Chapter 4: The Bayesian brain

The sensory data received by the brain provide incomplete and noisy information about the environment state. This chapter describes models of how the brain computes a probability distribution over (or point estimate of) hidden states. We immediately run into the problem that behavior seems to deviate from Bayes-optimal inference. These deviations can be understood through the lens of computational and representational constraints on inference.

In this chapter, we will carve off a piece of the general decision-theoretic setup introduced in Chapter 1. Specifically, we will ignore actions and rewards, focusing only on the updating of beliefs about hidden states (s) after observing data (x). This will set the stage for subsequent chapters in which we integrate belief updating with reward prediction and action selection.

In this chapter, we focus on problems where *exact* inference is possible. The next chapter will consider *approximate* inference algorithms applicable to a broader range of problems.

As described in Chapter 1, the normative standard for belief updating is Bayesian inference. The empirical question is whether the brain is (approximately) Bayesian. We will start by considering human behavior in simple probability judgment tasks, which shows systematic deviations from Bayesian inference. These deviations can be understood in terms of computational resource constraints on belief updating. We then consider what kinds of resource-constrained algorithms could give rise to these patterns of behavior, how they could be implemented using the neural building blocks introduced in Chapter 2, and what empirical evidence supports this neural implementation.

We also analyze the phenomenon of repulsion from high prior probability states in simple magnitude estimation tasks. This *appears* "anti-Bayesian" but in fact is consistent with Bayesian inference under certain representational assumptions about how the brain encodes magnitudes. We will examine neurophysiological data consistent with these assumptions, and discuss how neural networks can decode point estimates from the representations.

1 *Is behavior Bayesian?*

Answering this question is trickier than it might seem, because we need to know what (if any) prior, likelihood, posterior, and utility function the brain uses. One approach is to manufacture experimental tasks that tightly control all of these factors and impose them on human subjects. This approach has the advantage of allowing us to precisely answer the question, but it has the disadvantage of being

rather contrived, and it has been argued that people struggle with explicitly presented probability information (Gigerenzer and Hoffrage, 1995). We will first discuss one version of this approach (the urn task), and then describe other tasks which rely on implicit probability information learned through direct experience rather than verbal communication.

1.1 *The urn task*

Consider the following stylized task. I show you two urns filled with different compositions of green and yellow marbles (Figure 1). I choose one of the urns with a probability known to you (0.4 for urn A and 0.6 for urn B) and pull a marble from it. You get to see the marble but you don't get to see which urn I chose. Your task is to report the probability that I chose urn A.

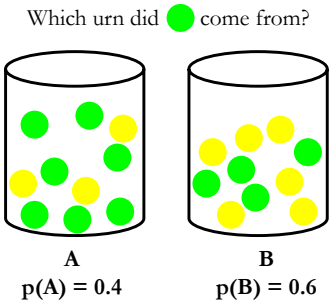

Figure 1: **The urn task.** Tasks like this have been used in many experiments with human subjects (e.g., Peterson et al., 1965; Phillips and Edwards, 1966), though differing in superficial details.

It's mathematically useful to convert the posterior probability into a "log odds" scale:

$$\log \frac{p(A|●)}{p(B|●)} = \log \frac{p(●|A)}{p(●|B)} + \log \frac{p(A)}{p(B)}, \tag{1}$$

where the first term is the likelihood log odds and the second term is the prior log odds. To quantitatively evaluate the Bayesian hypothesis, we can generalize this equation to a more flexible model with coefficients α and β (Grether, 1980):

$$y = \alpha \log \frac{p(●|A)}{p(●|B)} + \beta \log \frac{p(A)}{p(B)}, \tag{2}$$

where y is the response generated by human subjects. When the coefficients are fit to human response data, both are systematically below 1 (Figure 2), although to a first approximation reaction to the prior is close enough to 1 that we can safely ignore it. Thus, even in this idealized setting with full information and tight experimental control, people don't perfectly execute Bayes' rule. Although people update in the correct direction, they systematically under-react to the likelihood (i.e., the urn composition in this case). We will try to understand under-reaction and related deviations from Bayesian optimality in more detail.

The urn task is not representative of real-world inference problems in two ways. First, all the relevant information is given explicitly to subjects. In more realistic settings, this information often needs to be acquired from experience interacting with the environment. Second, the hypothesis space is very small compared to real-world problems. Consider, for example, the problem of inferring the three-dimensional configuration of a scene from the two-dimensional visual information conveyed by the retina. We aren't given explicit probabilities for

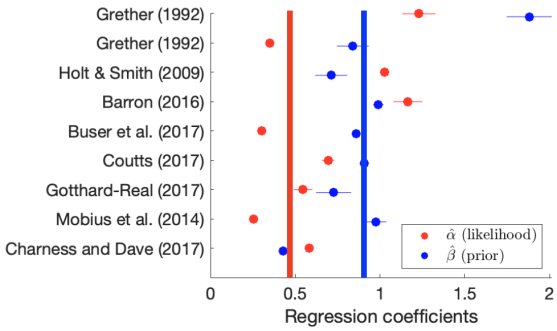

Figure 2: **Coefficient estimates from a meta-analysis of human responses in the urn task.** The vertical lines show meta-analytic estimates of the coefficients. Adapted from Zhu and Griffiths (2026).

each possible scene configuration, and the space of configurations is massive, precluding exhaustive enumeration. To understand probabilistic inference in the brain, will we need to go beyond the naive application of Bayes' rule and study scalable approximate inference algorithms (see next chapter).

1.2 *Magnitude estimation*

More realistic inference problems have been studied, with some surprising results. Before describing these results, let's talk about what seems like a generic property of Bayesian inference: the posterior mean should tend to be biased towards the prior mean. More formally, the bias is defined as $\mathbb{E}[\hat{s} - s|s]$, where $\hat{s} = \mathbb{E}[s|x]$ is the posterior mean. We expect bias to be positive whenever the state is below the prior mean, $s < \mathbb{E}[s]$, and negative whenever the state is above the prior mean, $s > \mathbb{E}[s]$. Indeed, many studies do show a bias towards the prior mean in simple magnitude estimation tasks (e.g., estimating the length of a line or the duration of a sound), sometimes called the *central tendency effect* (Hollingworth, 1910; Petzschner et al., 2015).

To make this concrete, let's consider a simple Gaussian model (Figure 3):

$$x \sim \mathcal{N}(s, \sigma_x^2), \quad s \sim \mathcal{N}(\bar{s}, \sigma_s^2). \tag{3}$$

The posterior mean is a convex combination of the signal x and the prior mean $\bar{s}$:

$$\hat{s} = wx + (1 - w)\bar{s}, \tag{4}$$

where

$$w = \frac{\sigma_s^2}{\sigma_x^2 + \sigma_s^2} \tag{5}$$

is the signal sensitivity, which is larger when the signal noise vari-

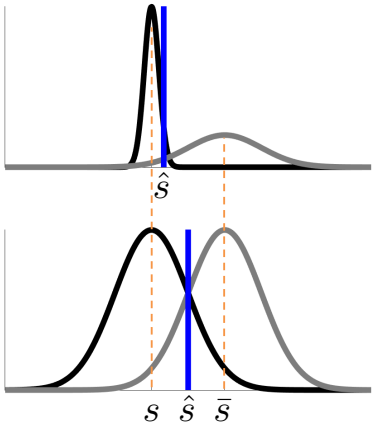

Figure 3: **Bayesian estimation for a Gaussian model.** The gray curve is the prior distribution, $p(s|\bar{s})$. The black curve is the signal distribution, $p(x|s)$. The blue line is the posterior mean. The top and bottom plots differ only in their prior variance, σ_s^2.

Sensitivity w is bounded between 0 (ignoring the signal entirely) and 1 (trusting the signal entirely).

ance is small relative to the prior variance. Using this expression, the bias is given by:

$$\mathbb{E}[\hat{s} - s|s] = \mathbb{E}[wx + (1-w)\bar{s} - s|s] = (1-w)(\bar{s} - s). \tag{6}$$

This demonstrates that the prior mean attracts the posterior mean (i.e., the bias is negative when the prior mean is less than s, and positive when the prior mean is greater than s), and the strength of this attraction is inversely proportional to the signal sensitivity.

The central tendency effect is our first glimpse of *inductive bias*: a preference for some hypotheses over others before observing data. This concept will show up in several places throughout the book.

One prediction of the model is that the central tendency effect should be stronger when the signal noise variance is relatively large and the prior variance is relatively small. Signal noise variance tends to increase with magnitude, possibly due to a nonlinear transformation from objective to subjective magnitude (more on this later); consistent with the model prediction, central tendency effects are stronger for larger magnitudes (Xiang et al., 2021). The study by Xiang et al. (2021) also found that increasing signal noise variance using shorter stimulus durations strengthened the central tendency effect. Similarly, the central tendency effect is strengthened by interposing a delay between the stimulus and judgment, or by adding noise to the stimulus (Olkkonen et al., 2014).

An empirical challenge to the Bayesian model of magnitude estimation comes from studies reporting human judgments that are *repulsed* from the prior mean—apparently an "anti-Bayesian" bias. For example, people judge a smaller object to be heavier than a larger object with the same mass (the *size-weight illusion*, first described by Charpentier, 1891). This seems to defy the prior that larger objects tend to be more massive. Repulsive biases have also been reported for orientation and spatial frequency judgments (Wei and Stocker, 2015). Taken at face value, these repulsive biases seem rather disastrous for Bayesian models of human inference.

See Peters et al. (2016) and Wolf et al. (2018) for more nuanced analyses of the size-weight illusion that take density into account.

2 *Explaining deviations from Bayesian inference*

We have seen that Bayesian inference matches behavior, at least qualitatively, in some ways but not in others. Can a resource-rational theory do better? To answer this question, we need to specify the cost function constraining the brain's inference algorithm.

2.1 Resource-rational analysis of costly inference

Intuitively, under-reaction suggests that updating from the prior to the (approximate) posterior is costly. To formalize this intuition, we first replace the true posterior $p(s|x)$ with an approximate posterior $q(s|x)$ in order to make explicit that we are no longer assuming exact

Bayesian inference. Next, we assume that the action a output deterministically by policy π is the approximate posterior: $a = q$. Finally, we quantify the cost of updating after observing signal x using the Kullback-Leibler (KL) divergence:

Later we will relax the determinism assumption.

$$\mathcal{D}[q(s|x)||p(s)] = \sum_s q(s|x) \log \frac{q(s|x)}{p(s)}. \tag{7}$$

According to this definition, belief updates that move the approximate posterior $q(s|x)$ farther from the prior $p(s)$ are more costly; when these two distributions are identical, the KL divergence achieves its minimum value of 0. We can now define the expected cost $c(\pi)$ under policy π, which averages over signals:

$$c(\pi) = \sum_x p(x)\mathcal{D}[q(s|x)||p(s)]. \tag{8}$$

When inference is exact, $q(s|x) = p(s|x)$, the expected KL cost is equal to the mutual information between the hidden state and the signal: $c(\pi) = \sum_x p(x) \sum_s p(s|x) \log \frac{p(s|x)}{p(s)}$. Thus, an ideal agent will pay a cost equal on average to the amount of information about the hidden state conveyed by the signal.

Turning now to the utility part of the optimization problem, if we're only concerned with the correctness beliefs (rather than external reward), then the utility should be higher when our beliefs are closer to the posterior. We can formalize this by stipulating that rewards are signals ($r = x$) and that the utility derived from these signals is the negative KL divergence between the approximate and true posterior:

$$u(r) = -\mathcal{D}[q(s|x)||p(s|x)]. \tag{9}$$

In Chapter 1, we defined resource rationality in terms of a bounded optimization problem:

$$\pi^* = \underset{\pi: c(\pi) \leq \mathcal{C}}{\operatorname{argmax}}\ \bar{u}(\pi), \tag{10}$$

where $\mathcal{C}$ is a capacity limit and

$$\bar{u}(\pi) = \mathbb{E}[u(r)|\pi] = -\sum_x p(x)\mathcal{D}[q(s|x)||p(s|x)] \tag{11}$$

is the expected utility. We can equivalently formulate this as an unbounded optimization using the method of Lagrange multipliers:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\ \bar{u}(\pi) - \lambda c(\pi), \tag{12}$$

where the Lagrange multiplier $\lambda \geq 0$ is given by:

$$\lambda = \frac{\partial \bar{u}(\pi^*)}{\partial c(\pi^*)}, \tag{13}$$

with $c(\pi^*) = \mathcal{C}$ (i.e., the optimal policy operates at the capacity limit). The Lagrange multiplier enforces the constraint that the optimal expected utility $\bar{u}(\pi^*)$ decreases with resource consumption. Eq. 12 is useful because it shows how resource units can be converted into

In general, λ decreases with capacity C; more capacity implies lower cost.

commensurable utility units, with λ acting as the conversion factor. It also motivates the linear cost function that is used in many models.

Plugging these terms into Eq. 12 and solving for π^* yields an approximate posterior q^* that maximizes utility relative to the information processing cost (Zhu and Griffiths, 2026):

$$q^*(s|x) \propto p(x|s)^{1/(1+\lambda)} p(s). \tag{14}$$

Notice that this is just Bayes' rule, except that the likelihood is down-weighted. This implies under-reaction to the likelihood, as seen experimentally. In arriving at this result, we have not made any specific algorithmic assumptions beyond requiring that the approximate inference algorithm (whatever it might be) must obey an information-theoretic capacity limit. Despite such weak assumptions, the result shows that we can make a broad and rigorous assertion about the behavioral consequences of resource rationality.

2.2 *Neural implementation of costly inference*

How might something like Eq. 14 be implemented in the brain? In this section, we'll start by describing some simple neural models of optimal Bayesian inference, and then consider how they can be modified to incorporate inference cost.

Let's start with a simple setup in which a single neuron approximates the posterior over two possible states $s \in \{A, B\}$. We will make use of the integrate-and-fire model from Chapter 1 (in this case without leak), where the dynamics of the neuron's membrane potential $\mu(t)$ at time t are described by the following linear differential equation:

$$C\dot{\mu} = I(t), \qquad \mu(0) = \mu^0, \tag{15}$$

where C is the membrane capacitance, μ^0 is the resting potential, and $I(t)$ is the input current, which we model as a linear combination of synaptic inputs $(z_1, \ldots, z_D)$ weighted by synaptic strengths $(w_1, \ldots, w_D)$:

$$I(t) = \sum_d w_d z_d(t). \tag{16}$$

By removing leak, we are now dealing with a *perfect integrator* of its synaptic inputs: the membrane potential reports the accumulated input current without any loss of information. More realistic models are "leaky" (decaying back to the resting potential), but for present purposes we will ignore leak (see Brunton et al., 2013, for behavioral evidence supporting perfect integration).

The inputs are generated by the Poisson spiking activity of presynaptic neurons; $z_d(t) = 1$ if presynaptic neuron d spiked at time t (0 otherwise).

Suppose we had only a single presynaptic neuron ($D = 1$), firing with rate $f(A)$ when $s = A$ and with rate $f(B)$ when $s = B$. After observing x spikes between time 0 and t, the log-likelihood ratio

under Poisson spiking is given by:

$$\log \frac{p(x|s=A)}{p(x|s=B)} = x_1 \log \frac{f(A)}{f(B)} + f(B) - f(A). \tag{17}$$

This expression is a linear function of the spike count. If we set the synaptic strength to be $w_1 = \log \frac{f(A)}{f(B)}$, the postsynaptic neuron will accumulate weighted spike counts over time such that its membrane potential represents the posterior log-odds:

$$\mu(t) = \log \frac{p(s=A|x_1)}{p(s=B|x_1)}, \tag{18}$$

provided the resting potential is given by:

$$\mu^0 = \log \frac{p(s=A)}{p(s=B)} + f(B) - f(A). \tag{19}$$

One problem with this model is that it requires the postsynaptic neuron to have precise knowledge about the firing rates of the presynaptic neuron. If $f(A)$ and $f(B)$ are imprecisely estimated, this will produce a bias in the posterior log-odds. One way to deal with this problem, proposed by Gold and Shadlen (2001), is to posit a second 'antineuron' with opposite tuning, firing with rate $f_2(A) = f_1(B)$ and with rate $f_2(B) = f_1(A)$), where we now designate the tuning function of the first neuron by f_1. The synaptic strength for the antineuron is the same as the strength for its paired neuron, but with opposite sign: $w_2 = -w_1 = \log \frac{f_2(A)}{f_2(B)}$. The log-likelihood ratio then becomes:

$$\log \frac{p(x_1, x_2|s=A)}{p(x_1, x_2|s=B)} = w_1 x_1 + w_2 x_2. \tag{20}$$

The resting potential for the postsynaptic neuron is simplified to:

$$\mu^0 = \log \frac{p(s=A)}{p(s=B)}. \tag{21}$$

The virtue of this scheme is that it doesn't require precise knowledge of the tuning functions, as long as the estimates have the correct sign (so that errors may affect the rate of evidence accumulation but won't systematically bias the posterior log-odds).

While an elegant idea, there isn't much direct evidence for neuron-antineuron pairs in the brain. Furthermore, the scheme doesn't take advantage of presynaptic populations with diverse tuning. Let's suppose instead, following Jazayeri and Movshon (2006), that there is a large population of presynaptic neurons, where neuron d has tuning function $f_d(s)$. The log-likelihood under Poisson spiking is given by:

$$\log p(x|s) = \sum_d x_d \log f_d(s) - f_d(s) - \log x_d! \tag{22}$$

The third term doesn't depend on s, so we can ignore it. We will also ignore the second term under the assumption that $\sum_d f_d(s)$ is a constant. This assumption is valid when the population uniformly covers the state space, so that the total response to a state is always the same (as we'll see, this uniformity assumption does not hold true in general, for interesting computational reasons). After discarding these terms, the log-likelihood ratio becomes:

$$\log \frac{p(x|s=A)}{p(x|s=B)} = \sum_d x_d \log \frac{f_d(A)}{f_d(B)}. \tag{23}$$

This is quite similar to the approach of Gold and Shadlen (2001), except that it doesn't require neuron/antineuron pairs. Again we have synaptic strengths equal to the log-transformed ratio between firing rates, so that the postsynaptic membrane potential tracks the posterior log odds.

We will systematically examine the decision-making aspects of sensory discrimination in Chapter 7.

To make things concrete, let's look at a canonical sensory discrimination task, where subjects (typically monkeys or humans) are asked to decide whether a cloud of moving dots is moving in one of two directions (Figure 4). A proportion of the dots are moving in the same direction, while the rest are moving in random directions. This proportion of coherently moving dots (*coherence* for short) allows the experimenter to manipulate the strength of evidence available for the decision.

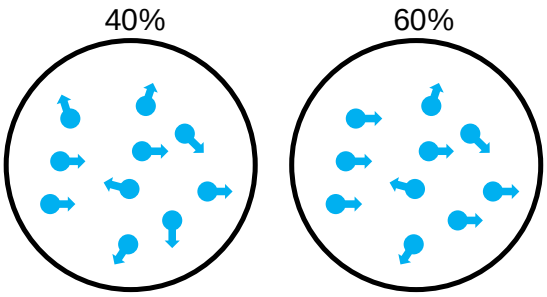

Figure 4: **Random dot motion stimuli.** Also known as a *random dot kinematograms*. The proportion of coherently moving dots is shown above each stimulus.

Neurophysiological studies have located the presynaptic population in extrastriate area MT (also known as V5), where most neurons are tuned to particular motion directions (Dubner and Zeki, 1971) and increase their firing in proportion to coherence in their preferred direction (Britten et al., 1992). The tuning functions of these neurons are modeled fairly well by a cosine function defined over the space of motion directions ($s \in [0, 360]$):

$$f_d(s) = \exp[\cos(s - s_d^*)/\nu], \tag{24}$$

where s_d^* is the preferred direction for neuron d and ν is the tuning width. One synapse downstream, neurons in parietal area LIP in-

tegrate the spiking of MT neurons, enabling them to compute the posterior log odds.

The motion discrimination setup can be used to test some empirical predictions of the Jazayeri and Movshon (2006) model. When subjects need to discriminate opposite directions of motion (e.g., left vs. right), the largest log-likelihood ratios will come from neurons tuned to the two target directions. In contrast, when subjects need to discriminate nearby directions, the largest log-likelihood ratios will come from neurons tuned to off-target directions. Intuitively, this is because neurons that respond similarly to the two alternatives will have small log-likelihood ratios—they provide little information about the state. These predictions can be tested by predicting choice behavior from the activity of MT neurons with different tuning. Consistent with the model, neurons tuned to target directions contribute more to choice behavior when discriminating opposite directions (Britten et al., 1996), whereas neurons tuned to off-target directions contribute more to choice behavior when discriminating nearby directions (Purushothaman and Bradley, 2005).

LIP neurons ramp up over time during viewing of the random dot motion stimulus (Figure 5). The slope of this ramp increases with coherence, consistent with the observation that MT neuron firing rates increase with coherence, thereby driving downstream LIP neurons more strongly.

One drawback of the random dot motion stimulus is that it is difficult to precisely quantify the information value of the stimulus at any given time. A study by Yang and Shadlen (2007) addressed this issue, recording LIP neurons while monkeys viewed a sequence of abstract shapes. At the end of the sequence, the monkey needed to choose (via an eye movement) one of two visual targets. The correct target (yielding a water reward) was determined by the shape sequence: each shape was associated with a particular log-likelihood ratio for one target vs. the other, such that the total log-likelihood ratio could be obtained by summing up the contributions of the shapes in the sequence. Changes in the firing rate of LIP neurons were linearly related to the log-likelihood ratio (Figure 6).

To summarize, neurophysiological data support the evidence accumulation mechanism formalized in this section: input neurons report the momentary evidence signals, which are linearly transformed by a set of synaptic strengths (log-likelihood ratios) into a postsynaptic membrane potential reporting the posterior log-odds. This simple model (as we'll see) is certainly not the whole story, but it provides an empirically defensible starting point for thinking about how inference costs could enter the picture.

Recall from the last section that the cost coefficient λ enters through

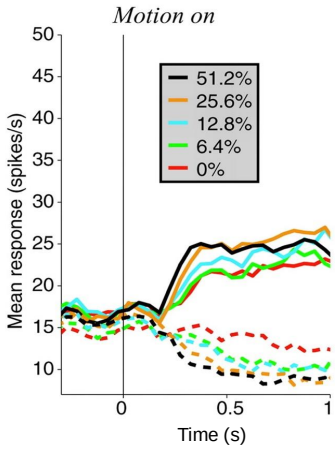

Figure 5: **LIP neurons accumulate evidence.** Firing rate of LIP neurons following stimulus onset in a motion discrimination task. Solid and dashed curves are from trials in which the monkey judged direction toward and away from a cell's preferred direction, respectively. Colors show different levels of motion coherence. Adapted from Shadlen and Newsome (2001).

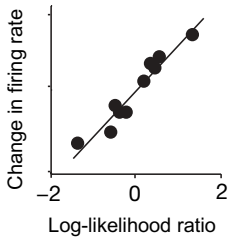

Figure 6: **Changes in the firing rate of LIP neurons tracks the log-likelihood ratio.** Each dot shows the average change in the firing rate of LIP neurons following the appearance of a shape that signals evidence about the correct visual target. The x-axis is the log-likelihood ratio associated with the shape. Adapted from Yang and Shadlen (2007).

an exponent of the likelihood. This implies that λ enters through a multiplier of the log-likelihood ratio. Using the Jazayeri and Movshon (2006) model, this can be interpreted as a global modulation:

$$\log \frac{p(x|s=A)^{1/(1+\lambda)}}{p(x|s=B)^{1/(1+\lambda)}} = \frac{1}{1+\lambda} \sum_d x_d \log \frac{f_d(A)}{f_d(B)}. \tag{25}$$

As λ increases (lower capacity C), the log-likelihood is suppressed. This could be interpreted mechanistically in several (not mutually exclusive) ways: suppression of firing, suppression of synaptic strengths, or suppression of the postsynaptic membrane potential. All of these are energetically costly (Niven, 2016), consistent with the idea that their suppression conserves energetic resources.

A study by Padamsey et al. (2022) grounds these theoretical speculations in empirical data (Figure 7). Mice were placed in a pool of water (which they don't like), where they searched for a submerged platform located in front of a particular visual cue (a grating with a particular orientation). They had to discriminate this visual cue from another visual cue at a different location with a different frequency. Thus, this is essentially a two-alternative sensory discrimination task of the form that we've already analyzed.

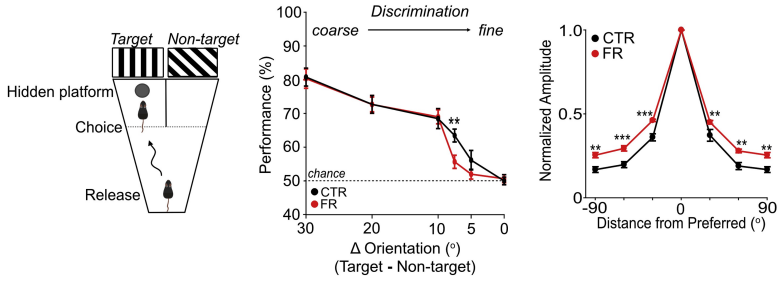

Figure 7: **Orientation discrimination under food restriction.** (Left) Discrimination task. (Middle) Discrimination performance as a function of the difference between target and non-target orientations. (Right) Orientation tuning in V1. CTR: control mice; FR: food-restricted mice. Adapted from Padamsey et al. (2022).

Food-restricted mice were impaired at discriminating similar frequencies, consistent with the hypothesis that capacity is reduced (and hence λ is increased) under food restriction. Recordings of orientation-tuned neurons in primary visual cortex (V1) revealed broadened tuning under food restriction. If we adapt the cosine function used by Jazayeri and Movshon (2006) to model direction tuning in MT (Eq. 24), we would say that the tuning width ν increases under food restriction.

The recordings from V1 during viewing of natural scene images also revealed that there was reduced discriminability of responses to similar images of natural scenes, as would be predicted by a general increase in tuning width.

The mechanism underlying this change was a reduction in AMPA receptor conductance, which was compensated for by increased input resistance and depolarization of the membrane potential. This had the effect of maintaining roughly the same firing rates but making firing more variable. The broader orientation tuning essentially reflects this higher variability (i.e., a higher probability of randomly responding to stimuli farther away from a neuron's preferred stimulus).

These findings can be captured qualitatively by the costly inference model (Figure 8).

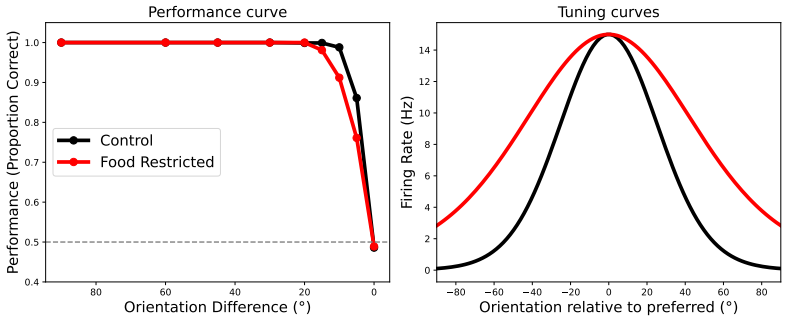

Figure 8: **Simulations of the costly inference model** (Left) Discrimination performance as a function of the difference between target and non-target orientations. (Right) Orientation tuning curves.

Summarizing the key insight from this section: the resource-rational solution may be realized neurally through a reduction in tuning precision.

2.3 *Representational effects on inference*

The previous section showed how optimally balancing inference cost and utility led to down-weighting the likelihood. This can explain under-reaction to data (as observed in the urn task), but it leaves unexplained how judgments in magnitude estimation tasks can sometimes be repulsed away from the prior mean. As pointed out above, this appears manifestly non-Bayesian, even if we allow for costly inference. We will see, however, that the internal representation of magnitudes can qualitatively change the predictions of the Bayesian analysis. This will lead us to the surprising result that Bayesian inference can produce both attractive and repulsive effects under certain representational assumptions.

We start by writing down a utility function that is appropriate for magnitude estimation. Let $a = \hat{s}$ denote a point estimate of a one-dimensional, continuous hidden state s. The policy π is assumed to be a deterministic mapping from the sensory signal x to the point estimate. We parametrize the utility as a one-dimensional function of the error $\epsilon = s - \hat{s}$: $u(\epsilon) = -|\epsilon|^\kappa$, where κ is an integer. The optimal policy, $\pi^* = \text{argmax}_\pi \bar{u}(\pi)$, outputs different point estimates under different parameter choices:

We no longer assume that $\hat{s}$ is the posterior mean; rather, the optimal point estimate depends on the utility function.

- $\kappa = 0$: posterior mode.
- $\kappa = 1$: posterior median.
- $\kappa = 2$: posterior mean.

The posterior mode is also known as the *maximum a posteriori* (MAP) estimate.

The next, crucial step is to define the measurement model. We assume that the state is mapped into an encoding $f(s)$ and then cor-

rupted by Gaussian noise: We will not make strong assumptions about the prior $p(s)$.

Technically, the mathematical analysis requires that $\log p(s)$ be twice differentiable.

$$x = f(s) + \epsilon, \tag{26}$$

where $\epsilon \sim \mathcal{N}(0, \sigma_s^2)$.

A useful result, derived by Hahn and Wei (2024), is an approximation of the estimation bias:

See also Prat-Carrabin and Woodford (2021).

$$\mathbb{E}[\hat{s} - s|s] \approx b(s) = \frac{1}{J(s)}(\log p(s))' + \frac{\kappa + 2}{4}\left(\frac{1}{J(s)}\right)', \tag{27}$$

where $J(s)$ is the *Fisher information* (see Chapter 3). Under the measurement model of Eq. 26, this is given by:

For compactness, we use $p'(s) = \frac{\partial}{\partial s}p(s)$, and similarly for other functions.

$$J(s) = \frac{f'(s)^2}{\sigma_s^2}. \tag{28}$$

Intuitively, the Fisher information can interpreted as the quantity of representational resources devoted to encoding information about state s. When $f(s)$ varies quickly around s (large Fisher information), it means that the encoding is highly sensitive to (carries more information about) state s.

To predict attraction vs. repulsion effects, we first need to define these effects more precisely. In the context of magnitude estimation, we define bias as attractive when it's pointing in the direction of the local prior mode (i.e., when it has the same sign as the derivative of the prior) and as repulsive when it's pointing in the opposite direction. Mathematically, this means the bias is attractive when $b(s)p'(s) > 0$ and repulsive when $b(s)p'(s) < 0$. This is not the only reasonable way to define attraction/repulsion. For example, one could define it in reference to the prior mean or median. By adopting the "local mode" definition, we are implicitly endorsing the posterior mode as the relevant point estimate, with $\kappa = 0$ as the relevant utility parametrization. We will continue with this assumption, while noting that alternative assumptions will be considered later.

In other words: attraction when the signs of $b(s)$ and $p'(s)$ agree, repulsion when they disagree.

Using Eq. 27, we get:

$$b(s)p'(s) = \frac{1}{J(s)}\left[\frac{p'(s)^2}{p(s)} - \frac{J'(s)p'(s)}{J(s)}\right]. \tag{29}$$

Since $J(s)$ and $p(s)$ are both non-negative, repulsion will only occur when $J'(s)$ and $p'(s)$ have the same sign and their product is large enough to outweigh the first term in the brackets.

How should we interpret $J'(s)$? To answer this question, we need to understand the encoding function better. In many magnitude estimation tasks, estimation accuracy diminishes with magnitude. Fechner (1860) proposed that this arises from a logarithmic encoding,

$f(s) = \omega + \psi \log s$, where ω and ψ are constants. The logarithmic encoding produces diminishing sensitivity by virtue of its concavity, which implies $f''(s) < 0$ (i.e., the second derivative is negative). In other words, increasing s produces smaller and smaller changes in $f(s)$ as s gets bigger. More generally, diminishing sensitivity can be obtained by any concave, monotonically increasing encoding function. For example, Stevens (1961) proposed a power-law encoding function, $f(s) \propto s^{\alpha}$, which produces diminishing sensitivity when $\alpha < 1$.

Neural evidence for an encoding function with diminishing sensitivity will be discussed in the next section.

For a logarithmic encoding function, the Fisher information is given by:

$$J(s) = \frac{\psi^2}{s^2 \sigma_s^2}, \tag{30}$$

which has power-law (s^{-2}) scaling in magnitude (a point we will return to in the next section). We can now relate $J'(s)$ to the concavity of the encoding function, by differentiating Eq. 30:

$$J'(s) = \frac{2f''(s)f'(s)}{\sigma^2}. \tag{31}$$

Thus, $J'(s)$ linearly increases with the second derivative of the encoding function, inheriting its diminishing sensitivity: $J'(s) < 0$.

Returning to Eq. 29, the assumption that $J'(s) < 0$ means that repulsion will occur when $p'(s) < 0$. This suggests that we should tend to see repulsion effects when magnitude probability is a decreasing function of magnitude. It turns out that many natural magnitudes have this property. A good example is spatial frequency: the distribution of spatial frequencies in natural images falls off according to a power law (Dong and Atick, 1995): $p(s) \propto s^{-\alpha}$ with α between 1 and 2 (note that this is conceptually distinct from the power-law encoding function discussed above). In this case, the approximate bias is given by:

$$\begin{aligned} b(s) &\propto s^2 \left[-\frac{\alpha}{s} + \frac{\kappa + 2}{s} \right] \\ &= s \left[-\alpha + \kappa + 2 \right]. \end{aligned} \tag{32}$$

For $\alpha < 2 + \kappa$, the bias is always positive (i.e., repulsive, since the mode is at 0).

In accordance with the theory, repulsion effects have been observed in spatial frequency estimation (Figure 9). Here the bias is positive, pointing in the direction opposite the local mode. Two other features of the data are notable. First, when stimuli are low contrast, the bias increases with spatial frequency. This is consistent with the

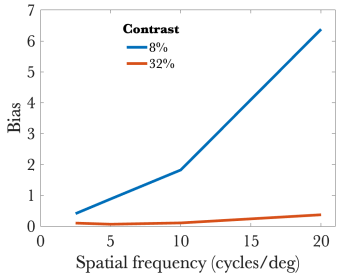

Figure 9: **Bias in spatial frequency estimation.** Adapted from Georgeson and Ruddock (1980).

fact that $J(s)$ decreases with magnitude under a logarithmic encoding, amplifying all bias effects (both attraction and repulsion). Second, when stimuli are high contrast, bias flattens out as a function of magnitude and approaches 0. This is consistent with the fact that $J(s)$ decreases with sensory noise, thereby suppressing all bias effects.

2.4 *Neural encoding of magnitude*

Treating the output of the encoding function as a scalar is a useful abstraction for deriving behavioral predictions, but it's unrealistic as a hypothesis about the brain. Behavioral judgments of subjective magnitude are computed on the basis of neural population activity, not single neurons. Thus, $f(s)$ should properly be thought of as vector-valued. As before, we will use $f_d(s)$ denote the tuning function for neuron d, with spike count $x_d \sim \text{Poisson}(f_d(s))$. Our first goal is to understand how diminishing sensitivity can be realized at the neural level. We will do this by analyzing the derivative of the Fisher information, since the analysis in the previous section showed how this is the critical factor in generating behavioral predictions.

For independent Poisson neurons, the mean and variance are identical. Thus the Fisher information is given by:

$$J(s) = \sum_d \frac{f_d'(s)^2}{f_d(s)}. \tag{33}$$

Things get more complicated when the noise is correlated, a more plausible assumption in networks of interconnected neurons (Abbott and Dayan, 1999). We will ignore this for present purposes.

Note that the Fisher information (under the uncorrelated noise assumption) decomposes into the sum of neuron-specific terms.

Next, we need to make some assumptions about tuning functions. A common finding is that magnitudes are represented in the brain by neurons with radial tuning functions. A simple parametrization of such functions is the squared exponential (or Gaussian):

$$f_d(s) = \exp\left[-\frac{(s - s_d^*)^2}{2v_d^2}\right], \tag{34}$$

where s_d^* is the preferred stimulus of neuron d (the peak of its tuning function) and v_d^2 is its tuning width. One way to get diminishing sensitivity ($J'(s) < 0$) is to concentrate the tuning functions around lower magnitudes, so that $s > s_d^*$ for most magnitudes. Another way is to increase tuning width with magnitude, which can be achieved by defining the tuning function on a logarithmic scale:

$$f_d(s) = \exp\left[-\frac{(\log s - \log s_d^*)^2}{2v_d}\right]. \tag{35}$$

Is this a good description of magnitude encoding by real neurons? Let's return to the case of spatial frequency, which has been extensively studied. Figure 10 shows tuning functions of neurons recorded

from the primary visual cortex of cats. To a first approximation, the average firing rates are well-described by Gaussians on the log scale. This kind of tuning is not unique to spatial frequency; for example, Gaussian tuning on the log scale has also been found for numerosity-tuned neurons in monkey prefrontal and parietal cortex (Nieder and Miller, 2003), and for temporally-tuned neurons in rodent hippocampus (Cao et al., 2022). Importantly, magnitude estimation in all of these domains exhibits diminishing sensitivity.

The slope of the log-Gaussian tuning function is given by:

$$f'(s) = -\frac{f(s)}{s\nu_d} \log \frac{s}{s_d^*}, \tag{36}$$

which shows that the tuning width is scaled linearly by s. The Fisher information becomes:

$$J(s) = \frac{1}{s^2} \sum_d f_d(s) \left(\frac{1}{\nu_d^2} \log \frac{s}{s_d^*} \right)^2. \tag{37}$$

We can gain further insight into this expression by taking the limit of infinite population size, assuming a uniform density of preferred stimuli in log space and a constant tuning width, ν^2. The sum can then be approximated by an integral:

$$J(s) = \frac{1}{s^2\nu^4} \int_{-\infty}^{\infty} \exp\left[-\frac{(\log s - s^*)^2}{2\nu^2} \right] (\log s - s^*)^2 ds^*. \tag{38}$$

Using a standard result for Gaussian integrals of quadratic functions, the integral yields the following scaling law for the Fisher information:

$$J(s) \propto \frac{1}{s^2}. \tag{39}$$

We will refer to neural populations obeying this law as *power-law codes*. The key takeaway from this analysis is that power-law neural codes show the same s^{-2} scaling as the logarithmic (Fechnerian) encoding function described above. In other words, we have derived an equivalence relation between a psychophysical representation (the logarithmic encoding function) and a neural representation (log-Gaussian tuning). The Fisher information allows us to bridge these levels of analysis in a unified way.

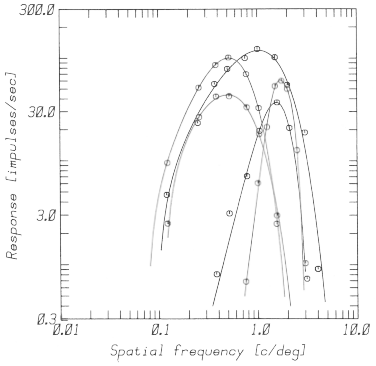

Figure 10: **Spatial frequency tuning in cat primary visual cortex.** Unpublished data courtesy of Tony Movshon.

2.5 *Bayesian decoding*

Now that we understand some of the ways that neural populations encode magnitudes, we can ask how a downstream "readout" population could decode these magnitudes. Let's return to Eq. 22, which says that the log-likelihood is a linear function of the presynaptic

See Pouget et al. (2013) for a survey of Bayesian encoding and decoding in neural populations.

spike counts $x = (x_1, \ldots, x_D)$, implicitly accumulated over some time window up to time t. If we have a set of neurons selective for different magnitudes $\{\tilde{s}_j\}$, they can compute the log-likelihoods using the perfect integrator model in Eq. 15, with input current $I_j(t)$, weights w_{jd}, and resting potential μ_j^0 given by:

Here j indexes a discrete set of magnitudes.

$$I_j(t) = \sum_d w_{jd} z_d(t), \tag{40}$$

$$w_{jd} = \log f_d(\tilde{s}_j) \tag{41}$$

$$\mu_j^0 = -\sum_d f_d(\tilde{s}_j). \tag{42}$$

As before, $z_d(t) = 1$ if presynaptic neuron d spiked at time t (o otherwise). Under these dynamics, the membrane potential $\mu_s(t)$ reports the log likelihood of s up to a constant. The maximum likelihood estimate then corresponds to the most active neuron's designated magnitude.

What about the prior? One approach is to encode it in the resting potential:

$$\mu_j^0 = \log p(\tilde{s}_j) - \sum_d f_d(\tilde{s}_j). \tag{43}$$

Under this assumption, the posterior mode (MAP estimate) corresponds to the most active neuron's designated magnitude. We can also compute the full posterior distribution using softmax normalization:

$$p(s = \tilde{s}_j | x) = \frac{\exp[\mu_j(t)]}{\sum_{j'} \exp[\mu_{j'}(t)]}, \tag{44}$$

which can be implemented by assuming that the firing rate of each postsynaptic neuron is an exponential function of its membrane potential, divisively normalized by the activation of the entire postsynaptic population.

This model predicts that the firing rate of readout neurons should increase monotonically with the posterior probability of their designated magnitude. A study of monkey superior colliculus neurons provides an illustrative example. The superior colliculus (in particular its intermediate layers) is a subcortical structure that is involved in the programming of eye movements based on input from cortex, and thus is a plausible readout area. Cells in this area represent a map of possible target locations for eye movement. "Buildup" neurons in this area increase their activity gradually prior to movement onset (Munoz and Wurtz, 1995). Basso and Wurtz (1997) manipulated visual target uncertainty by presenting different numbers of possible visual targets. After a delay, one of these targets was selected and

the monkey was trained to move its eye to that target. When more targets were present, the activity of buildup neurons was reduced; firing abruptly increased when the target was revealed. In a second study, the number of previewed targets was held fixed, and uncertainty was manipulated by comparing a condition where the same target was always selected with a condition where different targets were randomly selected on each trial. Again, higher uncertainty reduced the firing rate of target-selective neurons. Later work showed that this reduction of firing is correlated with reductions in behavioral performance on a trial-by-trial basis (Kim and Basso, 2008). In summary, these studies are consistent with the hypothesis that the superior colliculus functions as a probabilistic readout area sensitive to uncertainty in cortical populations.

The superior colliculus is not the only candidate readout area in the brain. For example, activity in the orbitofrontal cortex correlates with decision confidence (Masset et al., 2020), and activity in several prefrontal areas correlates with confidence in perceptual judgments (Geurts et al., 2022).

2.6 *Weber's law*

So far we have been discussing magnitude estimation tasks, but another important class of tasks is magnitude discrimination. A classic result from this class of tasks is *Weber's law* (Weber, 1834): the discrimination threshold (the magnitude difference needed to produce a criterion level of average discrimination accuracy) is linear in the reference magnitude (typically the lower or average magnitude of the two stimuli). An example of this phenomenon is shown in Figure 11. In this section, we show how the power-law neural code gives rise to Weber's law.

The discrimination threshold is also known as the *just noticeable difference*.

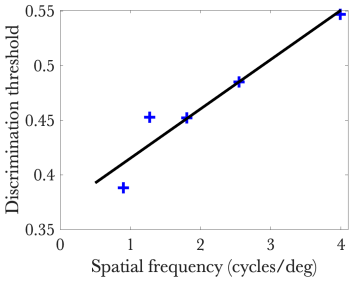

Figure 11: **Weber's law in spatial frequency discrimination.** Discrimination threshold (stimulus difference yielding 85% accuracy) as a function of the lower spatial frequency. Black line shows the least-squares regression fit. Adapted from Campbell et al. (1970).

An observer is asked to judge whether $s_1 > s_2$. Given the sensory evidence, the observer forms point estimates $\hat{s}_1$ and $\hat{s}_2$. If the evidence is sufficiently strong, the point estimates will be approximately Gaussian-distributed (due to sensory noise, for example from stochastic spiking):

$$\hat{s}_i \sim \mathcal{N}(s_i, 1/J(s_i)). \tag{45}$$

We assume that the observer selects stimulus 1 whenever $\hat{s}_1 > \hat{s}_2$. By marginalizing over the estimation noise, we obtain the accuracy rate (assuming $s_1 > s_2$):

$$p(\hat{s}_1 > \hat{s}_2) = \Phi\left(\frac{\Delta}{\sqrt{1/J(s_1) + 1/J(s_2)}}\right), \tag{46}$$

where $\Phi(\cdot)$ is the standard Gaussian CDF and $\Delta = s_1 - s_2$. Let D denote the criterion accuracy rate. We can then solve for Δ (the discrimination threshold):

$$\Delta = \Phi^{-1}(D)\sqrt{1/J(s_1) + 1/J(s_2)}. \tag{47}$$

Assuming that Δ is close to 0 and therefore $1/J(s_1) + 1/J(s_2) \approx 1/J(s)$ with $s = (s_1 + s_2)/2$, the discrimination threshold is given by:

$$\Delta \approx \Phi^{-1}(D)\sqrt{1/J(s)}. \tag{48}$$

Finally, we apply the power-law code (Eq. 39), yielding:

$$\Delta \propto s. \tag{49}$$

We thus obtain Weber's law: the discrimination threshold increases linearly with magnitude. Once again, the Fisher information provides a bridge between neural representation and psychophysical performance.

Seriès et al. (2009) derive a closely related result using the Cramér-Rao bound, which states that the variance of any unbiased estimator is lower-bounded by the inverse Fisher information.

3 *Conclusion*

In this chapter, we started with the normative ideal of Bayesian inference, and then tried to explain both the successes and failures of this ideal as a model of inference in the brain. The key idea is that computational and representational constraints shape inference in ways that comport with empirical observations. We also saw how these constraints can be realized in simple neural networks.

Despite the elegance of these solutions, their scope is quite limited. Realistic states are high-dimensional, with complex interdependencies. To deal with these more realistic scenarios, we need to think about how to implement *approximate* algorithms in neural networks. This is the subject of the next chapter.

Study questions

1. What other cost functions besides Kullback-Leibler divergence might be plausible? What are the advantages and disadvantages of different cost functions?
2. How would you modify the random dot motion discrimination task to directly test predictions of the resource-rational inference model?
3. In what ways might resource-rational inference vary systematically across individuals (e.g., children, older adults, clinical populations)? How would you test this empirically?

References

Abbott, L. F. and Dayan, P. (1999). The effect of correlated variability on the accuracy of a population code. *Neural Computation*, 11:91–101.

Basso, M. A. and Wurtz, R. H. (1997). Modulation of neuronal activity by target uncertainty. *Nature*, 389:66–69.

Britten, K., Newsome, W., Shadlen, M., Celebrini, S., and Movshon, J. (1996). A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Visual Neuroscience*, 13:87–100.

Britten, K. H., Shadlen, M. N., Newsome, W. T., and Movshon, J. A. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. *Journal of Neuroscience*, 12:4745–4765.

Brunton, B. W., Botvinick, M. M., and Brody, C. D. (2013). Rats and humans can optimally accumulate evidence for decision-making. *Science*, 340:95–98.

Campbell, F. W., Nachmias, J., and Jukes, J. (1970). Spatial-frequency discrimination in human vision. *JOSA*, 60:555–559.

Cao, R., Bladon, J. H., Charczynski, S. J., Hasselmo, M. E., and Howard, M. W. (2022). Internally generated time in the rodent hippocampus is logarithmically compressed. *Elife*, 11:e75353.

Charpentier, A. (1891). Analyse experimentale: De quelques elements de la sensation de poids. *Archives de Phisiologie Normale et Pathologique*, 3:122–135.

Dong, D. W. and Atick, J. J. (1995). Statistics of natural time-varying images. *Network: Computation in Neural Systems*, 6:345.

Dubner, R. and Zeki, S. (1971). Response properties and receptive fields of cells in an anatomically defined region of the superior temporal sulcus in the monkey. *Brain Research*, 35:528–532.

Fechner, G. T. (1860). *Elemente der Psychophysik*. Breitkopf u. Härtel.

Georgeson, M. and Ruddock, K. (1980). Spatial frequency analysis in early visual processing. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 290:11–22.

Geurts, L. S., Cooke, J. R., van Bergen, R. S., and Jehee, J. F. (2022). Subjective confidence reflects representation of bayesian probability in cortex. *Nature Human Behaviour*, 6:294–305.

Gigerenzer, G. and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102:684–704.

Gold, J. I. and Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences*, 5:10–16.

Grether, D. M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly Journal of Economics*, 95:537–557.

Hahn, M. and Wei, X.-X. (2024). A unifying theory explains seemingly contradictory biases in perceptual estimation. *Nature Neuroscience*, 27:793–804.

Hollingworth, H. (1910). The central tendency of judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, 7:461–469.

Jazayeri, M. and Movshon, J. A. (2006). Optimal representation of sensory information by neural populations. *Nature Neuroscience*, 9:690–696.

Kim, B. and Basso, M. A. (2008). Saccade target selection in the superior colliculus: a signal detection theory approach. *Journal of Neuroscience*, 28:2991–3007.

Masset, P., Ott, T., Lak, A., Hirokawa, J., and Kepecs, A. (2020). Behavior-and modality-general representation of confidence in orbitofrontal cortex. *Cell*, 182:112–126.

Munoz, D. and Wurtz, R. (1995). Saccade-related activity in monkey superior colliculus. I. Characteristics of burst and buildup cells. *Journal of Neurophysiology*, 73:2313–2333.

Nieder, A. and Miller, E. K. (2003). Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. *Neuron*, 37:149–157.

Niven, J. E. (2016). Neuronal energy consumption: biophysics, efficiency and evolution. *Current Opinion in Neurobiology*, 41:129–135.

Olkkonen, M., McCarthy, P. F., and Allred, S. R. (2014). The central tendency bias in color perception: Effects of internal and external noise. *Journal of Vision*, 14:5–5.

Padamsey, Z., Katsanevaki, D., Dupuy, N., and Rochefort, N. L. (2022). Neocortex saves energy by reducing coding precision during food scarcity. *Neuron*, 110:280–296.

Peters, M. A., Ma, W. J., and Shams, L. (2016). The size-weight illusion is not anti-Bayesian after all: a unifying Bayesian account. *PeerJ*, 4:e2124.

Peterson, C. R., Schneider, R. J., and Miller, A. J. (1965). Sample size and the revision of subjective probabilities. *Journal of Experimental Psychology*, 69:522–527.

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

Phillips, L. and Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, 72:346–354.

Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, 16:1170–1178.

Prat-Carrabin, A. and Woodford, M. (2021). Bias and variance of the bayesian-mean decoder. *Advances in Neural Information Processing Systems*, 34:23793–23805.

Purushothaman, G. and Bradley, D. C. (2005). Neural population code for fine perceptual decisions in area MT. *Nature Neuroscience*, 8:99–106.

Seriès, P., Stocker, A. A., and Simoncelli, E. P. (2009). Is the homunculus "aware" of sensory adaptation? *Neural Computation*, 21:3271–3304.

Shadlen, M. N. and Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, 86:1916–1936.

Stevens, S. S. (1961). To honor Fechner and repeal his law: A power function, not a log function, describes the operating characteristic of a sensory system. *Science*, 133:80–86.

Weber, E. H. (1834). *De Pulsu, resorptione, auditu et tactu: Annotationes anatomicae et physiologicae, auctore*. CF Koehler.

Wei, X.-X. and Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nature Neuroscience*, 18:1509–1517.

Wolf, C., Bergmann Tiest, W. M., and Drewing, K. (2018). A mass-density model can account for the size-weight illusion. *PloS One*, 13:e0190624.

Xiang, Y., Graeber, T., Enke, B., and Gershman, S. J. (2021). Confidence and central tendency in perceptual judgment. *Attention, Perception, & Psychophysics*, 83:3024–3034.

Yang, T. and Shadlen, M. N. (2007). Probabilistic reasoning by neurons. *Nature*, 447:1075–1080.

Zhu, J.-Q. and Griffiths, T. L. (2026). Computation-limited Bayesian updating: A resource-rational analysis of approximate Bayesian inference. *Psychological Review*, 133(3):619–635.