing the viability of a direct representation of probability.

Two forces cause the encoded distributions to change over time. As we interact with the world and receive new inputs, the probabilities of variables change to reflect the new observations. Moreover, even for a fixed set of observed variable assignments, the probabilities of latent variables will change as we perform inference. We show how a simple, iterative inference algorithm can be implemented with biologically plausible neural dynamics.

We assume that as an organism receives new inputs from the world, it updates its posterior distribution over the values of latent variables. Doing so requires a probabilistic model that relates hidden and observed variables via a joint probability distribution. Whereas in previous chapters we have considered directed graphical models, here we assume that the probabilistic models implemented in the brain are best described in terms of a *factor graph*,

$$p(\boldsymbol{z} \,|\, \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \prod_{j \in \mathcal{G}} \phi(z_j \,|\, \boldsymbol{\theta}) \prod_{i,j \in \mathcal{G}} \phi(z_i, z_j \,|\, \boldsymbol{\theta}) \tag{9.2}$$

The graph, $\mathcal{G}$, specifies unary and pairwise probabilistic dependencies between variables. Each unary factor, $\phi(\cdot \,|\, \boldsymbol{\theta})$ is a function that maps a variable assignment to a nonnegative real number, and each pairwise factor, $\phi(\cdot, \cdot \,|\, \boldsymbol{\theta})$, is a function that maps a pair of assignments, say $(z_i = k, z_j = k')$ to a nonnegative real number. The normalizing constant, $Z(\boldsymbol{\theta})$, ensures that the joint probability distribution sums to one. The probabilistic model in Eq. 9.2 reflects a specific assumption about the types of dependency structures neural populations can represent.

Assumption 2. *Neural populations perform inference in probabilistic models that factor into the product of unary and pairwise dependencies.*

A general probabilistic model need not factor into pairwise terms. It may instead have factors that relate three or more latent variables. As we will see, unary and pairwise factors map naturally onto neural biases and synaptic weights. In our proposed neural implementation, higher order factors would require the interaction of three or more neurons. While this may be realized with dendritic computation or interneurons, these more sophisticated implementations are beyond the scope of this chapter.

In general, the posterior distribution of a subset of hidden variables $\boldsymbol{z}_H \subseteq \boldsymbol{z}$ given the observed variables $\boldsymbol{z}_O = \boldsymbol{z} \setminus \boldsymbol{z}_H$ is,

$$p(\boldsymbol{z}_H \,|\, \boldsymbol{z}_O, \boldsymbol{\theta}) = \frac{p(\boldsymbol{z}_H, \boldsymbol{z}_O \,|\, \boldsymbol{\theta})}{\sum_{\boldsymbol{z}_H} p(\boldsymbol{z}_H, \boldsymbol{z}_O \,|\, \boldsymbol{\theta})}.$$

Typically, this cannot be efficiently computed since it requires a sum over all possible hidden variable assignments. However, as we have seen in previous chapters, there are many methods of approximating posterior distributions. Mean field variational inference maps particularly nicely onto the natural constraints of neural dynamics. In mean field variational inference, the intractable exact posterior distribution is approximated with a tractable, factorized distribution,

$$p(\boldsymbol{z}_H \,|\, \boldsymbol{z}_O, \boldsymbol{\theta}) \approx q(\boldsymbol{z}_H) \equiv \prod_{z_j \in \boldsymbol{z}_H} q(z_j).$$

The terms in this product are called *variational factors*. We solve for the variational factors that minimize KL-divergence between the true and approximate posterior, $\mathrm{KL}(q(\boldsymbol{z}_H) \,||\, p(\boldsymbol{z}_H \,|\, \boldsymbol{z}_O, \boldsymbol{\theta}))$. In minimizing the KL-divergence, we simultaneously maximize a lower bound on the log marginal likelihood, $\log p(\boldsymbol{z}_O \,|\, \boldsymbol{\theta})$.

The simplest method of minimizing this objective is via coordinate descent, iteratively updating the probability of one hidden variable given the probabilities of the rest. Since our variational distribution is factorized, the variational factor for variable z_j must satisfy the mean field consistency equation:

$$\log q(z_j) \simeq \mathbb{E}_{q(\boldsymbol{z}_{\neg j})} \left[\log p(\boldsymbol{z}_H, \boldsymbol{z}_O \,|\, \boldsymbol{\theta})\right], \tag{9.3}$$

where $\simeq$ denotes equality up to an additive constant and the expectations are taken with respect to the variational distribution over other hidden variables,

$$q(\boldsymbol{z}_{\neg j}) = \prod_{i \neq j} q(z_i).$$

The additive constant ensures normalization of the probabilities, and will be discussed subsequently.

For discrete random variables, the variational factors are simply vectors specifying the posterior probability of each variable, z_j. Under the direct representation described above, the instantaneous values of these factors are encoded in the relative spike counts of populations of neurons,

$$q_t(z_j = k) = \widehat{\pi}_{t,k}^{(j)}$$

To perform inference, the neuronal dynamics must be such that at each time step, the relative spike counts satisfy Eq. 9.3. Explicitly writing the additive constant, $-\log \nu_t^{(j)}$, we have,

$$\begin{aligned}
\log \widehat{\pi}_{t,k}^{(j)} &= -\log \nu_t^{(j)} + \log \phi(z_j = k \,|\, \boldsymbol{\theta}) \\
&\qquad + \mathbb{E}_{q_{t-1}(\boldsymbol{z}_{\neg j})}\Bigg[\sum_{i \in \mathsf{ne}(j)} \log \phi(z_i, z_j = k \,|\, \boldsymbol{\theta})\Bigg] \\
&= -\log \nu_t^{(j)} + \log \phi(z_j = k \,|\, \boldsymbol{\theta}) \\
&\qquad + \sum_{i \in \mathsf{ne}(j)} \sum_{k'=1}^{K} \left[\log \phi(z_i = k', z_j = k \,|\, \boldsymbol{\theta}) \cdot \widehat{\pi}_{t-1,k'}^{(i)}\right] \qquad (9.4) \\
&= -\log \nu_t^{(j)} + \psi_{t,k}^{(j)},
\end{aligned}$$

where

$$\psi_{t,k}^{(j)} = \log \phi(z_j = k \,|\, \boldsymbol{\theta}) + \sum_{i \in \mathsf{ne}(j)} \sum_{k'=1}^{K} \log \phi(z_i = k', z_j = k \,|\, \boldsymbol{\theta}) \cdot \widehat{\pi}_{t-1,k'}^{(i)}.$$

Since $\widehat{\boldsymbol{\pi}}_t^{(j)}$ is a probability distribution, the additive constant must be set to it is normalized. Thus,

$$\begin{aligned}
\widehat{\pi}_{t,k}^{(j)} &= \exp\left\{\psi_{t,k}^{(j)} - \log \nu_t^{(j)}\right\} \\
\implies \nu_t^{(j)} &= \sum_{k'} \exp\left\{\psi_{t,k'}^{(j)}\right\}.
\end{aligned}$$

Now that we have derived theoretically exact mean field updates, we must show how they can be approximated with plausible neural dynamics. We assume that inference occurs on a characteristic time scale of T_I time steps. This reflects the window of time over which neurons estimate probability distributions. From Lemma 2, we know that if the firing rates of neurons the variable-value pair (z_j, k) are proportional to $\widehat{\pi}_{t,k}^{(j)}$, then in expectation, the empirical distribution represented by the spike counts will be equal to the desired distribution. Thus, we aim to set,

$$\lambda_{t,n} = \lambda_{\mathsf{max}} \, \widehat{\pi}_{t,k_n}^{(j_n)} = \lambda_{\mathsf{max}} \exp\left\{ \psi_{t,k_n}^{(j_n)} - \log \nu_t^{(j_n)} \right\},$$

for maximum firing rate, λ_{max}. While this rate function is nearly a linear-nonlinear cascade, as we studied in previous chapters, there is one major impediment to realizing this calculation in biological neurons. Specifically, to compute the activation, a neuron must have access to the *normalized* probabilities of other hidden and visible variables. In practice, a neuron only observes the spike counts of the neurons it receives input from. However, these can be used to estimate the desired probabilities. This motivates our next assumptions,

Assumption 3. *Neurons are sparsely connected to one another. For each ordered pair of neurons, (m, n), the variable $a_{m\to n} \in \{0, 1\}$ indicates whether or not there exists a synaptic connection from neuron m to neuron n. These connections are modeled as independent and identically distributed Bernoulli random variables,*

$$a_{m\to n} \sim \mathrm{Bern}(\rho).$$

We combine these variables into a binary adjacency matrix, $\boldsymbol{A} \in \{0, 1\}^{N\times N}$.

Assumption 4. *All neurons in the population share the same gain, λ_{max}. Thus, neuron n's* *estimate of $\boldsymbol{\pi}^{(j)}$ is informed by $(\lambda_{\mathsf{max}} T_I)(\rho R)$ spikes, in expectation:*

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{A},\boldsymbol{s}}\left[\sum_{\Delta=1}^{T_I}\sum_{m=1}^{N}\mathbb{I}[j_m = j]\, a_{m\to n}\cdot s_{t-\Delta,m}\right] \\
&\qquad = \mathbb{E}_{\boldsymbol{A},\boldsymbol{s}}\left[\sum_{\Delta=1}^{T_I}\sum_{m=1}^{N}\sum_{k=1}^{K}\mathbb{I}[j_m = j, k_m = k]\, a_{m\to n}\cdot s_{t,m}\right] \\
&\qquad = \lambda_{\mathsf{max}} T_I \rho \sum_{m=1}^{N}\sum_{k=1}^{K}\mathbb{I}[j_m = j, k_m = k]\, \widehat{\pi}_{t,k}^{(j)} \\
&\qquad = (\lambda_{\mathsf{max}} T_I)(\rho R).
\end{aligned}$$

Moreover, the instantaneous probability is well-approximated by,

$$\begin{aligned}
\widehat{\pi}_{t,k}^{(j)} &= \frac{\sum_{\Delta=1}^{T_I}\sum_{m=1}^{N}\mathbb{I}[j_m = j, k_m = k]\, a_{m\to n}\cdot s_{t-\Delta,m}}{\sum_{\Delta=1}^{T_I}\sum_{m=1}^{N}\mathbb{I}[j_m = j]\, a_{m\to n}\cdot s_{t-\Delta,m}} \\
&\approx (\lambda_{\mathsf{max}} T_I \rho R)^{-1}\sum_{\Delta=1}^{T_I}\sum_{m=1}^{N}\mathbb{I}[j_m = j, k_m = k]\, a_{m\to n}\cdot s_{t-\Delta,m}.
\end{aligned}$$

In other words, the total spike count is concentrated around its mean.

Under the assumption of shared gain, the desired dynamics in Eq. 9.4 simplify to,

$$\lambda_{t,n} = \lambda_{\mathsf{max}} \exp\left\{ b_n + (\lambda_{\mathsf{max}} T_I \rho R)^{-1}\sum_{\Delta=1}^{T_I}\sum_{m=1}^{N} a_{m\to n}\cdot w_{m\to n}\cdot s_{t-\Delta,m} - \log \nu_t^{(j_n)} \right\} \tag{9.5}$$

where

$$b_n = \log \phi(z_{j_n} = k_n \,|\, \boldsymbol{\theta}),$$

and

$$w_{m \to n} = \begin{cases} \log \phi(z_{j_n} = k_n, z_{j_m} = k_m \,|\, \boldsymbol{\theta}) & \text{if } j_m \in \mathsf{ne}(j_n) \\ 0 & \text{o.w.} \end{cases}$$

Thus, the theory provides a normative interpretation of synaptic weights: they reflect the conditional log probabilities for the variable-value pairs represented by the pre- and post-synaptic neurons.

The last step is to compute the normalizing input, $\nu_t^{(j)}$. This requires summing the instantaneous rates of all neurons representing the random variable, z_j. While this is clearly implausible, we may derive a gain controller from an alternative perspective. Normalizing the probability distribution ensures that the expected spike count at any time step for neurons representing z_j is equal to $\lambda_{\mathsf{max}} R$. If the distribution is not properly normalized, the expected spike count will deviate. Thus, a reasonable gain controller can estimate the population can estimate the population rate,

$$\widehat{\lambda}_t^{(j)} = \sum_{\Delta=1}^{T_G} \sum_{n=1}^{N} \mathbb{I}[j_n = j] s_{t-\Delta,n},$$

and set the control input to,

$$\nu_t^{(j)} = \frac{\widehat{\lambda}_t^{(j)}}{\lambda_{\mathsf{max}} T_G R}.$$

The time scale of the gain controller is typically less than the time scale of inference, that is $T_G < T_I$.

Having shown that variational inference is theoretically plausible, we consider a simple example.

Example of a Simple Mixture Model

Consider a simple mixture model with a single latent variable denoting the mixture component, $z \in \{1, \ldots, K\}$, and a set of conditionally independent observations, $\{x_j\}_{j=1}^{J}$.

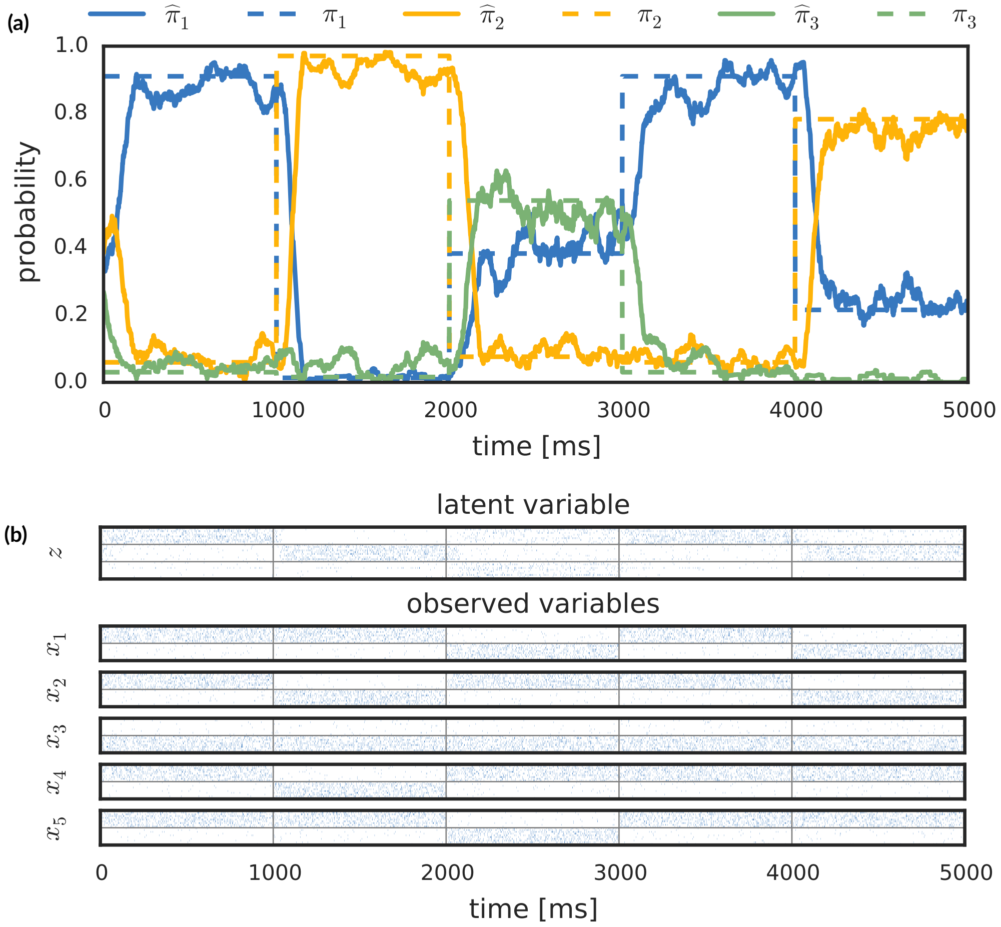

Figure 9.4: Example of neural inference a simple mixture model with one latent variable, $z \in \{1, 2, 3\}$, indicating which of the three mixture components gave rise to the data. The observations consist of 5 conditionally independent binary variables, $x_1, \ldots, x_5$, whose values change every second. **(a)** The empirical probabilities (solid lines) decoded from the spike train, and the true posterior (dashed line). Stochasticity arises from noisy inputs and Poisson spike counts. **(b)** The underlying spike trains of the neurons representing z and x_i. Horizontal gray lines distinguish subpopulations of $R = 10$ neurons for each value; vertical lines denote times of stimulus change.

In this example, we let the observations be Bernoulli random variables. The model is parameterized by a marginal class probability vector, $\boldsymbol{\alpha}$, which we assume is uniform, and class-conditional probabilities for each observation, $p_{j,k} = \Pr(x_j = 1 \mid z = k)$. Together, these specify the probabilistic model,

$$
\begin{aligned}
\boldsymbol{\alpha} &= \tfrac{1}{K}\mathbf{1}, & z &\sim \mathrm{Discrete}(\boldsymbol{\alpha}), \\
p_{j,k} &\sim \mathrm{Beta}(\tfrac{1}{2}, \tfrac{1}{2}), & x_j &\sim \mathrm{Bern}(p_{j,z}).
\end{aligned}
$$

This corresponds to a factor graph with,

$$
\begin{aligned}
\phi(z = k) &= \alpha_k, \\
\phi(x_j = 1, z = k) &= p_{j,k}, \\
\phi(x_j = 0, z = k) &= 1 - p_{j,k}.
\end{aligned}
$$

We simulate inference in this model with a population of neurons. Each variable-value pair is represented by $R = 10$ neurons, and each pair of neurons is connected with probability $\rho = 0.5$. The maximum firing rate is set to $\lambda_{\mathsf{max}} = 100\mathrm{Hz}$, and the integration time window is set to $T_I = 100\mathrm{ms}$. Hence, the expected spike count used to estimate probabilities is $(\rho R)(\lambda_{\mathsf{max}} T_I) = 50$ spikes. The neurons representing the observed variable assignment $x_j = k$ are externally driven at a rate of $0.95\lambda_{\mathsf{max}}$ if $x_j = k$, and a rate of $0.05\lambda_{\mathsf{max}}$ if $x_j \neq k$. We assume that the synaptic weights have already been learned and reflect the exact log of the pairwise factors.

Figure 9.4 illustrates inference dynamics for a changing stimulus. Every second, the observed variables are driven with a new pattern, which leads to a new posterior distribution over the latent variable. With each change in input, the neurons representing the latent variable adjust their firing rates to reflect this new probability. Fig. 9.4a shows the decoded probabilities over time (solid lines) along with the true posterior (dashed lines). Despite many sources of stochasticity, the decoded probabilities do converge to the correct posterior values. The $100\mathrm{ms}$ integration time is reflected in the delayed convergence upon each change in stimulus. Fig. 9.4b shows the spike trains from which these probabilities were decoded. The neurons are ordered according to the value they encode and the subpopulations of R neurons are separated by horizontal light gray lines. Vertical lines indicate changes in input. Overall, the neurons fire at between 30 and $40\mathrm{Hz}$, with a dynamic range of about 0 to $100\mathrm{Hz}$, as expected.

Reverse Engineering the Probabilistic Model from Spike Trains

Given this "top-down" theory of neural computation, can we reverse engineer the probabilistic model from neural recordings? To do so, we need to infer the subpopulations of neurons that encode each variable-value pair, as well as the characteristic weights that connect each subpopulation. We show that this is possible using the generalized linear models and structured network priors described in Chapter 5.

Recall the theoretical dynamics proposed in Section 9.4, reproduced here in slightly simplified form,

$$\lambda_{t,n} = \lambda_{\mathsf{max}} \exp\left\{ b_n + \gamma \sum_{\Delta=1}^{T_I} \sum_{m=1}^{N} a_{m\to n} \cdot w_{m\to n} \cdot s_{t-\Delta,m} - \log \nu_t^{(j_n)} \right\}.$$

The instantaneous firing rate take the form of a generalized linear model. Each neuron has a baseline rate that is a function of b_n and λ_{max}. Moreover, the rate is influenced by recent spiking activity through the network, $\boldsymbol{A} \odot \boldsymbol{W}$, and through the local normalization, ν.

According to our theory of neural inference, the sparsity pattern of the network should follow an independent Bernoulli model. That is, each edge is present with the same probability, $a_{m\to n} \sim \mathrm{Bern}(\rho)$. Moreover, the weights of network should encode the pairwise log probabilities, $\log \phi(z_i = k, z_j = k')$, and the weights from the R neurons representing $z_i = k$ to the R neurons representing $z_j = k'$ should all be approximately equal. This corresponds to stochastic block structure in the weight matrix. Thus, to reverse engineer the probabilistic model from the observed spike train, we fit a generalized linear model with a spike-and-slab network prior that has an independent Bernoulli model for the adjacency matrix, and a stochastic block model (SBM) for the weight matrix. The SBM has latent variables for each neuron that indicate the cluster assignment, and parameters that specify the average weight between each pair of clusters:

$$\begin{aligned} c_n &\sim \mathrm{Discrete}(\alpha \mathbf{1}), \\ \mu_{c\to c'} &\sim \mathcal{N}(0, \sigma_0^2), \\ w_{m\to n} &\sim \mathcal{N}(\mu_{c_m \to c_n}, \sigma^2). \end{aligned}$$

By performing Bayesian inference in this model, we recover a posterior distribution over biases, $\{b_n\}_{n=1}^N$; weighted adjacency matrices, $\boldsymbol{A}$ and $\boldsymbol{W}$; latent variables, $\{c_n\}_{n=1}^N$; and parameters, $\{\mu_{c \to c'}\}_{c,c'=1}^C$.

The normalization presents a minor complication. According to the theory, this is most likely computed by local inhibitory neurons that estimate the population rate of neurons representing z_j and deliver a common, normalizing input to stabilize the rate at the desired level. This can be roughly approximated with direct, excitatory connections between neurons representing the same variable-value pair, and inhibitory connections between neurons representing competing values of the same variable. In other words, if we focus solely on the activity of neurons representing the variable-value pairs, we should expect additional *functional* connections that encode the mutually exclusive nature of the distinct values of a given variable.

We demonstrate this approach by fitting the hierarchical model to a neural spike train simulated from the population described in Section 9.5. The population performs inference in a simple mixture model with three latent mixture components, and five binary observations. An observation consists of an assignment of the five observations, indicated by the variables $\{x_j\}_{j=1}^5$, and given an observation, the dynamics perform posterior inference of z, the latent variable indicating the underlying mixture component. The population consists of 130 neurons, 10 for each variable-value pair.

Figure 9.5 shows the inferred parameters of the hierarchical model. The posterior mean of the weighted adjacency matrix is shown in Figure 9.5a, with neurons sorted by variable and then by value. The block diagonal structure shows the normalizing connections between neurons representing the same variable, and the region highlighted in yellow shows the inferred connections from neurons representing x_j to neurons representing z. These encode the pairwise log probabilities of the mixture model.

Figure 9.5b shows the inferred posterior probability of two neurons belonging to the same cluster. While the true model has 13 clusters, we allowed our model to use as many as 20 clusters. If the variable-value subpopulations were recovered perfectly, this matrix would be block diagonal. We see that it is nearly so; only a handful of neurons are misclassified and some blocks are split in two.

Finally, Figure 9.5c shows the true and inferred mean weights under the stochastic block

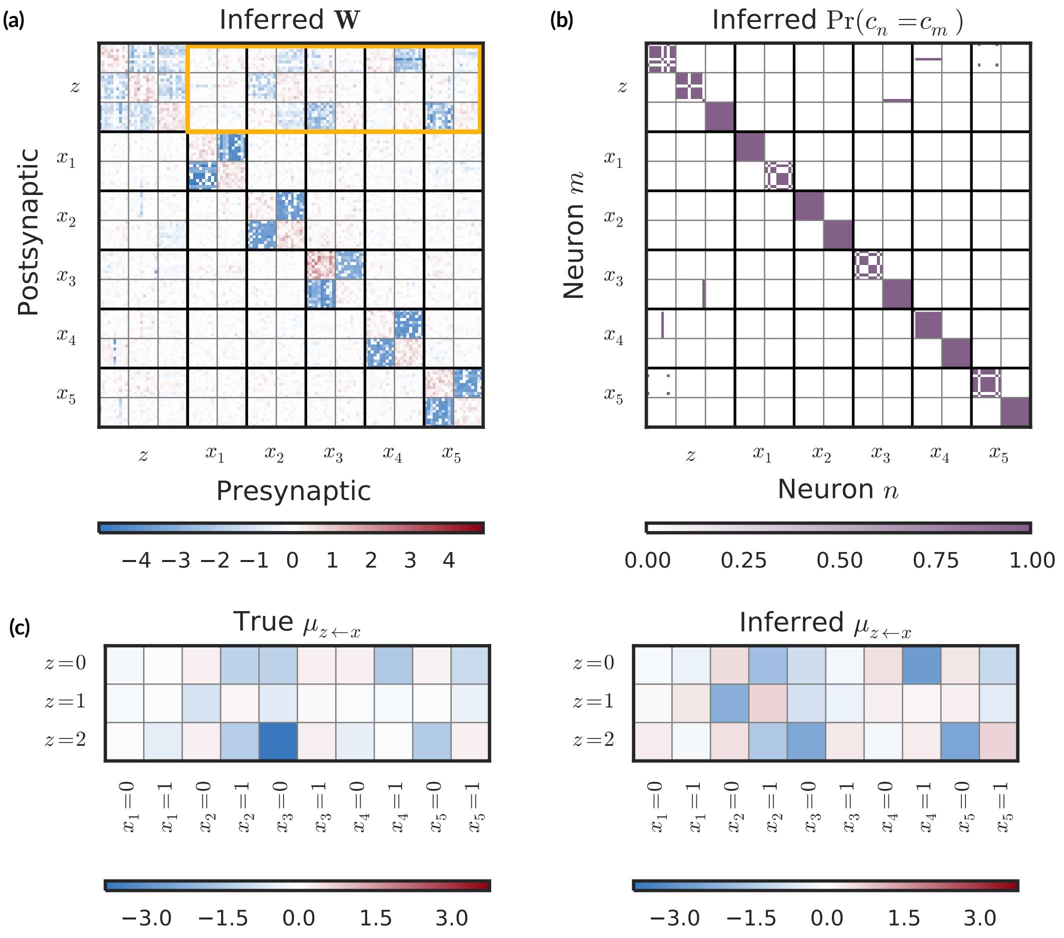

Figure 9.5: The probabilistic model can be reverse engineered from the neural spike train. **(a)** Inferred weighted adjacency matrix for the population of 130 neurons. Thin lines delineate boundaries between subpopulations for each value of z and x_j; bold lines separate populations for each variable. **(b)** Inferred probability that each pair of neurons belongs to the same cluster under a stochastic block model. The block diagonal structure shows that the variable-value subpopulations are clearly recovered. **(c)** True and inferred weights from x_j to z (yellow square in **(a)**). Inferred weights are the mean weights under the stochastic block model. They accurately recover the true weights.

model. First, we found the permutation of inferred cluster labels that best matched the true cluster labels. Then we found the linear transformation that best matched the true and inferred weights. The result shows that the true pattern of weights from x_j to z are recovered with high fidelity.

While this procedure for reverse engineering probabilistic models from observed spike trains is not foolproof, this simple example illustrates that much can be learned by combining top-down theories with bottom-up analysis. To further improve this inference procedure, we should include two sets of latent cluster assignments in our hierarchical model: one set of variables that specifies the variable that a cluster of neurons represents (i.e. j_n in our theory), and another that indicates the value (i.e. k_n). Incorporating the knowledge that different values of the same variable are mutually exclusive, we can build a strong prior distribution over weights given these two variables.

What else can be learned from the results of this approach? In practice, we only observe a fraction of the neurons in a particular region If our recording method samples N neurons out of N_{total}, then given an inferred block size, $\widehat{R}$, we can estimate the true representation size to be roughly, $R \approx \widehat{R}N_{\mathsf{total}}/N$. If variable-value subpopulations are truly disjoint, this provides an estimate of the number of subpopulations the region could encode. Combined with the complexity theoretic bounds developed in Section 9.3, these top-down and bottom-up approaches provide two tacks by which we may converge on a theory of probabilistic inference in neural circuits.

Future Work

This chapter has illustrated how theoretical models of neural computation may be assessed from a "top-down" perspective by analyzing the complexity and comparing it to biological parameters, and from the "bottom-up" perspective, by incorporating theoretical dynamics into probabilistic models of neural population activity. This is only a first step toward closing the gap between these two perspectives, and it opens many important questions. We enumerate and partially answer a few of them here.

Unsupervised Learning via Synaptic Plasticity

Perhaps the most pressing question is that of learning: how are these subpopulations of neurons and their weighted connections established? While we do not have a concrete answer to this question, we speculate that subpopulations are primarily allocated in a supervised fashion by a process like those of Valiant (1994). Once the neurons have been allocated, their weights may be tuned in an unsupervised manner. We suggest one way in which this unsupervised learning may be related to the process of spike-timing dependent plasticity, based on the work of Nessler et al. (2013).

The parameters of the model, $\boldsymbol{\theta}$, specify the conditional probabilities for pairs of hidden and visible variables. Rather than treating the parameters as given, we now treat them as part of the model.

$$\begin{aligned} p(\boldsymbol{z}, \boldsymbol{\theta}) &= p(\boldsymbol{\theta})\, p(\boldsymbol{z} \mid \boldsymbol{\theta}) \\ &= \frac{p(\boldsymbol{\theta})}{Z(\boldsymbol{\theta})} \prod_{j \in \mathcal{G}} \phi(z_j \mid \boldsymbol{\theta}) \prod_{i,j \in \mathcal{G}} \phi(z_i, z_j \mid \boldsymbol{\theta}). \end{aligned}$$

The challenge with learning is that the parameters appear in the normalizing constant, $Z(\boldsymbol{\theta})$, which is typically an intractable summation over variable assignments. For this simple example, we will only consider learning in a subset of models that can formulated as directed graphical models.

Assumption 5. *The following unsupervised learning algorithm assumes that the probabilistic model not only factors into the product of unary and pairwise potentials, but that this factorization corresponds to a directed graphical model in which the variables have at most one "parent" variable. That is, the variables are ordered such that the joint probability is equal to,*

$$p(\boldsymbol{z}, \boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{j=1}^{J} p(z_j \mid \mathsf{pa}(z_j), \boldsymbol{\theta}),$$

where $\mathsf{pa}(z_j) \in \{\varnothing, z_1, \ldots, z_{j-1}\}$. *Each conditional distribution in this product is properly normalized, which implies that the joint distribution is normalized as well.*

While this is clearly a strict assumption, it allows for some realistic models like mixtures and hidden Markov models. The advantage is that, here, the distribution is normalized such that the parameters appear only in their prior and in the conditional distributions, which depend on at most two variables. This will map nicely onto synaptic plasticity rules.

Since we are assuming the variables are discrete, the parameters $\boldsymbol{\theta}$ specify either the marginal probability of z_j (if $\mathsf{pa}(z_j) = \varnothing$) or the rows of a conditional probability table (if $\mathsf{pa}(z_j) \in \{z_1, \dots, z_{j-1}\}$). We make this explicit with the following notation,

$$\begin{aligned} p(z_j \mid \mathsf{pa}(z_j) = \varnothing, \boldsymbol{\theta}) &= \mathrm{Discrete}(\boldsymbol{\theta}^{(j)}), \\ p(z_j \mid \mathsf{pa}(z_j) = z_{j'} = k, \boldsymbol{\theta}) &= \mathrm{Discrete}(\boldsymbol{\theta}^{(j,k)}). \end{aligned}$$

In words, if the variable z_j has no parent, it is marginally distributed according to a categorical distribution with parameter $\boldsymbol{\theta}^{(j)}$. If variable z_j has parent $z_{j'}$, then when $z_{j'} = k$, the variable z_j follows a categorical distribution with parameter $\boldsymbol{\theta}^{(j,k)}$.

To incorporate these parameters into the model, we introduce Dirichlet priors over the probability vectors,

$$\boldsymbol{\theta}^{(j)} \sim \mathrm{Dir}(\alpha \mathbf{1}), \qquad \boldsymbol{\theta}^{(j,k)} \sim \mathrm{Dir}(\alpha \mathbf{1}).$$

Learning in a Bayesian framework corresponds to performing posterior inference over the parameters. Thus, we introduce a variational factor for $\boldsymbol{\theta}$ as well,

$$q(\boldsymbol{\theta}) = \prod_{j : \mathsf{pa}(z_j) = \varnothing} q(\boldsymbol{\theta}^{(j)}) \prod_{j : \mathsf{pa}(z_j) \neq \varnothing} \prod_{k} q(\boldsymbol{\theta}^{(j,k)})$$

Consider the variational factor for $\boldsymbol{\theta}^{(j,k)}$. Omitting the details, we can show that this factor takes the form of a Dirichlet distribution,

$$\begin{aligned} q_t(\boldsymbol{\theta}^{(j,k)}) &= \mathrm{Dir}(\boldsymbol{\theta}^{(j,k)} \mid \boldsymbol{\alpha}_t^{(j,k)}), \\ \boldsymbol{\alpha}_t^{(j,k)} &= \alpha + \widehat{\boldsymbol{\pi}}_{t-1}^{(j)} \cdot \widehat{\pi}_{t-1,k}^{(\mathsf{pa}(z_j))}. \end{aligned}$$

The updates of $q(z_j)$ must consider expectations with respect to this Dirichlet factor:

$$\log q(z_j) \simeq \mathbb{E}_{q(\boldsymbol{z}_{\neg j})} \mathbb{E}_{q(\boldsymbol{\theta})} \left[p(\boldsymbol{z}, \boldsymbol{\theta}) \right].$$

This expecation is given by,

$$\begin{aligned}\mathbb{E}_{q_t(\boldsymbol{\theta})}\left[\log p(z_j = k \mid \mathsf{pa}(z_j) = k', \boldsymbol{\theta})\right] &= \mathbb{E}_{q_t(\boldsymbol{\theta})}\left[\log \theta_k^{(j,k')}\right] \\ &= \psi\big(\alpha_{t,k}^{(j,k')}\big) - \psi\big(\sum_{i=1}^{K} \alpha_{t,i}^{(j,k')}\big) \\ &= \psi\big(\alpha_{t,k}^{(j,k')}\big) - \psi\big(K\alpha + \widehat{\pi}_{t-1,k'}^{(\mathsf{pa}(z_j))}\big). \end{aligned} \tag{9.6}$$

How could this be implemented biologically? First, we assume that learning occurs on a timescale of T_L time steps, which is relatively slow compared to the time scales of inference and behavior. That is, $T_I < T_L$. This allows the learning algorithm to generalize from many input rather than overfitting to a single example.

We want the synaptic weights to equal the expected log parameter value, as in (9.6). In theory, the weights should be identical for all synapses between neurons representing $(z_j = k)$ and neurons representing $(z_{j'} = k')$. It is unreasonable to assume this in practice, since these synapses exist between different neurons and are updated independently. However, we can specify a simple learning rule that would give rise to the same weights in expectation.

Assume that each synapse has a two latent state variabless, $\alpha_{t,m\to n}$, and $\beta_{t,m\to n}$. These will enable us to compute the expectation with respect to the variational parameter. We propose the following learning rule for the first state,

$$\begin{aligned}\alpha_{t,m\to n} &= \alpha + (\lambda_{\mathsf{max}}^2 T_L)^{-1} \sum_{\Delta=1}^{T_L} s_{t-\Delta,n} \cdot s_{t-\Delta,m} \\ &\approx \alpha + \widehat{\pi}_{t-1,k_n}^{(j_n)} \cdot \widehat{\pi}_{t-1,k_m}^{(j_m)}.\end{aligned}$$

Suppose neuron n represents the child variable and neuron m represents the parent variable. Then,[†]

$$\begin{aligned}\beta_{t,m\to n} &= K\alpha + (\lambda_{\mathsf{max}} T_L)^{-1} \sum_{\Delta=1}^{T_L} s_{t-\Delta,m} \\ &\approx K\alpha + \widehat{\pi}_{t-1,k_m}^{(z_{jm})}.\end{aligned}$$

The synaptic weight is then a deterministic function of these two state variables,

$$w_{t,m\to n} = \psi(\alpha_{t,m\to n}) - \psi(\beta_{t,m\to n}).$$

This state-based learning rule is Hebbian in that correlated spiking activity leads to increases in $\alpha_{t,m\to n}$, which in turn lead to larger weights (since the digamma function is increasing on the nonnegative reals). This is counteracted by the accrual of $\beta_{t,m\to n}$, which counts pre- or post-synaptic spikes, depending on whether the post-synaptic neuron represents the child or parent variable, respectively. If this value is large relative to $\alpha_{t,m\to n}$, the spike correlation is low relative to the background rate, which implies a low probability and a strongly negative weight.

Moreover, this learning rule is nonlinear. While the state variables are linear functions of pre- and post-synaptic spike counts, their effect on the weight is highly nonlinear due to the digamma functions. We could instead write this learning rule as a nonlinear dynamical system on the weights alone since the digamma function is also invertible on this range. Using the tools developed in Chapter 6, this dynamic learning process could potentially be incorporated in a probabilistic model for neural activity as well. We leave this for future work.

Representing Continuous Random Variables

While this chapter has focused on representing discrete random variables, many of the quantities we need to infer and reason about are continuous in nature. Suppose that we wish to represent a random variable $z \in \mathbb{R}^D$. Rather than representing the parameters

[†] Here, the synaptic state variable counts spikes on the pre-synaptic neuron. If the parent-child order was flipped, the synapse would have to count post-synaptic spikes instead. This asymmetry is admittedly somewhat unsatisfying.

of a standard distribution, like the mean and variance of a Gaussian, the brain may use a nonparametric representation like a kernel density estimate for the variational factors (Anderson and Essen, 1994; Barber et al., 2003). Suppose,

$$q_t(z_j) \propto \sum_{k=1}^{K} \eta_{t,k}^{(j)} \, \zeta(z_j; \mu_k),$$

where $\{\eta_{t,k}^{(j)}\}_{k=1}^{K}$ is a set of nonnegative weights that sum to one, $\zeta(z; \mu)$ is a nonnegative "kernel function" that integrates to one and has mean μ, and $\{\mu_k\}$ is the set of means at which these kernels are located. This defines a proper density function because $q_t(z_j)$ is nonnegative and integrates to one,

$$\int q_t(z_j) \, \mathrm{d}z_j = \sum_{k=1}^{K} \int \eta_{t,k}^{(j)} \, \zeta(z; \mu_k) \, \mathrm{d}z_j = \sum_{k=1}^{K} \eta_{t,k}^{(j)} = 1.$$

To implement this with a distributed population of neurons, let

$$\eta_{t,k}^{(j)} = \frac{\sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{I}[j_n = j, k_n = k] \, s_{t,n}}{\sum_{n=1}^{N} \mathbb{I}[j_n = j] s_{t,n}}.$$

The mean field variational inference algorithm is no longer as simple as in Section 9.4, but the key quantities of the variational lower bound, namely the entropy of the variational factor and the expected log probability, are still tractable. Using an approach similar to that of Gershman et al. (2012a), we have,

$$\begin{aligned}
\mathbb{E}_{q_t(\boldsymbol{z})}\left[\log p(\boldsymbol{z} \mid \boldsymbol{\theta})\right] &= \mathbb{E}_{q_t(\boldsymbol{z})}\left[\sum_{j\in\mathcal{G}} \log \phi(z_j \mid \boldsymbol{\theta}) + \sum_{i,j\in\mathcal{G}} \log \phi(z_i, z_j \mid \boldsymbol{\theta})\right] \\
&= \sum_{j\in\mathcal{G}} \sum_{k=1}^{K} \eta_{t,k}^{(j)} \int \log \phi(z_j \mid \boldsymbol{\theta}) \zeta(z_j; \mu_k) \, \mathrm{d}z_j \\
&\quad + \sum_{i,j\in\mathcal{G}} \sum_{k=1}^{K} \sum_{k'=1}^{K} \eta_{t,k'}^{(i)} \, \eta_{t,k}^{(j)} \iint \log \phi(z_i, z_j \mid \boldsymbol{\theta}) \, \zeta(z_i; \mu_k) \, \zeta(z_j; \mu_{k'}) \, \mathrm{d}z_i \, \mathrm{d}z_j \\
&= \sum_{j\in\mathcal{G}} \sum_{k=1}^{K} \eta_{t,k}^{(j)} \, \widetilde{b}_k^{(j)} + \sum_{i,j\in\mathcal{G}} \sum_{k'=1}^{K} \sum_{k=1}^{K} \eta_{t,k'}^{(i)} \, \eta_{t,k}^{(j)} \, \widetilde{w}_{k',k}^{(i,j)}
\end{aligned}$$

The second term of the variational lower bound is the entropy of the variational distribution,

$$\begin{aligned}
\mathcal{H}[q_t(z_j)] &= -\int q_t(z_j) \log q_t(z_j) \, \mathrm{d}z_j \\
&= -\int q_t(z_j) \log \sum_{k=1}^{K} \eta_{t,k}^{(j)} \, \zeta(z_j; \mu_k) \, \mathrm{d}z_j.
\end{aligned}$$

We can lower bound this with Jensen's inequality to get,

$$\begin{aligned}
\mathcal{H}[q_t(z_j)] &\geq \sum_{k=1}^{K} \log \int q_t(z_j) \, \eta_{t,k}^{(j)} \, \zeta(z_j; \mu_k) \, \mathrm{d}z_j \\
&= \sum_{k=1}^{K} \log \left(\eta_{t,k}^{(j)} \sum_{k'=1}^{K} \eta_{t,k'}^{(j)} \int \zeta(z_j; \mu_{k'}) \, \zeta(z_j; \mu_k) \, \mathrm{d}z_j \right) \\
&= \sum_{k=1}^{K} \log \left(\eta_{t,k}^{(j)} \sum_{k'=1}^{K} \eta_{t,k'}^{(j)} \, \zeta_{k,k'}^{*} \right)
\end{aligned}$$

where we have defined $\zeta_{k,k'}^{*} = \zeta_{k',k}^{*} = \int \zeta(z_j; \mu_{k'}) \, \zeta(z_j; \mu_k) \, \mathrm{d}z_j$ as the convolution of a pair of kernel functions.

Now, to perform inference, we can perform gradient ascent directly on the *evidence lower bound* (ELBO) on the log marginal likelihood, which is just the sum of these two terms. That is, we drive firing rates such that,

$$\eta^{(j)}_{t,k} = \eta^{(j)}_{t-1,k} + \alpha \nabla_{\boldsymbol{\eta}} \left(\mathbb{E}_{q_{t-1}(\boldsymbol{z})} \left[\log p(\boldsymbol{z} \mid \boldsymbol{\theta}) \right] + \mathcal{H}[q_{t-1}(z_j)] \right).$$

The gradient of the ELBO has two components, the first from the expected log probability and the second from the entropy. The first is quite intuitive,

$$\frac{\partial}{\partial \eta^{(j)}_{t,k}} \mathbb{E}_{q_t(\boldsymbol{z})} \left[\log p(\boldsymbol{z} \mid \boldsymbol{\theta}) \right] = \widetilde{b}^{(j)}_k + \sum_{i \in \mathsf{ne}(j)} \sum_{k'=1}^{K} \eta^{(i)}_{t,k'} \, \widetilde{w}^{(i,j)}_{k',k}.$$

As in the discrete random variable case, the firing rates of neurons representing the pair (j, k), are driven by a bias and a weighted sum of activity from connected neurons. Now, however, the bias and the weights reflect integrations with respect to the basis functions.

The second term comes from the lower bound on the entropy,

$$\begin{aligned} \frac{\partial}{\partial \eta^{(j)}_{t,k}} \mathcal{H}[q_t(z_j)] &= \frac{1}{\eta^{(j)}_{t,k}} + \sum_{k' \neq k} \frac{\partial}{\partial \eta^{(j)}_{t,k}} \log \left(\eta^{(j)}_{t,k} \, \zeta^{*}_{k',k} + \sum_{k'' \neq k} \eta^{(j)}_{t,k''} \, \zeta^{*}_{k',k''} \right) \\ &= \frac{1}{\eta^{(j)}_{t,k}} + \sum_{k' \neq k} \zeta^{*}_{k',k} \left(\textstyle\sum_{k''=1}^{K} \eta^{(j)}_{t,k''} \, \zeta^{*}_{k',k''} \right)^{-1}. \end{aligned} \tag{9.7}$$

While less intuitive, this term effectively provides a damping signal that prevents one kernel from dominating the rest. As the rates approach one, the first term in (9.7) diminishes. We leave more detailed studies of the biological plausibility of this approach to future work.

Alternative Representations of Probability

The introduction enumerated a host of potential neural representations of probability and corresponding inference algorithms. Eventually, these combinations of representation and algorithm lead to some prediction of the dynamics of neural spiking. Often, these dynamics follow standard forms, like the linear-nonlinear cascade of the GLM. If this is the case, then we should be able to derive probabilistic models for neural data that incorporate the hypothesized dynamics of Bayesian inference.

One popular theory of representation is the *probabilistic population code* (PPC) (Ma et al., 2006). According to this theory, the Poisson-like variability of neurons leads to a likelihood of a random variable for any particular spike train, $p(\boldsymbol{s} \mid z_j)$. Combined with a prior, this yields a posterior distribution, $p(z_j \mid \boldsymbol{s})$. In their theory, the encoded distribution is exactly this posterior, $\widehat{p}(z_j) = p(z_j \mid \boldsymbol{s})$.

This leads to two levels of randomness. While the neurons may be driven with firing rates that, in expectation, encode the distribution $\bar{p}(z_j)$, the randomness in $\boldsymbol{s}$ implies a distribution over encoded distributions, $p(\widehat{p}(z_j))$. This doubly stochastic nature has been explored by Zemel et al. (1998), Sahani and Dayan (2003), and others. Ma et al. (2006) skirt this issue by assuming that as the number of neurons grows, this distribution over distributions collapses to its mode, $p(\widehat{p}(z_j)) = \delta_{p^*(z_j)}$, which is presumed to be approximately equal to the desired distribution. That is, $p^*(z_j) \approx \bar{p}(z_j)$. In their example, a Gaussian distribution is encoded by a population of neurons with radial basis function tuning curves. The mean is encoded by the relative firing rate of the activity, and the precision is encoded by the absolute firing rate, or gain, of the population.

Their main contribution is a demonstration of how inference in some simple probabilistic models, like a naïve Bayes model for cue combination, can be performed with simple linear functions on PPCs. For example, in simple naïve Bayes models, downstream neurons only need to sum population activity to combine evidence and compute an updated posterior. However, other probabilistic computations, like marginalization and variational inference, do require nonlinear operations (Beck et al., 2011; 2012).

If neural populations are performing inference with PPCs using linear operations, then the neural spike trains recorded from these populations should follow linear Hawkes process dynamics. Thus, the tools of Chapters 2 and 3 could provide a mechanism for inferring the connection weights. Given these connection weights and an estimate of the tuning curves, it should be possible to reverse engineer the underlying probabilistic model. This provides one more avenue toward connecting theory and experiment.

Conclusion

This chapter has taken a novel look at the problem of connecting the theory of neural computation to experimental recording. While the traditional approach of making specific, testable predictions of a theory remains invaluable, this is a process we would like to automate as much as possible. With the advent of large-scale recording technologies, the bottleneck in the scientific process moves from collecting evidence to designing experiments and revising theories. Here, we have suggested that the scientific loop of theorizing, experimenting, and revising may be closed by formulating our theories in the language of prior distributions in a Bayesian probabilistic model of neural data. With such a model, we could hypothetically measure the marginal likelihood of a theory given the data, suggest experiments to refine our estimate of theoretical values, and revise our theories in an automated fashion. This chapter has not nearly closed this loop, but it has provided a framework for thinking about a future in which theoretical, computational, and systems neuroscience are tightly tethered.

References

Yashar Ahmadian, Jonathan W Pillow, and Liam Paninski. Efficient Markov chain Monte Carlo methods for decoding neural spike trains. *Neural Computation*, 23(1):46–96, 2011.

Misha B Ahrens, Michael B Orger, Drew N Robson, Jennifer M Li, and Philipp J Keller. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature Methods*, 10(5):413–420, 2013.

Laurence Aitchison and Peter E Latham. Synaptic sampling: A connection between PSP variability and uncertainty explains neurophysiological observations. *arXiv preprint arXiv:1505.04544*, 2015.

Laurence Aitchison and Máté Lengyel. The Hamiltonian brain. *arXiv preprint arXiv:1407.0973*, 2014.

David J Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

Charles H Anderson and David C Van Essen. Neurobiological computational systems. *Computational Intelligence Imitating Life*, pages 1–11, 1994.

Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50(1-2):5–43, 2003.

Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

Michael J Barber, John W Clark, and Charles H Anderson. Neural representation of probabilistic information. *Neural Computation*, 15(8):1843–64, August 2003.

Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, 1966.

Matthew J. Beal, Zoubin Ghahramani, and Carl E. Rasmussen. The infinite hidden Markov model. *Advances in Neural Information Processing Systems 14*, pages 577–585, 2002.

Jeffrey M Beck and Alexandre Pouget. Exact inferences in a neural implementation of a hidden Markov model. *Neural Computation*, 19(5):1344–1361, 2007.

Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Marginalization in neural circuits with divisive normalization. *The Journal of Neuroscience*, 31(43):15310–15319, 2011.

Jeffrey M Beck, Katherine A Heller, and Alexandre Pouget. Complex inference in neural circuits with probabilistic population codes and topic models. *Advances in Neural Information Processing Systems*, pages 3059–3067, 2012.

Yoshua Bengio and Paolo Frasconi. An input output HMM architecture. *Advances in Neural Information Processing Systems*, pages 427–434, 1995.

Pietro Berkes, Gergo Orbán, Máté Lengyel, and József Fiser. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331(6013): 83–7, January 2011.

Gordon J Berman, Daniel M Choi, William Bialek, and Joshua W Shaevitz. Mapping the stereotyped behaviour of freely moving fruit flies. *Journal of The Royal Society Interface*, 11(99):20140672, 2014.

Philippe Biane, Jim Pitman, and Marc Yor. Probability laws related to the Jacobi theta and Riemann zeta functions, and Brownian excursions. *Bulletin of the American Mathematical Society*, 38(4):435–465, 2001.

Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

David M Blei. Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203–232, 2014.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

Carolyn R Block and Richard Block. *Street gang crime in Chicago*. US Department of Justice, Office of Justice Programs, National Institute of Justice, 1993.

Carolyn R Block, Richard Block, and Illinois Criminal Justice Information Authority. Homicides in Chicago, 1965-1995. ICPSR06399-v5. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], July 2005.

Charles Blundell, Katherine A Heller, and Jeffrey M Beck. Modelling reciprocating relationships with Hawkes processes. *Advances in Neural Information Processing Systems*, pages 2600–2608, 2012.

George EP Box. Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, pages 383–430, 1980.

David H Brainard and William T Freeman. Bayesian color constancy. *Journal of the Optical Society of America A*, 14(7):1393–1411, 1997.

Kevin L Briggman, Henry DI Abarbanel, and William B Kristan. Optical imaging of neuronal populations during decision-making. *Science*, 307(5711):896–901, 2005.

David R. Brillinger. Maximum likelihood analysis of spike trains of interacting nerve cells. *Biological Cybernetics*, 59(3):189–200, August 1988.

David R Brillinger, Hugh L Bryant Jr, and Jose P Segundo. Identification of synaptic interactions. *Biological Cybernetics*, 22(4):213–228, 1976.

Michael Bryant and Erik B Sudderth. Truly nonparametric online variational inference for hierarchical Dirichlet processes. *Advances in Neural Information Processing Systems 25*, pages 2699–2707, 2012.

Lars Buesing, Johannes Bill, Bernhard Nessler, and Wolfgang Maass. Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology*, 7(11):e1002211, November 2011.

Lars Buesing, Jakob H. Macke, and Maneesh Sahani. Learning stable, regularised latent models of neural population dynamics. *Network: Computation in Neural Systems*, 23: 24–47, 2012a.

Lars Buesing, Jakob H Macke, and Maneesh Sahani. Spectral learning of linear dynamics from generalised-linear observations with application to neural population data. *Advances in Neural Information Processing Systems*, pages 1682–1690, 2012b.

Lars Buesing, Timothy A Machado, John P Cunningham, and Liam Paninski. Clustered factor analysis of multineuronal spike data. *Advances in Neural Information Processing Systems*, pages 3500–3508, 2014.

Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.

Santiago Ramón Cajal. *Textura del Sistema Nervioso del Hombre y los Vertebrados*, volume 1. Imprenta y Librería de Nicolás Moya, Madrid, Spain, 1899.

Natalia Caporale and Yang Dan. Spike timing-dependent plasticity: a Hebbian learning rule. *Annual Review of Neuroscience*, 31:25–46, 2008.

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

Zhe Chen, Fabian Kloosterman, Emery N Brown, and Matthew A Wilson. Uncovering spatial topology represented by rat hippocampal population neuronal codes. *Journal of Computational Neuroscience*, 33(2):227–255, 2012.

Zhe Chen, Stephen N Gomperts, Jun Yamamoto, and Matthew A Wilson. Neural representation of spatial topology in the rodent hippocampus. *Neural Computation*, 26(1): 1–39, 2014.

Sharat Chikkerur, Thomas Serre, Cheston Tan, and Tomaso Poggio. What and where: A Bayesian inference theory of attention. *Vision Research*, 50(22):2233–2247, 2010.

Yoon Sik Cho, Aram Galstyan, Jeff Brantingham, and George Tita. Latent point process models for spatial-temporal networks. *arXiv:1302.2671*, 2013.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

Aaron C Courville, Nathaniel D Daw, and David S Touretzky. Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, 10(7):294–300, 2006.

Ronald L Cowan and Charles J Wilson. Spontaneous firing patterns and axonal projections of single corticostriatal neurons in the rat medial agranular cortex. *Journal of Neurophysiology*, 71(1):17–32, 1994.

W Maxwell Cowan, Thomas C Südhof, and Charles F Stevens. *Synapses*. Johns Hopkins University Press, 2003.

Mary Kathryn Cowles and Bradley P Carlin. Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91: 883–904, 1996.

John P Cunningham and Byron M Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, 2014.

Paul Dagum and Michael Luby. Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence*, 60(1):141–153, 1993.

Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: Volume I: Elementary Theory and Methods*. Springer Science & Business Media, 2 edition, 2003.

Peter Dayan and Larry F Abbott. *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. MIT Press, 2001.

Peter Dayan and Joshua A Solomon. Selective Bayes: Attentional load and crowding. *Vision Research*, 50(22):2248–2260, 2010.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

Sophie Deneve. Bayesian spiking neurons I: inference. *Neural Computation*, 20(1):91–117, January 2008.

Luc Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, USA, 1986.

Christopher DuBois, Carter Butts, and Padhraic Smyth. Stochastic block modeling of relational event dynamics. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 238–246, 2013.

Seif Eldawlatly, Yang Zhou, Rong Jin, and Karim G Oweiss. On the use of dynamic Bayesian networks in reconstructing functional neuronal networks from spike train ensembles. *Neural Computation*, 22(1):158–189, 2010.

Marc O Ernst and Martin S Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.

Sean Escola, Alfredo Fontanini, Don Katz, and Liam Paninski. Hidden Markov models for the stimulus-response relationships of multistate neural systems. *Neural Computation*, 23(5):1071–1132, 2011.

Warren John Ewens. Population genetics theory—the past and the future. In S. Lessard, editor, *Mathematical and Statistical Developments of Evolutionary Theory*, pages 177–227. Springer, 1990.

Daniel E Feldman. The spike-timing dependence of plasticity. *Neuron*, 75(4):556–71, August 2012.

Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47, 1991.

Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

Christopher R Fetsch, Amanda H Turner, Gregory C DeAngelis, and Dora E Angelaki. Dynamic reweighting of visual and vestibular cues during self-motion perception. *The Journal of Neuroscience*, 29(49):15601–15612, 2009.

Christopher R Fetsch, Alexandre Pouget, Gregory C DeAngelis, and Dora E Angelaki. Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience*, 15(1):146–154, 2012.

József Fiser, Pietro Berkes, Gergő Orbán, and Máté Lengyel. Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, 14(3):119–130, 2010.

Alyson K Fletcher, Sundeep Rangan, Lav R Varshney, and Aniruddha Bhargava. Neural reconstruction with approximate message passing (neuramp). *Advances in Neural Information Processing Systems*, pages 2555–2563, 2011.

Emily B Fox. *Bayesian nonparametric learning of complex dynamical phenomena*. PhD thesis, Massachusetts Institute of Technology, 2009.

Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. An HDP-HMM for systems with state persistence. *Proceedings of the International Conference on Machine Learning*, pages 312–319, 2008.

Jeremy Freeman, Greg D Field, Peter H Li, Martin Greschner, Deborah E Gunning, Keith Mathieson, Alexander Sher, Alan M Litke, Liam Paninski, Eero P Simoncelli, et al. Mapping nonlinear receptive field structure in primate retina at single cone resolution. *eLife*, 4:e05241, 2015.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, 11(2):127–38, February 2010.

Karl J Friston. Functional and effective connectivity in neuroimaging: a synthesis. *Human Brain Mapping*, 2(1-2):56–78, 1994.

Deep Ganguli and Eero P Simoncelli. Implicit encoding of prior probabilities in optimal neural populations. *Advances in Neural Information Processing Systems*, pages 6–9, 2010.

Peiran Gao and Surya Ganguli. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current Opinion in Neurobiology*, 32:148–155, 2015.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. CRC press, 3rd edition, 2013.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

Felipe Gerhard, Tilman Kispersky, Gabrielle J Gutierrez, Eve Marder, Mark Kramer, and Uri Eden. Successful reconstruction of a physiological circuit with known connectivity from spiking activity alone. *PLoS Computational Biology*, 9(7):e1003138, 2013.

Samuel J Gershman, Matthew D Hoffman, and David M Blei. Nonparametric variational inference. *Proceedings of the International Conference on Machine Learning*, pages 663–670, 2012a.

Samuel J Gershman, Edward Vul, and Joshua B Tenenbaum. Multistability and perceptual inference. *Neural Computation*, 24(1):1–24, 2012b.

Sebastian Gerwinn, Jakob Macke, Matthias Seeger, and Matthias Bethge. Bayesian inference for spiking neuron models with a sparsity prior. *Advances in Neural Information Processing Systems*, pages 529–536, 2008.

Charles J Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, pages 473–483, 1992.

Walter R Gilks. *Markov Chain Monte Carlo*. Wiley Online Library, 2005.

Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, and Edoardo M Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2010.

Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1019–1028, 2010.

Noah Goodman, Vikash Mansinghka, Daniel M Roy, Keith Bonawitz, and Joshua B Tenenbaum. Church: a language for generative models. *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 220––229, 2008.

Noah D Goodman, Joshua B Tenenbaum, and Tobias Gerstenberg. Concepts in a probabilistic language of thought. Technical report, Center for Brains, Minds and Machines (CBMM), 2014.

Agnieszka Grabska-Barwinska, Jeff Beck, Alexandre Pouget, and Peter Latham. Demixing odors-fast inference in olfaction. *Advances in Neural Information Processing Systems*, pages 1968–1976, 2013.

SG Gregory, KF Barlow, KE McLay, R Kaul, D Swarbreck, A Dunham, CE Scott, KL Howe, K Woodfine, CCA Spencer, et al. The DNA sequence and biological annotation of human chromosome 1. *Nature*, 441(7091):315–321, 2006.

Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. Bayesian models of cognition. In Ron Sun, editor, *The Cambridge Handbook of Computational Psychology*. Cambridge University Press, 2008.

Roger B Grosse, Chris J Maddison, and Ruslan R Salakhutdinov. Annealing between distributions by averaging moments. *Advances in Neural Information Processing Systems*, pages 2769–2777, 2013.

Roger B Grosse, Zoubin Ghahramani, and Ryan P Adams. Sandwiching the marginal likelihood using bidirectional Monte Carlo. *arXiv preprint arXiv:1511.02543*, 2015.

Yong Gu, Dora E Angelaki, and Gregory C DeAngelis. Neural correlates of multisensory cue integration in macaque MSTd. *Nature Neuroscience*, 11(10):1201–1210, 2008.

Fangjian Guo, Charles Blundell, Hanna Wallach, and Katherine A Heller. The Bayesian echo chamber: Modeling influence in conversations. *arXiv preprint arXiv:1411.2674*, 2014.

Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83, 1971.

Moritz Helmstaedter, Kevin L Briggman, Srinivas C Turaga, Viren Jain, H Sebastian Seung, and Winfried Denk. Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature*, 500(7461):168–174, 2013.

Geoffrey E Hinton. How neural networks learn from experience. *Scientific American*, 1992.

Geoffrey E Hinton and Terrence J Sejnowski. Optimal perceptual inference. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1983.

Daniel R Hochbaum, Yongxin Zhao, Samouil L Farhi, Nathan Klapoetke, Christopher A Werley, Vikrant Kapoor, Peng Zou, Joel M Kralj, Dougal Maclaurin, Niklas Smedemark-Margulies, et al. All-optical electrophysiology in mammalian neurons using engineered microbial rhodopsins. *Nature Methods*, 2014.

Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117 (4):500, 1952.

Peter D Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. *Advances in Neural Information Processing Systems*, 20:1–8, 2008.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Douglas N. Hoover. Relations on probability spaces and arrays of random variables. Technical report, Institute for Advanced Study, Princeton, 1979.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

Patrik O Hoyer and Aapo Hyvarinen. Interpreting neural response variability as Monte Carlo sampling of the posterior. *Advances in neural information processing systems*, pages 293–300, 2003.

Yanping Huang and Rajesh P. N. Rao. Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(5):580–593, September 2011.

David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, 1962.

Hemant Ishwaran and Mahmoud Zarepour. Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30(2):269–283, 2002.

Tomoharu Iwata, Amar Shah, and Zoubin Ghahramani. Discovering latent influence in online social activities via shared cascade Poisson processes. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 266–274, 2013.

Mehrdad Jazayeri and Michael N Shadlen. Temporal context calibrates interval timing. *Nature Neuroscience*, 13(8):1020–1026, 2010.

Mehrdad Jazayeri and Michael N Shadlen. A neural mechanism for sensing and reproducing a time interval. *Current Biology*, 25(20):2599–2609, 2015.

Matthew J Johnson. *Bayesian time series models and scalable inference*. PhD thesis, Massachusetts Institute of Technology, June 2014.

Matthew J Johnson and Alan S Willsky. Bayesian nonparametric hidden semi-Markov models. *Journal of Machine Learning Research*, 14(1):673–701, 2013.

Matthew J Johnson and Alan S Willsky. Stochastic variational inference for Bayesian time series models. *Proceedings of the International Conference on Machine Learning*, 32:1854–1862, 2014.

Matthew J Johnson, Scott W Linderman, Sandeep R Datta, and Ryan P Adams. Discovering switching autoregressive dynamics in neural spike train recordings. *Computational and Systems Neuroscience (Cosyne) Abstracts*, 2015.

Lauren M Jones, Alfredo Fontanini, Brian F Sadacca, Paul Miller, and Donald B Katz. Natural stimuli evoke dynamic sequences of states in sensory cortical ensembles. *Proceedings of the National Academy of Sciences*, 104(47):18772–18777, 2007.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

Eric R Kandel, James H Schwartz, Thomas M Jessell, et al. *Principles of neural science*, volume 4. McGraw-Hill New York, 2000.

David Kappel, Stefan Habenschuss, Robert Legenstein, and Wolfgang Maass. Network plasticity as Bayesian inference. *PLoS Computational Biology*, 11(11):e1004485, 2015a.

David Kappel, Stefan Habenschuss, Robert Legenstein, and Wolfgang Maass. Synaptic sampling: A Bayesian approach to neural network plasticity and rewiring. *Advances in Neural Information Processing Systems*, pages 370–378, 2015b.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

Jason ND Kerr and Winfried Denk. Imaging in vivo: watching the brain in action. *Nature Reviews Neuroscience*, 9(3):195–205, 2008.

Roozbeh Kiani and Michael N Shadlen. Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, 324(5928):759–64, May 2009.

John F. C. Kingman. *Poisson Processes (Oxford Studies in Probability).* Oxford University Press, January 1993. ISBN 0198536933.

David C Knill and Whitman Richards. *Perception as Bayesian inference.* Cambridge University Press, 1996.

Konrad P Körding and Daniel M Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–7, January 2004.

Alp Kucukelbir, Rajesh Ranganath, Andrew Gelman, and David Blei. Automatic variational inference in Stan. *Advances in Neural Information Processing Systems*, pages 568–576, 2015.

Stephen W Kuffler. Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, 16(1):37–68, 1953.

Harold W Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science*, 349(6244):184–187, 2015.

Tai Sing Lee and David Mumford. Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, 20(7):1434–1448, 2003.

Robert Legenstein and Wolfgang Maass. Ensembles of spiking neurons with noise support optimal probabilistic inference in a dynamically changing environment. *PLoS Computational Biology*, 10(10):e1003859, 2014.

William C Lemon, Stefan R Pulver, Burkhard Höckendorf, Katie McDole, Kristin Branson, Jeremy Freeman, and Philipp J Keller. Whole-central nervous system functional imaging in larval Drosophila. *Nature Communications*, 6, 2015.

Michael S Lewicki. A review of methods for spike sorting: the detection and classification of neural action potentials. *Network: Computation in Neural Systems*, 9(4):R53–R78, 1998.

Percy Liang, Slav Petrov, Michael I Jordan, and Dan Klein. The infinite PCFG using hierarchical Dirichlet processes. *Proceedings of Empirical Methods in Natural Language Processing*, pages 688–697, 2007.

David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

Jeff W Lichtman, Jean Livet, and Joshua R Sanes. A technicolour approach to the connectome. *Nature Reviews Neuroscience*, 9(6):417–422, 2008.

Scott W Linderman and Ryan P. Adams. Discovering latent network structure in point process data. *Proceedings of the International Conference on Machine Learning*, pages 1413–1421, 2014.

Scott W Linderman and Ryan P Adams. Scalable Bayesian inference for excitatory point process networks. *arXiv preprint arXiv:1507.03228*, 2015.

Scott W Linderman and Ryan P Johnson, Matthew Jand Adams. Dependent multinomial models made easy: Stick-breaking with the Pólya-gamma augmentation. *Advances in Neural Information Processing Systems*, pages 3438–3446, 2015.

Scott W Linderman, Christopher H Stock, and Ryan P Adams. A framework for studying synaptic plasticity with neural spike train data. *Advances in Neural Information Processing Systems*, pages 2330–2338, 2014.

Scott W Linderman, Ryan P Adams, and Jonathan W Pillow. Inferring structured connectivity from spike trains under negative-binomial generalized linear models. *Computational and Systems Neuroscience (Cosyne) Abstracts*, 2015.

Scott W Linderman, Matthew J Johnson, Matthew W Wilson, and Zhe Chen. A nonparametric Bayesian approach to uncovering rat hippocampal population codes during spatial navigation. *Journal of Neuroscience Methods*, 263:36–47, 2016a.

Scott W Linderman, Aaron Tucker, and Matthew J Johnson. Bayesian latent state space models of neural activity. *Computational and Systems Neuroscience (Cosyne) Abstracts*, 2016b.

Fredrik Lindsten, Michael I Jordan, and Thomas B Schön. Ancestor sampling for particle Gibbs. *Advances in Neural Information Processing Systems*, pages 2600–2608, 2012.

Shai Litvak and Shimon Ullman. Cortical circuitry implementing graphical models. *Neural Computation*, 21(11):3010–3056, 2009.

James Robert Lloyd, Peter Orbanz, Zoubin Ghahramani, and Daniel M Roy. Random function priors for exchangeable arrays with applications to graphs and relational data. *Advances in Neural Information Processing Systems*, 2012.

Wei Ji Ma and Mehrdad Jazayeri. Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, 37:205–220, 2014.

Wei Ji Ma, Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–8, November 2006.

David JC MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

Jakob H Macke, Lars Buesing, John P Cunningham, M Yu Byron, Krishna V Shenoy, and Maneesh Sahani. Empirical models of spiking in neural populations. *Advances in neural information processing systems*, pages 1350–1358, 2011.

Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

Vikash Mansinghka, Daniel Selsam, and Yura Perov. Venture: a higher-order probabilistic programming platform with programmable inference. *arXiv preprint arXiv:1404.0099*, 2014.

David Marr. *Vision: A computational investigation into the human representation and processing of visual information.* MIT Press, 1982.

Paul Miller and Donald B Katz. Stochastic transitions between neural states in taste processing and decision-making. *The Journal of Neuroscience*, 30(7):2559–2570, 2010.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023—-1032, 1988.

Shakir Mohamed, Zoubin Ghahramani, and Katherine A Heller. Bayesian and L1 approaches for sparse unsupervised learning. *Proceedings of the International Conference on Machine Learning*, pages 751–758, 2012.

Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482, 1998.

Michael L Morgan, Gregory C DeAngelis, and Dora E Angelaki. Multisensory integration in macaque visual cortex depends on cue reliability. *Neuron*, 59(4):662–673, 2008.

Abigail Morrison, Markus Diesmann, and Wulfram Gerstner. Phenomenological models of synaptic plasticity based on spike timing. *Biological Cybernetics*, 98(6):459–478, 2008.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Radford M Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

Radford M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, pages 113–162, 2010.

John A Nelder and R Jacob Baker. Generalized linear models. *Encyclopedia of Statistical Sciences*, 1972.

Bernhard Nessler, Michael Pfeiffer, Lars Buesing, and Wolfgang Maass. Bayesian computation emerges in generic cortical microcircuits through spike-timing-dependent plasticity. *PLoS Computational Biology*, 9(4):e1003037, 2013.

Mark EJ Newman. The structure and function of complex networks. *Society for Industrial and Applied Mathematics (SIAM) Review*, 45(2):167–256, 2003.

Krzysztof Nowicki and Tom A B Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

Seung Wook Oh, Julie A Harris, Lydia Ng, Brent Winslow, Nicholas Cain, Stefan Mihalas, Quanxin Wang, Chris Lau, Leonard Kuan, Alex M Henry, et al. A mesoscale connectome of the mouse brain. *Nature*, 508(7495):207–214, 2014.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, 1982.

John O'Keefe and Lynn Nadel. *The Hippocampus as a Cognitive Map*, volume 3. Clarendon Press, 1978.

Peter Orbanz and Daniel M Roy. Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):437–461, 2015.

Peter Orbanz and Yee Whye Teh. Bayesian nonparametric models. In *Encyclopedia of Machine Learning*, pages 81–89. Springer, 2011.

Adam M Packer, Darcy S Peterka, Jan J Hirtz, Rohit Prakash, Karl Deisseroth, and Rafael Yuste. Two-photon optogenetics of dendritic spines and neural circuits. *Nature Methods*, 9(12):1202–1205, 2012.

Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15(4):243–262, January 2004.

Liam Paninski, Yashar Ahmadian, Daniel Gil Ferreira, Shinsuke Koyama, Kamiar Rahnama Rad, Michael Vidne, Joshua Vogelstein, and Wei Wu. A new look at state-space models for neural data. *Journal of Computational Neuroscience*, 29(1-2):107–126, 2010.

Andrew V Papachristos. Murder by structure: Dominance relations and the social structure of gang homicide. *American Journal of Sociology*, 115(1):74–128, 2009.

Il Memming Park and Jonathan W Pillow. Bayesian spike-triggered covariance analysis. *Advances in Neural Information Processing Systems*, pages 1692–1700, 2011.

Patrick O Perry and Patrick J Wolfe. Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2013.

Biljana Petreska, Byron Yu, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and Maneesh Sahani. Dynamical segmentation of single trials from population neural data. *Advances in Neural Information Processing Systems*, pages 756–764, 2011.

David Pfau, Eftychios A Pnevmatikakis, and Liam Paninski. Robust learning of low-dimensional dynamics from large neural ensembles. *Advances in Neural Information Processing Systems*, pages 2391–2399, 2013.

Jonathan W. Pillow and James Scott. Fully Bayesian inference for neural models with negative-binomial spiking. *Advances in Neural Information Processing Systems*, pages 1898–1906, 2012.

Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky, and Eero P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008.

Eftychios A Pnevmatikakis, Daniel Soudry, Yuanjun Gao, Timothy A Machado, Josh Merel, David Pfau, Thomas Reardon, Yu Mu, Clay Lacefield, Weijian Yang, et al. Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron*, 2016.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using Pólya-gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.

Ruben Portugues, Claudia E Feierstein, Florian Engert, and Michael B Orger. Whole-brain activity maps reveal stereotyped, distributed networks for visuomotor behavior. *Neuron*, 81(6):1328–1343, 2014.

Alexandre Pouget, Jeffrey M Beck, Wei Ji Ma, and Peter E Latham. Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, 16(9):1170–1178, 2013.

Robert Prevedel, Young-Gyu Yoon, Maximilian Hoffmann, Nikita Pak, Gordon Wetzstein, Saul Kato, Tina Schrödel, Ramesh Raskar, Manuel Zimmer, Edward S Boyden, et al. Simultaneous whole-animal 3d imaging of neuronal activity using light-field microscopy. *Nature Methods*, 11(7):727–730, 2014.

Lawrence R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Adrian E Raftery and Steven Lewis. How many iterations in the Gibbs sampler? *Bayesian Statistics*, pages 763–773, 1992.

Rajesh Ranganath, Sean Gerrish, and David M Blei. Black box variational inference. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 33:275––283, 2014.

Rajesh P. N. Rao. Bayesian computation in recurrent neural circuits. *Neural Computation*, 16(1):1–38, January 2004.

Rajesh P. N. Rao. Neural models of Bayesian belief propagation. In *Bayesian brain: Probabilistic approaches to neural computation*, pages 236–264. MIT Press Cambridge, MA, 2007.

Rajesh P. N. Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1): 79–87, January 1999.

Danilo J Rezende, Daan Wierstra, and Wulfram Gerstner. Variational learning for recurrent spiking networks. *Advances in Neural Information Processing Systems*, pages 136–144, 2011.

Fred Rieke, David Warland, Rob de Ruyter van Steveninck, and William Bialek. *Spikes: exploring the neural code*. MIT press, 1999.

Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

Dan Roth. On the hardness of approximate reasoning. *Artificial Intelligence*, 82(1):273–302, 1996.

Maneesh Sahani. *Latent variable models for neural data analysis*. PhD thesis, California Institute of Technology, 1999.

Maneesh Sahani and Peter Dayan. Doubly distributional population codes: simultaneous representation of uncertainty and multiplicity. *Neural Computation*, 2279:2255–2279, 2003.

Joshua R Sanes and Richard H Masland. The types of retinal ganglion cells: current status and implications for neuronal classification. *Annual Review of Neuroscience*, 38:221–246, 2015.

Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4: 639–650, 1994.

Ben Shababo, Brooks Paige, Ari Pakman, and Liam Paninski. Bayesian inference and online experimental design for mapping neural microcircuits. *Advances in Neural Information Processing Systems*, pages 1304–1312, 2013.

Vahid Shalchyan and Dario Farina. A non-parametric Bayesian approach for clustering and tracking non-stationarities of neural spikes. *Journal of Neuroscience Methods*, 223: 85–91, 2014.

Lei Shi and Thomas L Griffiths. Neural implementation of hierarchical Bayesian inference by importance sampling. *Advances in Neural Information Processing Systems*, 2009.

Yousheng Shu, Andrea Hasenstaub, and David A McCormick. Turning on and off recurrent balanced cortical activity. *Nature*, 423(6937):288–293, 2003.

Jack W Silverstein. The spectral radii and norms of large dimensional non-central random matrices. *Stochastic Models*, 10(3):525–532, 1994.

Aleksandr Simma and Michael I Jordan. Modeling events with cascades of Poisson processes. *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2010.

Eero P Simoncelli. Optimal estimation in sensory systems. *The Cognitive Neurosciences, IV*, 2009.

Anne C Smith and Emery N Brown. Estimating a state-space model from point process observations. *Neural Computation*, 15(5):965–91, May 2003.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, pages 2951–2959, 2012.

Sen Song, Kenneth D Miller, and Lawerence F Abbott. Competitive Hebbian learning through spike-timing-dependent synaptic plasticitye. *Nature Neuroscience*, 3(9):919–26, September 2000. ISSN 1097-6256.

Daniel Soudry, Suraj Keshri, Patrick Stinson, Min-hwan Oh, Garud Iyengar, and Liam Paninski. Efficient ”shotgun” inference of neural connectivity from highly sub-sampled activity data. *PLoS Computational Biology*, 11(10):1–30, 10 2015. doi: 10.1371/journal.pcbi.1004464.

Olaf Sporns, Giulio Tononi, and Rolf Kötter. The human connectome: a structural description of the human brain. *PLoS Computational Biology*, 1(4):e42, 2005.

Olav Stetter, Demian Battaglia, Jordi Soriano, and Theo Geisel. Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals. *PLoS Computational Biology*, 8(8):e1002653, 2012.

Ian Stevenson and Konrad Koerding. Inferring spike-timing-dependent plasticity from spike train data. *Advances in Neural Information Processing Systems*, pages 2582–2590, 2011.

Ian H Stevenson, James M Rebesco, Nicholas G Hatsopoulos, Zach Haga, Lee E Miller, and Konrad P Körding. Bayesian inference of functional connectivity and network structure from spikes. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 17(3):203–213, 2009.

Alan A Stocker and Eero P Simoncelli. Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9(4):578–85, April 2006.

Yee Whye Teh and Michael I Jordan. Hierarchical Bayesian nonparametric models with applications. *Bayesian Nonparametrics*, pages 158–207, 2010.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581, 2006.

Joshua B Tenenbaum, Thomas L Griffiths, and Charles Kemp. Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7):309–318, 2006.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011.

Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

Wilson Truccolo, Uri T. Eden, Matthew R. Fellows, John P. Donoghue, and Emery N. Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of Neurophysiology*, 93(2):1074–1089, 2005. doi: 10.1152/jn.00697.2004.

Philip Tully, Matthias Hennig, and Anders Lansner. Synaptic and nonsynaptic plasticity approximating probabilistic inference. *Frontiers in Synaptic Neuroscience*, 6(8), 2014.

Srini Turaga, Lars Buesing, Adam M Packer, Henry Dalgleish, Noah Pettit, Michael Hausser, and Jakob Macke. Inferring neural population dynamics from multiple partial recordings of the same neural circuit. *Advances in Neural Information Processing Systems*, pages 539–547, 2013.

Leslie G Valiant. *Circuits of the Mind*. Oxford University Press, Inc., 1994.

Leslie G Valiant. Memorization and association on a realistic neural model. *Neural Computation*, 17(3):527–555, 2005.

Leslie G Valiant. A quantitative theory of neural computation. *Biological Cybernetics*, 95 (3):205–211, 2006.

Jurgen Van Gael, Yunus Saatci, Yee Whye Teh, and Zoubin Ghahramani. Beam sampling for the infinite hidden Markov model. *Proceedings of the International Conference on Machine Learning*, pages 1088–1095, 2008.

Michael Vidne, Yashar Ahmadian, Jonathon Shlens, Jonathan W Pillow, Jayant Kulkarni, Alan M Litke, EJ Chichilnisky, Eero Simoncelli, and Liam Paninski. Modeling the impact of common noise inputs on the network activity of retinal ganglion cells. *Journal of Computational Neuroscience*, 33(1):97–121, 2012.

Joshua T Vogelstein, Brendon O Watson, Adam M Packer, Rafael Yuste, Bruno Jedynak, and Liam Paninski. Spike inference from calcium imaging using sequential Monte Carlo methods. *Biophysical Journal*, 97(2):636–655, 2009.

Joshua T Vogelstein, Adam M Packer, Timothy A Machado, Tanya Sippy, Baktash Babadi, Rafael Yuste, and Liam Paninski. Fast nonnegative deconvolution for spike train inference from population calcium imaging. *Journal of Neurophysiology*, 104(6):3691–3704, 2010.

Hermann von Helmholtz and James Powell Cocke Southall. *Treatise on Physiological Optics: Translated from the 3rd German Ed.* Optical Society of America, 1925.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

Yair Weiss, Eero P Simoncelli, and Edward H Adelson. Motion illusions as optimal percepts. *Nature Neuroscience*, 5(6):598–604, 2002.

Mike West, P Jeff Harrison, and Helio S Migon. Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association*, 80(389):73–83, 1985.

John G White, Eileen Southgate, J Nichol Thomson, and Sydney Brenner. The structure of the nervous system of the nematode Caenorhabditis elegans: the mind of a worm. *Philosophical Transactions of the Royal Society of London: Series B (Biological Sciences)*, 314:1–340, 1986.

Louise Whiteley and Maneesh Sahani. Attention in a Bayesian framework. *Frontiers in Human Neuroscience*, 6, 2012.

Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta. Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.

Jesse Windle, Nicholas G Polson, and James G Scott. Sampling Pólya-gamma random variates: alternate and approximate techniques. *arXiv preprint arXiv:1405.0506*, 2014.

Frank Wood and Michael J Black. A nonparametric Bayesian alternative to spike sorting. *Journal of Neuroscience Methods*, 173(1):1–12, 2008.

Frank Wood, Jan Willem van de Meent, and Vikash Mansinghka. A new approach to probabilistic programming inference. *arXiv preprint arXiv:1507.00996*, 2015.

Tianming Yang and Michael N Shadlen. Probabilistic reasoning by neurons. *Nature*, 447 (7148):1075–80, June 2007.

Byron M. Yu, John P. Cunningham, Gopal Santhanam, Stephen I. Ryu, Krishna V. Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of Neurophysiology*, 102:614–635, 2009.

Alan Yuille and Daniel Kersten. Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10(7):301–308, 2006.

Richard S Zemel, Peter Dayan, and Alexandre Pouget. Probabilistic interpretation of population codes. *Neural Computation*, 10(2):403–30, February 1998.

Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 16, 2013.

Mingyuan Zhou, Lingbo Li, Lawrence Carin, and David B Dunson. Lognormal and gamma mixed negative binomial regression. *Proceedings of the International Conference on Machine Learning*, pages 1343–1350, 2012.