Chapter 5: Approximate inference

Exact inference is typically intractable. Can the brain get close to the right answer? This chapter discusses two classes of approximation. Sampling approximations harness randomness, achieving asymptotic correctness in the limit of many samples. Sampling offers one functional explanation for the ubiquity of noise in the brain. Variational approximations replace complex posteriors with simpler parametric forms, converting inference into a more tractable optimization problem (minimizing free energy). Under some assumptions, a variational approximation can be implemented using a hierarchical architecture in which feedback signals convey predictions and feedforward signals convey prediction errors.

To get a sense of what makes inference hard, consider the following everyday problem. You get up at night and look for the light switch in the dark. Your room is filled with dimly illuminated contours, and your brain's job is to identify contours that look like light switches. We can think of a contour as a collection of adjacent edges that vary smoothly over space in a one-dimensional pattern (see Figure 1). Thus, the inference problem is potentially high-dimensional: the brain has to infer the edge orientation at each point in space (the orientation field). In principle, this problem could be parallelized if we assume that each orientation is independent, but this violates the key assumption that the edges covary with their neighbors. There is no escaping the exponentially large number of possible orientation fields.

Space here should be understood as *retinotopic space*, a 2D map where each position corresponds to a location on the retina.

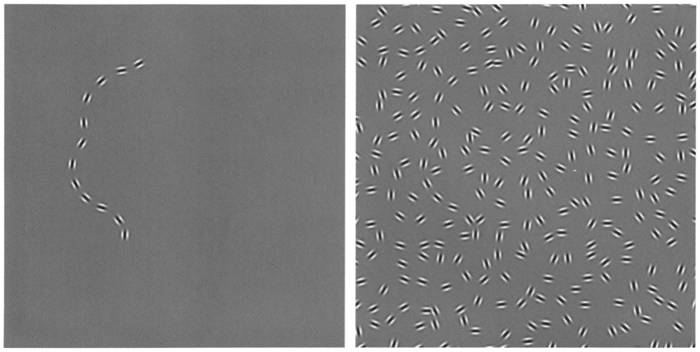

Figure 1: **Contour detection**. (Left) A contour defined by a set of Gabor patches. (Right) The same contour embedded in "noise" (randomly oriented) patches. The contours are identifiable by humans as long as the elements are not too far apart and the contour is relatively smooth. Reproduced from Field et al. (1993).

Recall our problem: we want to compute the posterior $p(s|x)$ over hidden state s given data x. As stipulated by Bayes' rule,

$$p(s|x) = \frac{p(x|s)p(s)}{\sum_{s'} p(x|s')p(s')} \tag{1}$$

In the contour detection example, x is an image and s is the orientation at each location in the image (we will shortly formalize this in a more biologically plausible way). Typically, the easy part is evaluating the numerator of Bayes' rule for any particular value of s, since we are assuming access to a computationally tractable joint distribution, $p(x, s) = p(x|s)p(s)$. The hard part is the denominator (the marginal likelihood), which involves summing (or integrating, in the case of continuous variables) over all possible states. When states are multi-dimensional, we run into the *curse of dimensionality* (Bellman, 1957): there is an exponentially large number of states, making marginalization through exhaustive enumeration intractable. For example, if there are M possible orientations and N locations, then the number of possible orientation fields is M^N.

Many other problems have a similar nature due to the coupling of many variables—solving crossword puzzles and segmenting events, to name a few. Nonetheless, there is a glimmer of hope due to the underlying structure of these problems. It is precisely the coupling that constrains inference. Algorithms that make efficient use of this structure can render inference tractable—albeit approximate.

In this chapter, we delve into two classes of approximation: sampling (also known as *Monte Carlo*) and variational algorithms. It's no coincidence that these are widely used in machine learning, statistics, and physics. They are the most effective engineering tools we have for tackling difficult inference problems. Intriguingly, nature may have hit upon similar solutions.

1 *Sampling approximations*

The idea behind sampling is to replace exhaustive enumeration of the hidden states with a set of K samples, $\{s^1, \dots, s^K\}$, drawn from $p(s|x)$:

$$p(s|x) \approx \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}[s^k = s], \tag{2}$$

For continuous variables, replace $\mathbb{I}[s^k = s]$ with the Dirac delta function, $\delta(s^k - s)$.

where $\mathbb{I}[\cdot] = 1$ if its argument is true, and 0 otherwise. Eq. 2 says that the probability of state s is approximately equal to the proportion of samples with that value. As K gets larger, the approximation gets increasingly accurate.

This is essentially the same idea as a histogram in data analysis.

The rub is that generating samples from the posterior is nontrivial. One general strategy is to generate samples from a dynamical system that can be proven to converge to the posterior. When the dynamical system is characterized by a transition probability $T(s'|s)$ that doesn't depend on any previous samples, it is called a *Markov*

For a general introduction to MCMC algorithms, see MacKay (2003).

chain, and the sampling algorithm is a form of *Markov chain Monte Carlo* (MCMC).

1.1 Markov chain Monte Carlo

How do we guarantee that a particular Markov chain converges to the posterior? A sufficient (but not necessary) condition is known as *detailed balance*:

$$\frac{T(s'|s)}{T(s|s')} = \frac{p(s'|x)}{p(s|x)}. \tag{3}$$

We can construct a versatile family of MCMC algorithms using the following construction. First, we factor the transition probability into a *proposal distribution* $G(s'|s)$ and an *acceptance distribution* $A(s'|s)$:

$$T(s'|s) = G(s'|s,x)A(s'|s). \tag{4}$$

Plugging this into Eq. 3, we get:

$$\frac{A(s'|s)}{A(s|s')} = \frac{p(s'|x)G(s|s',x)}{p(s|x)G(s'|s,x)}. \tag{5}$$

Finally, we need to specify an acceptance distribution that satisfies this equation. The classical Metropolis-Hastings algorithm uses

$$A(s'|s) = \min\left[1, \frac{p(s'|x)G(s|s',x)}{p(s|x)G(s'|s,x)}\right]. \tag{6}$$

We're still left with the problem that the acceptance distribution depends on the unknown posterior. However, notice that the posterior only enters as a ratio between $p(s'|x)$ and $p(s|x)$. Using Bayes' rule, we can write this ratio as:

$$\frac{p(s'|x)}{p(s|x)} = \frac{p(x|s')p(s')}{p(x|s)p(s)}. \tag{7}$$

This is much easier to evaluate, because (as stated above) typically we can tractably compute the likelihood $p(x|s)$ and prior $p(s)$ for a given sample.

1.2 Gibbs sampling

An important special case of the Metropolis-Hastings algorithm is *Gibbs sampling*, where the proposal distribution is the distribution over part of the state space conditional on the rest of the state space. It's easiest to think about this in the case where partition of the state space corresponds to single "sites" (state features). In the contour detection example, this means sampling a new orientation at location n conditional on the inferred orientations at all the other locations (denoted $s_{/n}$). Let s'_n denote a copy of s with only site n modified. The proposal distribution can then be formalized as follows:

$$G(s'|s,x) = \sum_n p(n)p(s'_n|s_{/n},x), \tag{8}$$

where $p(n)$ is the probability of a modification at site n. At each iteration, a single site is selected from $p(n)$ and then the modification is sampled from $p(s'_n|s_{/n},x)$. The Gibbs proposal is always accepted (to see this, plug the proposal distribution into Eq. 6).

We now bring these ideas back to neural computation, using the contour detection problem as an example. We can formalize the contour detection problem in the following way. Let $s_n \in [0,2\pi]$ denote the orientation at location n. The state vector s is the orientation field. In natural images, edges at nearby locations tend to have similar orientations (Sigman et al., 2001). We therefore impose a smoothness prior on the orientation field:

Sensitivity to smooth contours in human perception is exemplified by the Gestalt *Law of good continuation* (Wertheimer, 1938), which states that the visual system tends to prefer smooth over non-smooth contours. This preference has been characterized quantitatively by psychophysical experiments (Field et al., 1993).

$$p(s) \propto \exp\left[\sum_n \sum_m H_{nm}\cos(s_n - s_m)\right], \tag{9}$$

where $H_{nm} > 0$ if locations n and m are neighbors (0 otherwise). We combine this prior with sensory data $x = (x_1, \ldots, x_D)$, the spike counts of Poisson neurons with tuning curves $\{f_d(s)\}$. We will assume spatially localized cosine tuning functions (introduced in Chapter 3):

$$f_d(s_n) = \exp\left[\frac{1}{\nu}\cos(s_n - s^*_{dn})\right], \tag{10}$$

where s^*_{dn} is the preferred orientation for neuron d at location n (note that neuron d does not respond to inputs at other locations), and ν is the tuning width. Plugging this into the Poisson likelihood (see Chapter 4), we get:

$$p(x|s) \propto \exp\left[\frac{1}{\nu}\sum_d x_d \sum_n \cos(s_n - s^*_{dn})\right], \tag{11}$$

where (as in the last chapter) we have made use of the assumption that $\sum_d f_d(s)$ is a constant (satisfied if the tuning functions are shifted copies of one another and tile the state space).

Putting the pieces together, the posterior is given by:

$$\begin{aligned} p(s|x) &\propto p(x|s)p(s) \\ &\propto \exp\left[\frac{1}{\nu}\sum_d x_d \sum_n \cos(s_n - s^*_{dn}) + \sum_n \sum_m H_{nm}\cos(s_n - s_m)\right]. \end{aligned} \tag{12}$$

We can apply Gibbs sampling by iterating over locations and choosing new orientations conditional on the orientations at the other locations:

$$p(s_n|s_{/n}, x) \propto \exp\left[\frac{1}{\nu}\sum_d x_d \cos(s_n - s^*_{dn}) + \sum_m H_{nm}\cos(s_n - s_m)\right], \tag{13}$$

where d sums over all input neurons tuned to location n. If (as in the last chapter) we discretize the orientations into $\{\tilde{s}_j\}$, we can calculate the normalizing constant of this distribution and tractably sample from it. We can see that this is just another example of the softmax equation from the last chapter.

We can now see more clearly how to map this onto a neural circuit. Let $z_d(t) \in \{0, 1\}$ denote the spike train of input neuron d. Output neurons are indexed both by location (n) and orientation (j). Each output neuron integrates input spikes linearly, along with "lateral" contributions from other output neurons whose spike trains are presented by $\{y_{mj}(t)\}$:

$$I_{nj}(t) = \frac{1}{\nu}\sum_d z_d(t)\cos(s_n - s^*_{dn}) + \sum_m H_{nm}\cos(s_n - s_m)y_{mj}(t), \tag{14}$$

where again d sums over input neurons tuned to location n. As in the last chapter, we model the membrane potential $\mu_{nj}(t)$ as a perfect integrator, $C\dot{\mu}_{nj}(t) = I_{nj}(t)$, that is exponentiated to produce the intensity function (expected firing rate of a Poisson process). Finally, we assume that the neuron receives feedback inhibition reflecting the total activity of neurons tuned to the same location but different orientations. This yields a softmax intensity function:

$$\rho_{nj}(t) = \frac{\exp[\mu_{nj}(t)]}{\sum_{j'}\exp[\mu_{nj'}(t)]}. \tag{15}$$

From the MCMC perspective, we can view this circuit as implementing a stochastic dynamical system that converges to the posterior, so that the spikes of the output neurons can be viewed as posterior samples.

See Buesing et al. (2011) and Pecevski et al. (2011) for a description of spiking neuron circuits that implement a form of Gibbs sampling for a more general class of probabilistic models.

An example of the model's behavior is shown in Figure 2. Neural Gibbs sampling effectively denoises a contour hidden in a smoothly varying orientation field.

1.3 *Langevin sampling*

While Gibbs sampling is a powerful algorithm, it has the drawback that convergence to the posterior is often slow due to the fact that moves through the state space are small. What we'd ideally like is to

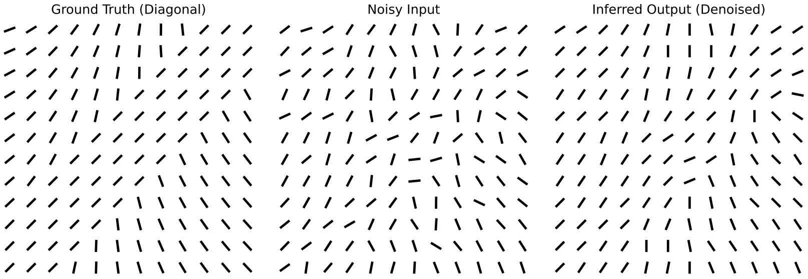

Figure 2: **Contour detection simulation**. (Left) Ground truth contour. (Middle) Noisy observations. (Right) Inferred contour using neural Gibbs sampling.

quickly move towards high probability states. Langevin sampling, another MCMC algorithm, achieves this using gradients. The core of Langevin sampling is the following dynamical system:

$$ds = \nabla_s \log p(s(t)|x)\, dt + \sqrt{2}\sigma\, dW(t), \tag{16}$$

where $s(t)$ denotes the state sampled at time t, and $W(t)$ is a Wiener process, which produces independent, Gaussian-distributed increments; the standard deviation parameter σ scales the increments. In other words, samples of the state are generated in continuous time by following the posterior gradient corrupted by white noise. It can be shown that these dynamics converge to a stationary distribution, $p(s|x)$. Thus, we can generate posterior samples by following a noisy gradient.

Continuing our example from the previous section, x represents the spike counts from an input population of Poisson neurons. Under this assumption, the gradient is given by:

$$\nabla_s \log p(s|x) = \sum_d \frac{x_d \nabla_s f_d(s)}{f_d(s)} + \frac{\nabla_s p(s)}{p(s)}. \tag{17}$$

We can map this equation onto a neural circuit by using the same probabilistic assumptions as in the previous section, and positing an output neuron for location n with perfect integration and noisy membrane potential dynamics (see Chapter 2):

$$C\dot{\mu}_n = I_n(t) + \sqrt{2}\sigma\, dW(t). \tag{18}$$

Continuing the contour detection example, the input current is given by:

The constant disappears under the gradient dynamics.

$$\begin{aligned} I_n(t) &= \sum_d \frac{z_d(t)}{f_d(s)} \frac{\partial}{\partial s_n} f_d(s) + \frac{1}{p(s)} \frac{\partial}{\partial s_n} p(s) \\ &= -\frac{1}{\nu} \sum_d z_d(t) \sin(s_n - s^*_{dn}) - \sum_m H_{nm} y_m(t) \sin(s_n - s_m) + \text{const.} \end{aligned} \tag{19}$$

Notice that we are now interpreting the membrane potential values (rather than the spikes) as the posterior samples (see Orbán et al., 2016).

Also notice that we are no longer discretizing the orientation space, instead treating it as a continuous variable.

How, then, should we interpret spikes? The probability of generating a spike within some infinitesimally small window of time (the instantaneous spike probability) is the probability of the membrane potential crossing the spiking threshold θ. From a sampling perspective, it's the proportion of samples above the threshold, an approximation of the posterior probability that $s_n > \theta$. This kind of representation is useful for detection tasks. If the task is to respond whenever $s_n > \theta$, then a response circuit need only listen to the activity of neurons with thresholds near θ. If the decision criterion is changed, then the response circuit can listen to a different set of neurons. Alternatively, if the task is to report the probability that s_n lies within some interval $[a, b]$, then a response circuit can compute the difference between neurons with thresholds near b and a.

We can also use the same representation to estimate the posterior mean. For non-negative variables ($s_n \geq 0$), the tail integral formula states that:

$$\mathbb{E}[s_n|x] = \int_0^\infty p(s_n > \theta|x)d\theta. \tag{20}$$

If we have output neurons with a fine enough range of thresholds, then we can approximate the expectation by simply counting the number of output neurons generating a spike within some short interval of time.

1.4 Neural evidence for sampling

Many models view neural noise (whatever its locus and origin) as a "fact of life" which must be mitigated for the purposes of computation. Sampling models offer a fundamentally different point of view—noise is a feature, not a bug (Maass, 2014). Hoyer and Hyvärinen (2002) were the first to suggest that neural variability might reflect posterior sampling, but their arguments were not strongly supported by data available at the time. Since then, new data have offered more direct tests of their hypothesis.

When testing such a general hypothesis, it's important to start with predictions that generalize across many different versions of the hypothesis (e.g., Gibbs, Langevin, etc.). Here are three general predictions (see Orbán et al., 2016, for more details).

First, "spontaneous" neural activity prior to stimulus onset should reflect samples from the prior, whereas stimulus-evoked activity should reflect samples from the posterior. Generally, the posterior will be narrower (lower variance) than the prior. Under the sampling hypothesis, neural variability should increase monotonically with posterior variance. Intuitively, this is because when variance is high the samples must explore a broader range of states. This implies that stimulus onset should "quench" neural variability, as observed experimentally (Figure 3).

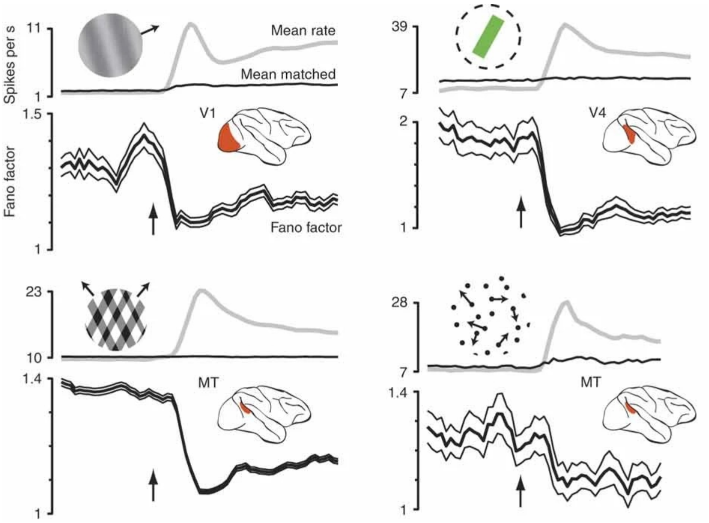

Figure 3: **Stimulus onset quenches neural variability**. Each panel shows experiments with a different stimulus and brain area. Fano factor is the ratio of the variance to the mean. Here it was computed for "mean-matched" activity traces, to avoid the confound of mean firing rate. Adapted from Churchland et al. (2010).

A second general prediction is that experimentally increasing uncertainty (e.g., by reducing stimulus contrast) should produce a corresponding increase in neural variability. Consistent with this prediction, neural variability in primary visual cortex (V1) is higher for low contrast stimuli (Figure 4).

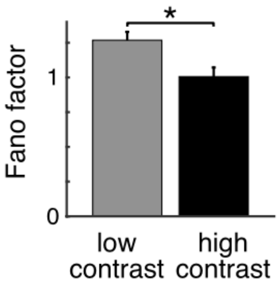

Figure 4: **Stimulus contrast decreases neural variability**. Adapted from Orbán et al. (2016).

A third general prediction is that stimulus-evoked activity, when averaged across trials, should resemble spontaneous activity. This follows from the fact that the expected posterior is just the prior:

$$\mathbb{E}[p(s|x)] = \sum_x p(x)p(s|x) = p(s). \tag{21}$$

Critically, this is only true if the stimulus distribution $p(x)$ reflects the "natural statistics" of stimuli in the real world, assuming that the brain's prior, $p(s)$, is adapted to these statistics. In other words, $p(x)$ and $p(s)$ must be related through $p(x) = \sum_s p(s)p(x|s)$. It's possible to break this relationship by presenting artificial stimuli like noise patterns or gratings. Orbán et al. (2016) confirmed this prediction for V1 by showing that the distribution of spontaneous activity was much more similar to the distribution of average stimulus-evoked activity for natural images compared to artificial images.

See also Berkes et al. (2011).

1.5 Behavioral evidence for sampling

The sampling hypothesis also makes predictions about behavior. As with the neural data, we focus on predictions that generalize across different versions of the hypothesis.

As pointed out by Hoyer and Hyvärinen (2002), perceptual multistability seems like a hallmark of sampling. Perceptual multistability arises when the brain switches repeatedly (and usually stochastically) between different interpretations of the same visual input. For example, ambiguous images like the Necker cube and the face-vase illusion (Figure 5) can be interpreted in different ways by the same observer over a short interval of time. Another example is binocular rivalry: when different images are presented to each eye, typically only one image is perceived at a time, with stochastic switches between the dominant image.

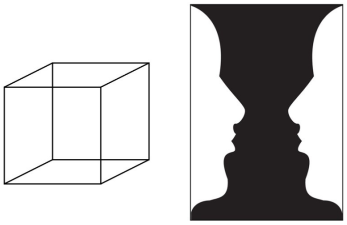

Figure 5: **Ambiguous images**. (Left) The Necker cube. (Right) The face-vase illusion. For both images, the interpretation will switch spontaneously if you look at it for long enough.

Gershman et al. (2012) used the sampling hypothesis to account for a number of subtle phenomena in binocular rivalry. One phenomenon is that switches are not always all-or-none; in particular, large images in binocular rivalry experiments produce "piecemeal" switches, where one part of the image switches before other parts (O'Shea et al., 1997). This makes sense if sampling is operating at the level of image parts (whatever those might be). For larger images, the dependencies between different parts become weaker, allowing piecemeal switches. The piecemeal nature of multistability is also revealed by the observation that for certain stimuli switching manifests as a traveling wave (Wilson et al., 2001). For example, if the two images in a binocular rivalry experiment are radial and concentric gratings, a transient increase in image contrast at one location on the non-dominant grating will produce a wave-like switching pattern, similar to lighting a fuse and watching it progressively ignite adjacent regions (Figure 6). The wave pattern can be quantified by asking observers to report when they perceived a switch at particular locations on the grating. The propagation is slower for concentric gratings, consistent with a smoothness prior that favors collinear edge orientations (as discussed above). The propagation can also be slowed down by introducing gaps in the grating. All of these results agree with a model in which the dynamics of sampling is governed by the stimulus structure.

Using brain imaging, Lee et al. (2005) showed that wave propagation during binocular rivalry is measurable in primary visual cortex.

Another subtle aspect of binocular rivalry is that sometimes the images fuse rather than rival. Several factors determine whether fusion or rivalry occurs. Fusion is more likely when both images are low contrast (Burke et al., 1999) and when they are similar (e.g., two gratings with slightly different orientations; Knapen et al., 2007). Under the sampling hypothesis, fusion arises when the two posterior

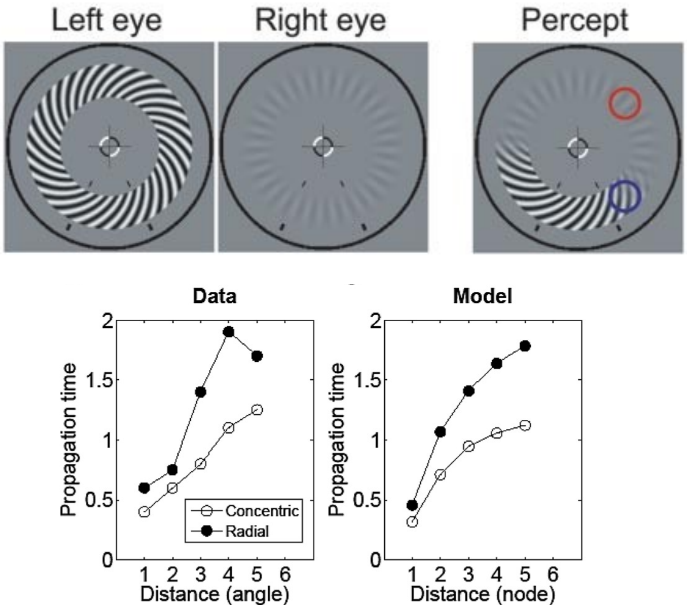

Figure 6: **Traveling waves in binocular rivalry**. (Top) Binocular stimulus and percept (from Lee et al., 2005). The low contrast stimulus appeared to spread around the annulus after being ignited by a contrast change at the red location. The blue location shows an example probe location. (Bottom) Estimated propagation time (from Wilson et al., 2001) and simulations (from Gershman et al., 2012).

modes are not well-separated, either because their variances are large (low contrast) or because their modes are near one another (high similarity). In this case, sampling doesn't bounce as much between the two modes, but instead spends more time in the high density area between them.

Outside of visual perception, the sampling hypothesis has been used to explain a wide range of cognitive phenomena. For example, when people are asked to estimate an unknown quantity (e.g., "In what year was Beethoven born?") after being primed with an irrelevant quantity (e.g., the last 4 digits of their social security number), they tend to anchor their estimates on the irrelevant quantity, adjusting insufficiently away from it (Tversky and Kahneman, 1974). This "anchoring and adjustment" heuristic arises naturally from an MCMC algorithm that is initialized at the irrelevant quantity and then generates only a small number of samples (Dasgupta et al., 2017; Lieder et al., 2018a).

See Sanborn and Chater (2016) for a review of the sampling hypothesis applied to human cognition.

The number of samples may be adaptive. As shown by Lieder et al. (2018b), people do more adjustment when the costs of time are low and the costs of errors are large.

Another cognitive phenomenon explained by sampling is the *unpacking effect*. When people are asked to judge the probability of dying from a natural cause (the packed condition), their average answer was 0.58, but when asked to separately judge the probability of dying from heart disease, cancer, or some other natural cause (the unpacked condition), the summed probabilities equaled 0.73 (Tversky and Koehler, 1994). Mathematically, these should be the same, since the marginal probability of dying from a natural cause is the sum of the conditional probabilities of dying from particular natural causes.

According to the sampling hypothesis, unpacking to typical examples like heart disease and cancer ensure that these contribute to the sample set and thus to probability judgments. In contrast, people are left to generate all the samples on their own in the packed condition, and therefore can potentially miss some if they are generating a small number of samples. In contrast, unpacking to atypical examples (e.g., pneumonia, diabetes) leads to marginal probabilities that are *lower* than in the packed condition (Sloman et al., 2004). According to the sampling hypothesis, this happens because the sampler gets stranded in a lower-probability region of the state space and has trouble recovering from this with limited samples. These effects are amplified by time pressure and cognitive load (Dasgupta et al., 2017), consistent with a reduction in the number of samples.

2 *Variational approximations*

MCMC algorithms are asymptotically correct: if you run them long enough, you'll approximate the posterior to an arbitrary degree of precision. However, you might need to run them a long time if the problem is complex. An alternative approach is to use an approximation algorithm that produces an answer more quickly, but doesn't enjoy asymptotic correctness. Variational approximations offer a general framework for doing this.

2.1 *Free energy minimization*

The basic idea is to turn inference into a constrained optimization problem. The goal is to find an approximate posterior $q \in \mathcal{Q}$ that gets closest to the posterior, where $\mathcal{Q}$ is a constrained family of probability distributions. This family should be chosen in such a way that both finding and evaluating q is relatively fast. More precisely, the optimization problem is defined as:

$$q^* = \operatorname*{argmin}_{q \in \mathcal{Q}} \mathcal{D}[q(s|x)||p(s|x)], \tag{22}$$

where $\mathcal{D}[q(s|x)||p(s|x)]$ is the Kullback-Leibler (KL) divergence between the approximate and exact posteriors:

Recall from Chapter 4 that we used the same KL divergence to define an objective function for resource-rational belief updating.

$$\mathcal{D}[q(s|x)||p(s|x)] = \sum_s q(s|x) \log \frac{q(s|x)}{p(s|x)}. \tag{23}$$

The KL divergence is always non-negative, and equals 0 when $q(s|x) = p(s|x)$.

The basic problem with Eq. 22 is that the KL divergence can't be optimized directly, since it is a functional of the posterior—precisely the thing we are trying to approximate. There is an alternative way of formulating this optimization problem, which turns out to be equivalent:

$$q^* = \operatorname*{argmin}_{q \in \mathcal{Q}} \mathcal{F}[q(s|x)], \tag{24}$$

where $\mathcal{F}[q(s|x)]$ is the *variational free energy*:

The terminology here derives from applications in physics.

$$\mathcal{F}[q(s|x)] = \sum_s q(s|x) \log \frac{q(s|x)}{p(x,s)}. \tag{25}$$

This is related to the original optimization problem via the following equation:

$$\log p(x) = \mathcal{D}[q(s|x)||p(s|x)] - \mathcal{F}[q(s|x)], \tag{26}$$

where $\log p(x)$ is the *evidence* (the log marginal likelihood). The equality implies that decreasing free energy by some amount forces the KL divergence to also decrease by the same amount. Thus, minimizing free energy is thus equivalent to minimizing KL divergence, in the sense that the optimal approximate posterior is the same.

Because the KL divergence is non-negative, $-\mathcal{F}[q(s|x)]$ is a lower bound on the evidence, which is why it is sometimes known as the *evidence lower bound*.

The idea that the brain minimizes free energy—the *free energy principle*—has spawned a large literature investigating many different dimensions of this idea. It has even been proposed as a 'unified brain theory' (Friston, 2010) because it subsumes and generalizes several other general principles (see Chapter 3). Our purpose here is narrower: to understand how free energy minimization can be leveraged for tractable approximate inference.

See Parr et al. (2022) for a comprehensive survey.

2.2 *The Laplace and mean-field approximations*

We need to restrict the approximation family $\mathcal{Q}$ in some way that makes inference more tractable. One technique (applicable to models with continuous states) is to restrict $\mathcal{Q}$ to the set of Gaussian posteriors: $q(s|x) = \mathcal{N}(s; \hat{s}, \Sigma)$, where the mean and covariance are known as variational parameters. It's important to keep in mind that these are not latent variables that the brain is inferring, but rather part of the brain's computational machinery for approximate inference. In particular, the optimization problem is to find the variational parameters that minimize free energy.

Restricting to a Gaussian is not on its own sufficient, because one still can't compute the free energy—the integral over s is intractable in the general case, due to nonlinearities in the joint distribution $p(x,s)$. We can, however, obtain a tractable integral if we linearize $\log p(x,s)$ with a second-order Taylor series expansion around $\hat{s}$:

$$\log p(x,s) \approx \log p(x,\hat{s}) + (s-\hat{s})^\top \nabla_s \log p(x,\hat{s}) - \frac{1}{2}(s-\hat{s})^\top \Lambda (s-\hat{s}), \tag{27}$$

where $\Lambda = -\nabla_s \nabla_s \log p(x, \hat{s})$ is the Hessian (matrix of 2nd derivatives) of the negative log likelihood evaluated at $\hat{s}$. This is known as the *Laplace approximation*, which can then be used to analytically approximate the free energy:

$$\mathcal{F}[q(s|x)] \approx -\log p(x, \hat{s}) - \frac{1}{2}\mathrm{Tr}[\Lambda\Sigma] + \frac{1}{2}\log|\Sigma| + \text{const.} \tag{28}$$

This result can be derived using a standard formula for the Gaussian integral of a quadratic form, combined with the entropy of a multivariate Gaussian.

where $\mathrm{Tr}[\cdot]$ is the trace operator, and $|\cdot|$ is the matrix determinant. Setting the gradient of the free energy to 0 and solving for the variational parameters gives:

$$\hat{s} = \operatorname*{argmax}_s p(x, s) = \operatorname*{argmax}_s p(s|x) \tag{29}$$

$$\Sigma = \Lambda^{-1}. \tag{30}$$

In other words, the optimal mean is the posterior mode, and the optimal covariance is the inverse Hessian.

Computing the inverse Hessian is non-trivial, and it's not clear how this could be implemented neurally. A typical move is to adopt a *mean-field approximation*, where the approximate posterior factorizes:

$$q(s|x) = \prod_n q_n(s_n|x). \tag{31}$$

When combined with the Laplace approximation, the factorization leads to a diagonal covariance: $\Sigma = \mathrm{diag}(1/\lambda_1, \dots, 1/\lambda_N)$, where $\lambda_n = -\frac{\partial^2}{\partial s^2}\log p(x, \hat{s}_n)$. This is the *mean-field Laplace* approximation.

The posterior mode can be found using gradient ascent. This is essentially a deterministic version of Langevin sampling (no membrane potential noise). As in our analysis of Langevin sampling, we interpret the input current as the gradient of the log posterior, the membrane potential as an integrator of this gradient, and the spiking activity as samples from the "tail" distribution $p(s_n > \theta|x)$. Under the Laplace approximation, this distribution is given by:

$$p(s_n > \theta|x) \approx q(s_n > \theta|x) = \Phi\left(\sqrt{\lambda_n}(\hat{s}_n - \theta)\right), \tag{32}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard Gaussian distribution, and λ_n is the posterior precision. The tail probability is a sigmoidal function of $\hat{s}$, with an inflection point at θ (the point at which the probability crosses 0.5). The posterior precision controls the slope of the sigmoid: greater precision results in a steeper slope.

How can a neuron compute the precision? One way is to use the interpretation of the input current for neuron n as $I_n(t) = \frac{\partial}{\partial s_n}\log p(s(t), x)$. This implies that the precision is the negative partial derivative of the input current with respect to s_n. Using our contour detection example, we can differentiate Eq. 19 around the posterior mode to obtain:

$$\lambda_n(t) = \frac{1}{\nu}\sum_d z_d(t)\cos(\hat{s}_n - s^*_{dn}) + \sum_m H_{nm} y_m(t)\cos(\hat{s}_n - s_m), \qquad (33)$$

where we have expressed the precision as a time-dependent function to make clear that it is being dynamically computed by the postsynaptic neuron. Intuitively, precision is largest when the posterior mode is close to the preferred stimuli of the input neurons (high likelihood) and when it is smooth (high prior). Precision is a linear function of the input and lateral spikes; thus, it could be plausibly computed by linear synaptic integration. Because both the input current and the precision are linear functions of the presynaptic spikes, an intriguing possibility is that these variables are represented separately in different parts of the dendritic tree. For example, "distal" dendritic input (far from the cell body) tends to have relatively weak direct effects on firing rate compared to "proximal" dendritic input (near the cell body), but the distal inputs can modulate gain (Larkum et al., 2004)—the role of precision suggested by Eq. 32. Thus, it is plausible that $I_n(t)$ represents inputs to proximal dendrites, while $\lambda_n(t)$ represents the inputs to distal dendrites.

2.3 *Predictive coding*

If we assume a Gaussian noise model and prior, it becomes possible to parametrize the variational posterior in a different way. Specifically, let us assume the following generative model:

$$s \sim \mathcal{N}(\bar{s}, \Omega) \qquad (34)$$

$$x_d \sim \mathcal{N}(f_d(s), \omega). \qquad (35)$$

The Gaussian noise model is closely related to the Poisson noise model, since the Poisson distribution becomes increasingly Gaussian as the firing rate increases.

The posterior mode can be updated by following the gradient of the joint log likelihood:

$$\Delta\hat{s} \propto \nabla_s \log p(x, s) = \omega \sum_d [x_d - f_d(s)]\nabla_s f_d(s) - \Omega(\hat{s} - \bar{s}). \qquad (36)$$

This formulation invites us to think about a *predictive coding* architecture in which "prediction" neurons $y(t)$ reporting the inferred state, $\hat{s}$, receive input from "error" neurons reporting the difference between observed and expected signals. There are two kinds of error neurons. The "bottom-up" (or "feedforward") error neurons report the difference between sensory signals $z(t)$ and the expected firing rate under the inferred state, $f(\hat{s})$. The "top-down" (or "feedback") error neurons report the difference between the inferred state and the expected state under the prior distribution, $\bar{s}$.

Predictive coding concepts have a long history in neuroscience. One of the earliest versions of this idea was proposed for the retina by Srinivasan et al. (1982). A more general version of this hypothesis for the entire visual system was proposed by Rao and Ballard (1999), and subsequently developed as a general principle for cortical computation by Friston (2005) within the framework of free energy minimization.

In this section, we assume that all neurons are perfect integrators of synaptic inputs, and that the firing rate is linear in the membrane potential. We will work directly with these firing rates (ignoring spikes). The dynamics for the prediction neuron population activity $y(t)$ can be written compactly in vector form:

$$\dot{y} = \frac{1}{\omega} h(t) \nabla_{\hat{s}} f(\hat{s}) - g(t)\Omega^{-1}, \tag{37}$$

where $h(t)$ and $g(t)$ are the firing rates of bottom-up and top-down error coding neurons, respectively, with the following dynamics:

$$\dot{g} = y(t) - \bar{s} - g(t)\Omega \tag{38}$$

$$\dot{h} = z(t) - f_d(\hat{s}) - \omega h(t). \tag{39}$$

The fixed points of these dynamics are:

$$y(\infty) = \hat{s} \tag{40}$$

$$g(\infty) = \Omega^{-1}(\hat{s} - \bar{s}) \tag{41}$$

$$h(\infty) = \frac{1}{\omega}(x - f(\hat{s})). \tag{42}$$

Thus, the posterior mode can be found by running the dynamics of this system and then observing the activity of the prediction neurons at steady state.

We can iterate this architecture hierarchically (Figure 7), where the top-down signals receive inputs from higher-level neurons encoding expectations at the next level of the generative model (Rao and Ballard, 1999). This fits with the general idea that the brain learns generative models at multiple levels of abstraction, where each level defines a prior distribution on the level below (Lee and Mumford, 2003; Friston, 2008). The hierarchical architecture is thought to be implemented in the laminar (layered) structure of cortex (Bastos et al., 2012), where neurons in layers 2/3 convey feedforward signals (errors) to higher levels of the hierarchy, while neurons in layers 5/6 convey feedback signals (predictions) from higher to lower levels.

Closely related ideas have been studied in cognitive science (e.g., Kemp et al., 2007).

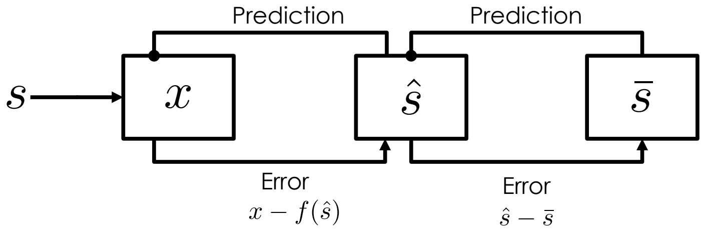

Figure 7: **Hierarchical predictive coding architecture**.

2.4 Evidence for predictive coding in the brain

A classical view of the brain, and of cortex in particular, posits a feedforward hierarchy of feature detectors (Rosenblatt, 1958; Fukushima, 1980; Marr, 1982). Each layer of detectors identifies more complex or abstract features based on the activity of the feature detectors in the layer below. For example, some models view retinal ganglion cells as light spot detectors, V1 neurons as oriented edge detectors, which feed into V2 contour detectors, which in turn feed into shape detectors in the lateral occipital cortex. While there's much to recommend such a view, it also seems to be missing something important. A few examples will illustrate this point.

Many kinds of neurons characterized as feature detectors have receptive fields with excitatory centers and inhibitory surrounds. What this means is that the neurons respond maximally when their preferred stimulus is presented but are inhibited by similar stimuli. For example, retinal ganglion cells are excited by spots of light presented at particular retinotopic locations; this response increases with the size of the spot, but starts to decrease when the spot gets to a certain size (Kuffler, 1953). Similarly, many neurons in V1 respond maximally to lines of a particular orientation, length, and location; if the line gets long enough, the response decreases—a phenomenon known as *endstopping* (Hubel and Wiesel, 1965). The interpretation of center-surround receptive field structure has been the subject of extensive theoretical speculation, but generally it has been challenging to come up with an interpretation that is general enough to encompass all the instances in which such tuning manifests (including motion and shape processing areas, among others).

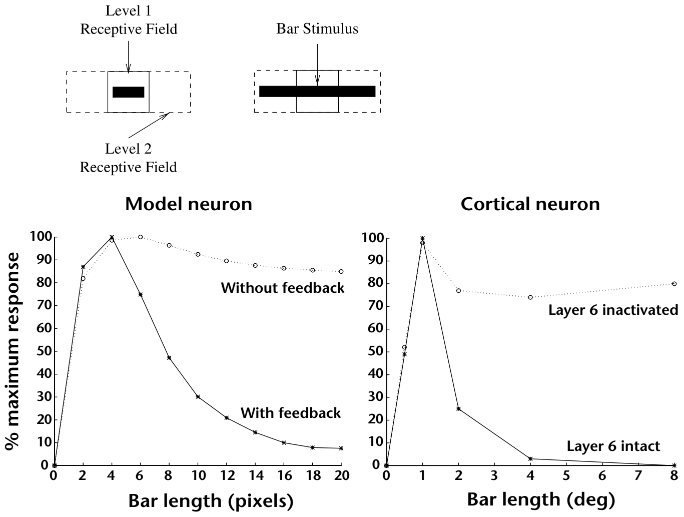

Figure 8: **Endstopping**. (Top) Receptive fields for "level 1" (V1) and "level 2" (V2) neurons. (Bottom) Activation as a function of bar length, with and without feedback. Adapted from Rao and Ballard (1999). The physiological data come from Sandell and Schiller (1982).

Predictive coding offers a normative account (Srinivasan et al., 1982; Rao and Ballard, 1999), interpreting neurons with center-surround receptive fields as error neurons, $g(t)$. The intuition is that error neurons will only respond to a stimulus when it can't be predicted by the activity of the prediction neurons. A small oriented edge is relatively unpredictable from the perspective of prediction neurons in V1, since there is no larger-scale spatial structure. However, a longer edge indicates a contour that extends beyond the receptive field of V1 prediction neurons, activating prediction neurons in higher visual areas (such as V2) that detect contours (Figure 8). The prediction neurons send suppressive feedback to V1 error neurons, thereby producing endstopping. Consistent with this hypothesis, inactivation of V2 strongly reduces endstopping (Sandell and Schiller, 1982; Nassi et al., 2013).

Suppressive effects of predictions on cortical activation have been observed in many experiments. The visual responses of V1 neurons in layers 2/3 (the feedforward pathway thought to convey errors) increase when novel images are presented, and these novelty responses decrease as the images are repeatedly presented (Homann et al., 2022). Similarly, visual responses of V1 neurons in layers 2/3 increase when an animal encounters unexpected disruptions in visual flow during locomotion (Keller et al., 2012). Higher-level visual areas become more active in response to images with coherent shape structure (compared to images with randomly assembled edges), and this is accompanied by *decreases* in the responses of lower-level regions (Figure 9). When predictive responses are identified in V1, these tend to originate in the deep layers thought to convey feedback from higher-level regions (Kok et al., 2016; Aitken et al., 2020).

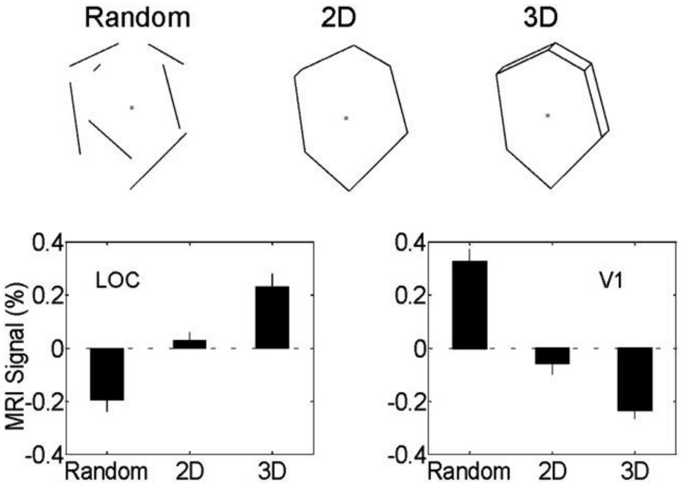

Figure 9: **Effects of visual structure on neural activity**. A high-level visual region (the lateral occipital complex, LOC) responds more to images with 3D structure compared to images with 2D and random structure. A low-level visual region (V1) has the opposite profile. Neural activity is measured here in humans using functional MRI. Adapted from Murray et al. (2002).

3 Conclusion

The brain has multiple biologically plausible options for approximate inference. These are not mutually exclusive. One possibility is that different algorithms are used by different parts of the brain, based on their complementary strengths and weaknesses for different tasks. For tasks requiring fast sensory processing, it may make sense to rely on primarily feedforward algorithms, whereas for tasks requiring context-sensitivity, it may make sense to rely on algorithms with recurrent dynamics and feedback. Another possibility is that these algorithms are integrated; for example, there are ways to use sampling methods in the service of variational inference and predictive coding (Oliviers et al., 2024). This may help solve the outstanding problem of how uncertainty is represented in predictive coding schemes.

The fact that evidence exists for all of these possibilities suggests that the complete picture is likely complex, not reducible to any single simple algorithm.

Study questions

1. Contrast Gibbs sampling and Langevin sampling in terms of computational effectiveness and biological plausibility.
2. What are the complementary strengths and weaknesses of sampling vs. variational approximations? How might the brain decide which to deploy in a given context?
3. The free energy principle has been proposed as a unified brain theory. Do you find that claim justified, or does the evidence suggest a patchwork of different strategies?

References

Aitken, F., Menelaou, G., Warrington, O., Koolschijn, R. S., Corbin, N., Callaghan, M. F., and Kok, P. (2020). Prior expectations evoke stimulus-specific activity in the deep layers of the primary visual cortex. *PLoS Biology*, 18:e3001023.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76:695–711.

Bellman, R. E. (1957). *Dynamic Programming*. Princeton University Press.

Berkes, P., Orbán, G., Lengyel, M., and Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331:83–87.

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

Burke, D., Alais, D., and Wenderoth, P. (1999). Determinants of fusion of dichoptically presented orthogonal gratings. *Perception*, 28:73–88.

Churchland, M. M., Yu, B. M., Cunningham, J. P., Sugrue, L. P., Cohen, M. R., Corrado, G. S., Newsome, W. T., Clark, A. M., Hosseini, P., Scott, B. B., et al. (2010). Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature Neuroscience*, 13:369–378.

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

Field, D. J., Hayes, A., and Hess, R. F. (1993). Contour integration by the human visual system: evidence for a local "association field". *Vision Research*, 33:173–193.

Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360:815–836.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology*, 4:e1000211.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11:127–138.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202.

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

Homann, J., Koay, S. A., Chen, K. S., Tank, D. W., and Berry, M. J. (2022). Novel stimuli evoke excess activity in the mouse primary visual cortex. *Proceedings of the National Academy of Sciences*, 119:e2108882119.

Hoyer, P. and Hyvärinen, A. (2002). Interpreting neural response variability as Monte Carlo sampling of the posterior. *Advances in Neural Information Processing Systems*, 15.

Hubel, D. H. and Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology*, 28:229–289.

Keller, G. B., Bonhoeffer, T., and Hübener, M. (2012). Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. *Neuron*, 74:809–815.

Kemp, C., Perfors, A., and Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, 10:307–321.

Knapen, T., Kanai, R., Brascamp, J., van Boxtel, J., and van Ee, R. (2007). Distance in feature space determines exclusivity in visual rivalry. *Vision Research*, 47:3269–3275.

Kok, P., Bains, L. J., Van Mourik, T., Norris, D. G., and de Lange, F. P. (2016). Selective activation of the deep layers of the human primary visual cortex by top-down feedback. *Current Biology*, 26:371–376.

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

Larkum, M. E., Senn, W., and Lüscher, H.-R. (2004). Top-down dendritic input increases the gain of layer 5 pyramidal neurons. *Cerebral Cortex*, 14:1059–1070.

Lee, S.-H., Blake, R., and Heeger, D. J. (2005). Traveling waves of activity in primary visual cortex during binocular rivalry. *Nature Neuroscience*, 8:22–23.

Lee, T. S. and Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, 20:1434–1448.

Lieder, F., Griffiths, T. L., M. Huys, Q. J., and Goodman, N. D. (2018a). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, 25:322–349.

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

Maass, W. (2014). Noise as a resource for computation and learning in networks of spiking neurons. *Proceedings of the IEEE*, 102:860–880.

MacKay, D. J. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

Marr, D. (1982). *Vision: A Computational Approach*. Freeman.

Murray, S. O., Kersten, D., Olshausen, B. A., Schrater, P., and Woods, D. L. (2002). Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences*, 99:15164–15169.

Nassi, J. J., Lomber, S. G., and Born, R. T. (2013). Corticocortical feedback contributes to surround suppression in V1 of the alert primate. *Journal of Neuroscience*, 33:8504–8517.

Oliviers, G., Bogacz, R., and Meulemans, A. (2024). Learning probability distributions of sensory inputs with Monte Carlo predictive coding. *PLOS Computational Biology*, 20:e1012532.

Orbán, G., Berkes, P., Fiser, J., and Lengyel, M. (2016). Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, 92:530–543.

O'Shea, R. P., Sims, A. J., and Govan, D. G. (1997). The effect of spatial frequency and field size on the spread of exclusive visibility in binocular rivalry. *Vision Research*, 37:175–183.

Parr, T., Pezzulo, G., and Friston, K. J. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press.

Pecevski, D., Buesing, L., and Maass, W. (2011). Probabilistic inference in general graphical models through sampling in stochastic networks of spiking neurons. *PLoS Computational Biology*, 7:e1002294.

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

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408.

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

Sandell, J. and Schiller, P. (1982). Effect of cooling area 18 on striate cortex cells in the squirrel monkey. *Journal of Neurophysiology*, 48:38–48.

Sigman, M., Cecchi, G. A., Gilbert, C. D., and Magnasco, M. O. (2001). On a common circle: natural scenes and Gestalt rules. *Proceedings of the National Academy of Sciences*, 98:1935–1940.

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

Srinivasan, M. V., Laughlin, S. B., and Dubs, A. (1982). Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 216:427–459.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185:1124–1131.

Tversky, A. and Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101:547–567.

Wertheimer, M. (1938). Laws of organization in perceptual forms. In Ellis, W., editor, *A Sourcebook of Gestalt Psychology*, pages 71–88. Harcourt, Brace.

Wilson, H. R., Blake, R., and Lee, S.-H. (2001). Dynamics of travelling waves in visual perception. *Nature*, 412:907–910.