Bayesian Encoding and Decoding as Distinct Perspectives on Neural Coding

Richard D. Lange, Sabyasachi Shivkumar*, Ankani Chattoraj*, Ralf M. Haefner

Abstract

The Bayesian Brain hypothesis, according to which the brain implements statistically optimal algorithms, is one of the leading theoretical frameworks in neuroscience. There are two distinct underlying philosophies: one in which the brain recovers experimenter-defined structures in the world from sensory neural activity (decoding), and another in which it represents latent quantities in an internal model (encoding). We argue that an implicit disagreement on this point underlies some of the debate surrounding the neural implementation of statistical algorithms, in particular the difference between sampling-based and parametric distributional codes. To demonstrate the complementary nature of the two approaches, we have shown mathematically that encoding by sampling can be equivalently interpreted as decoding task variables in a manner consistent with linear probabilistic population codes (PPCs), a popular decoding approach. Awareness of these differences in perspective helps misunderstandings and false dichotomies, and future research will benefit from an explicit discussion of the relative advantages and disadvantages of either approach to constructing models.

1 Introduction

According to the Bayesian Brain hypothesis, one of the main operations of neural circuits is to carry out statistical computations by flexibly combining prior knowledge with new evidence and evaluating quantities of interest with respect to the entire posterior distribution. In the case of perception, prior knowledge is assumed either to come from experience with the world during development or to be encoded genetically having been learned over the course of generations. While any given sensory measurement may be noisy or ambiguous – providing a wide likelihood function in Bayesian terms – prior knowledge is deployed to resolve these ambiguities when possible (von Helmholtz, 1925). The Bayesian framework has been instrumental for our understanding of perception (Knill and Richards, 1996; Pouget et al., 2013).

At the core of the Bayesian Brain hypothesis is the idea that neural activity corresponds to probability distributions rather than point estimates – such schemes are known as "distributional codes" (Zemel et al., 1998). Previous surveys of distributional codes have emphasized a distinction between sampling-based and parametric codes (Fiser et al., 2010; Pouget et al., 2013; Sanborn, 2015; Gershman and Beck, 2016). From a general theoretical standpoint, both sampling and parametric codes have advantages and disadvantages. In the context of neuroscience, sampling and parametric codes have also been compared with respect to the simplicity of implementing computations believed to be important for the brain, such as cue combination and marginalization (Fiser et al., 2010). Further, numerous studies have empirically tested for properties of sampling or parametric codes in neural responses. Sampling codes have been argued to explain spontaneous cortical activity (Berkes et al., 2011), neural variability (Hoyer and Hyvärinen, 2003), structure in noise correlations (Haefner et al., 2016; Bányai et al., 2019), onset transients and oscillations (Aitchison and Lengyel, 2016; Hennequin et al., 2018; Echeveste et al., 2019), and more (Orbán et al., 2016). Meanwhile, parametric codes have been cited in explanations of contrast-invariant tuning (Ma et al., 2006), near-linearity during cue-combination (Fetsch et al., 2011, 2013), evidence integration dynamics in parietal cortex (Beck et al., 2008; Hou et al., 2019), divisive normalization (Beck et al., 2011), and more (Pouget et al., 2013). Importantly, sampling and parametric codes have so far always been discussed and compared as competing

*equal contribution

and mutually exclusive mathematical models of the same neural circuits, with no decisive evidence presented favoring one over the other model.

Here, we describe how part of this debate can be resolved by considering that sampling and parametric codes, as they are usually discussed, reflect two distinct and *complementary* philosophies on how to construct models of inference in the brain. In particular, the primary goal of this paper is to clearly establish a distinction between what we call **Bayesian Encoding** and **Bayesian Decoding** perspectives on the Bayesian Brain hypothesis. These two perspectives constitute different ways of thinking about the kinds of inference problems faced by the brain and over what variables which inference is performed. Not making these differences explicit has led to confusion about how to interpret neural data. The distinction between an encoding and a decoding perspective has several components, an understanding of which we hope will clarify future research.

We illustrate the complementary nature of these two philosophies using a toy model, previously presented at NeurIPS (Shivkumar et al., 2018). In this example, we *construct* a sampling-based encoding over a linear Gaussian image model (Olshausen and Field, 1996, 1997), and show analytically that firing rates in this model are equivalent to a Probabilistic Population Code (PPC) over arbitrary scalar stimuli in a task. There is thus no inherent contradiction in saying that the brain is *both* sampling (in the "Bayesian Encoding" sense) *and* a parametric code (in the "Bayesian Decoding" sense). We conclude with a discussion of other possible connections between sampling and parametric codes and distributional neural codes in general.

2 Results

Both Bayesian Encoding and Bayesian Decoding fall under the umbrella of *distributional* neural codes. This means that any given pattern of neural activity is interpreted not as representing a point estimate of some quantity, but as representing an entire probability distribution over it. The nature of this "quantity" is key to the distinction between both frameworks.

2.1 Bayesian Encoding

We define **Bayesian Encoding** as the view that there exists a probability distribution over some quantity of interest to the brain, and that the primary function of sensory neurons is to compute and represent an approximation to this distribution. We use the term "encoding" because the probability distribution that is represented conceptually precedes the actual neural responses. That is, in Bayesian encoding models, there exists a reference distribution that is defined independently of how neurons actually respond, and which is approximately encoded by neural responses.

The Bayesian Encoding perspective requires a probabilistic model that defines the reference distribution. In the context of the sensory system, this model often takes the form of an internal generative model of sensory inputs (Figure 1a). With this perspective, the long-term goal of sensory areas of the brain is to develop a statistical model of its sensory inputs. Sensory data, such as an image on the retina, are *explained* as the result of higher order causes. Whereas an image on the retina is high-dimensional and complex, latent variables tell their story: objects, lights, textures, and optics interacted to create each image. A generative model makes this process explicit by assigning prior probabilities to the (co)occurrence of latent variables and by quantifying the likelihood of generating a particular sensory observation from a particular configuration of latent variables. The encoded distribution in this framework is defined over the variables in this statistical model.

For latent variables $\mathbf{x}$ and sensory input $\mathbf{I}$, optimal inference means computing the posterior distribution,

$$\mathrm{p_b}(\mathbf{x}|\mathbf{I}) = \frac{\mathrm{p_b}(\mathbf{I}|\mathbf{x})\mathrm{p_b}(\mathbf{x})}{\mathrm{p_b}(\mathbf{I})}\,. \tag{1}$$

We use the subscript b in $\mathrm{p_b}(\mathbf{x}, \mathbf{I})$ to refer to quantities in the brain's internal model to distinguish them from other types of probabilities such as a decoder's uncertainty. The Bayesian Encoding perspective poses the question of how neural circuits could compute and represent the posterior distribution $\mathrm{p_b}(\mathbf{x}|\mathbf{I})$ for any sensory $\mathbf{I}$, given the internal model that the brain has learned (Figure 1b). In general, exact inference is an intractable problem (Murphy, 2012; Wainwright and Jordan, 2008; Bishop, 2006), leading to the

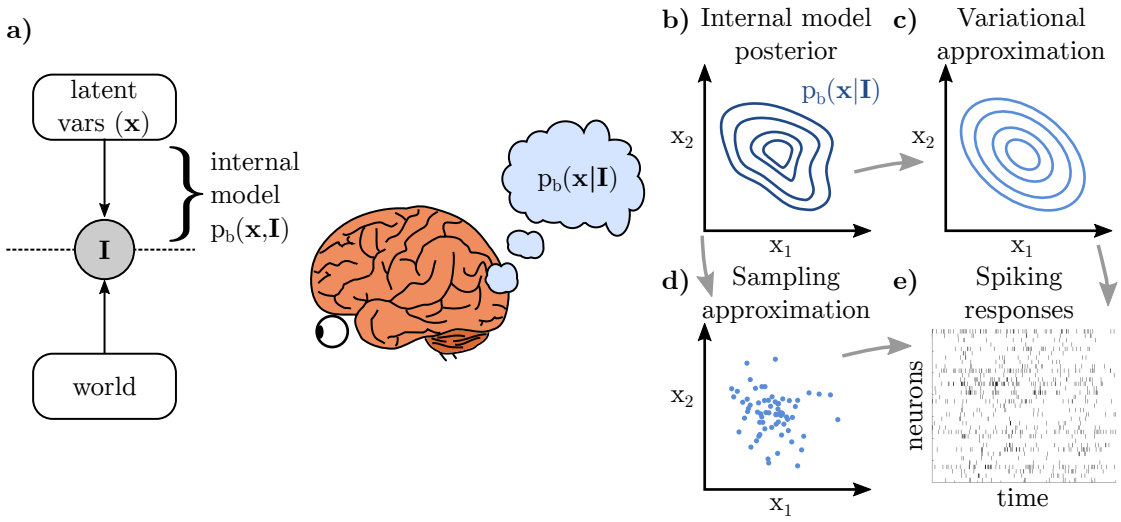

Figure 1: Visualization of Bayesian Encoding. **a)** A common assumption of Bayesian Encoding is that the brain constructs an internal model of the world, and that perceptual inferences are about quantities in the internal model, as opposed to being about external quantities in the world *per se*. This diagram emphasizes this distinction between the world and an internal model. Whether or not stimuli come from natural experience or from an artificial task, the brain computes a posterior over internal variables, $p_b(\mathbf{x}|\mathbf{I})$, in all cases. **b-e)** The defining feature of Bayesian Encoding is the existence of a "true" distribution (b), often the posterior over a latent variable, $\mathbf{x}$, given a sensory measurement, $\mathbf{I}$. One then typically assumes an approximation scheme such as variational inference (b→c) or sampling (b→d), and that this approximation is then realized in patterns of neural activity (e).

question of how the brain could compute and represent an *approximation* to the true posterior (Figure 1c-e). This line of reasoning motivates work on "neurally plausible approximate inference algorithms," including approaches with connections to sampling-based inference (Figure 1d), as well as approaches inspired by variational inference techniques (Figure 1c) (reviewed in Fiser et al. (2010); Sanborn (2015); Gershman and Beck (2016)).

2.2 Bayesian Decoding

We define **Bayesian Decoding** as the perspective in which neural activity is treated as *given*, and emphasis is placed on the statistical uncertainty of a decoder observing those neural responses. Bayesian Decoding is closely related to ideal observer models in psychophysics involving tasks that require the estimation of scalar aspects of a presented stimulus (e.g. its orientation or its contrast) or a decision whether the stimulus belongs to one of two or more discrete classes (e.g. "left" or "right"). Of course, any stimulus s that elicits neural responses $\mathbf{r}$ is optimally decoded by computing $p(s|\mathbf{r})$. In general, this decoder may be complex or sensitive to context or other "nuisance variables." The key question within the Bayesian Decoding framework is this: what conditions must the stimulus-driven neural activity ($p(\mathbf{r}|s)$) fulfill such that the decoder ($p(s|\mathbf{r})$) is both simple (e.g. linear) and invariant to changes in context? For instance, linearity and invariance constraints on the decoder imply constraints on tuning curves and the distribution of neural noise (Zemel et al., 1998; Ma et al., 2006).

There is little practical difference between this definition of Bayesian Decoding and familiar notions of optimal neural decoding, except in one's philosophical stance towards inference in the brain, and hence in the kinds of problems and tools that are emphasized. Classically, decoding is either a tool for assessing information content in neural responses or a mechanistic model of how they impact behavior. In the Bayesian setting, one might further invoke the language of ideal observers and priors. However, contrasting Bayesian versus classical decoding is not pertinent to our main argument; we are instead interested in the distinction of both with Bayesian Encoding.

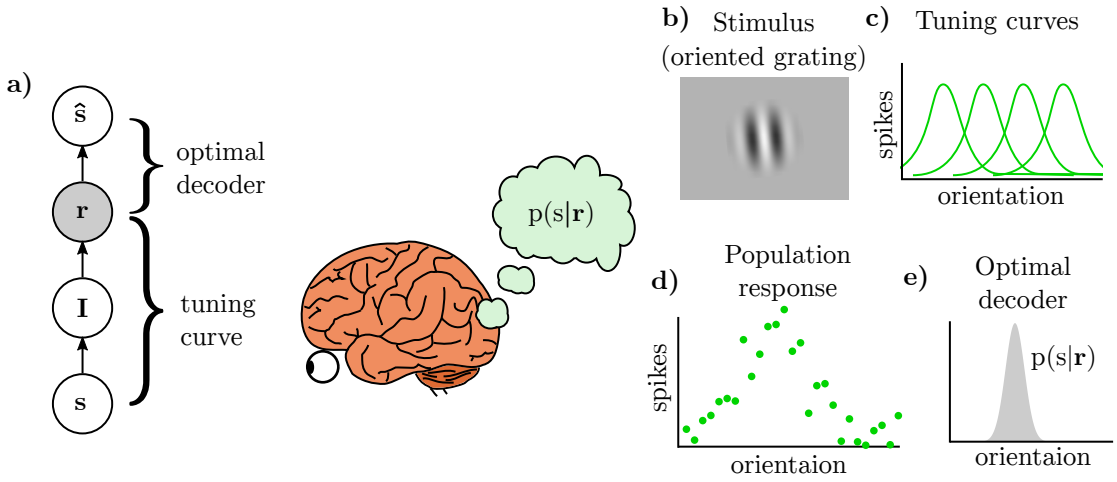

Figure 2: Visualization of Bayesian Decoding. **a)** Decoding is fundamentally a problem of estimating *external* quantities from internal (neural) representations. This diagram emphasizes the symmetry between stimuli that exist in the world, and quantities estimated or inferred in the brain. Here, a scalar stimulus, s, elicits neural responses, vr, mediated by an image, $\mathbf{I}$. The decoding question is how the brain forms an internal estimate, $\hat{s}$, from $\mathbf{r}$. **b)** The decoding problem usually begins with a stimulus, such as the direction of motion of dots viewed through an aperture. **c-e)** Given a population of neurons' tuning curves to s (c) and an observation of spikes on a single trial (d), an optimal decoder computes $\mathrm{p}(s|\mathbf{r})$ (e). A PPC is a decoder with two convenient properties: it is an exponential family with natural parameters linearly related to $\mathbf{r}$, and the decoder is invariant to irrelevant nuisance variables if they only scale the tuning curves.

Probabilistic Population Codes (PPCs), as introduced by Ma et al (2006), exemplify the Bayesian Decoding approach. PPCs imply one way to construct a Bayesian decoder that is both simple and invariant to nuisance: if a population of neurons tuned to s have "Poisson-like" variability, then the optimal decoder is part of the exponential family with firing rates as natural parameters. This is a particularly "convenient" representation for taking products of two distributions (Ma et al., 2006; Beck et al., 2008). Perhaps even more important is the notion of *invariance* afforded by a PPC: as long as nuisance variables such as image contrast or dot coherence only multiplicatively scale tuning curves, the decoder can ignore them.

Importantly, linearity for cue combination and multiplicative gain by nuisance variables are what constitute the *predictions* of PPCs. In classical decoding approaches, neural responses are simply "given," not prescribed by a theory. In the Bayesian Decoding framework generally, and in the case of PPCs in particular, imposing constraints on the decoder constrain the possible set of evoked response distributions, $\mathrm{p}(\mathbf{r}|s)$. These constraints are then formulated as predictions and tested empirically (Fetsch et al., 2011, 2013; Pouget et al., 2013; Hou et al., 2019).

2.3 Contrasting Bayesian Encoding and Bayesian Decoding

There are three key differences between the Bayesian Encoding and Bayesian Decoding perspectives involving (1) what they assume the brain is inferring, (2) implicit notions of causality, and (3) the empirical data and other arguments used to motivate them. As our goal is to summarize and categorize a large and diverse sub-field, there will be exceptions to each rule, but we expect these distinctions to be useful for framing further discussions.

2.3.1 Differences in what is assumed to be inferred

An integral part of the Bayesian Encoding framework is the existence of an abstract internal model that is defined independently of how neurons actually respond. The model is independent of neurons in the sense

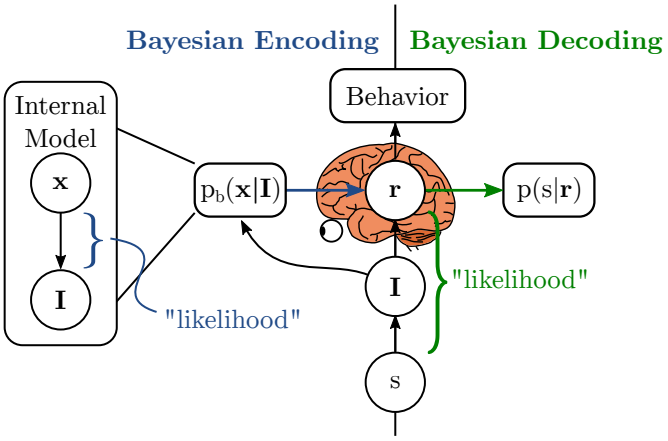

Figure 3: Side-by-side comparison of Bayesian Encoding and Bayesian Decoding. In both frameworks, it is understood that there exists a mechanistic connection between stimuli ($\mathbf{I}$), sensory neural responses ($\mathbf{r}$), and behavior. In the Bayesian Decoding framework, emphasis is placed on the uncertainty of a decoder estimating a (usually scalar) stimulus parameter s from $\mathbf{r}$ (green arrow). Bayesian Encoding posits the existence of an internal model with latent variables $\mathbf{x}$, and that neural responses ($\mathbf{r}$) encode the computation of a posterior distribution ($p_b(\mathbf{x}|\mathbf{I})$). The blue arrow from $p_b(\mathbf{x}|\mathbf{I})$ to $\mathbf{r}$ is an instance of *downward causation*, since changes to the posterior imply changes to neural responses. In Bayesian Decoding, the "likelihood" refers to $p(\mathbf{r}|s)$, and the inference problem is to recover s from $\mathbf{r}$. In Bayesian Encoding, the "likelihood" refers to the internal model's $p_b(\mathbf{I}|\mathbf{x})$, and the inference problem is to recover $\mathbf{x}$ from $\mathbf{I}$ and to embed the posterior over $\mathbf{x}$ in $\mathbf{r}$.

that the same model could in principle be implemented *in silico* or in the brains of other individuals or other species. Translating from inference in an internal model into predictions for neural data usually requires an additional linking hypothesis on the nature of distributional codes, such as whether neurons sample or encode variational parameters, and how either samples or parameters correspond to observable biophysical quantities like membrane potentials, spike times or spike counts.

The brain's internal model is typically assumed to have been calibrated through exposure to natural stimuli (Berkes et al., 2011) and to only change slowly with extensive exposure to new stimuli. For this reason, the generative model in Bayesian Encoding models is often assumed to be task-independent; *what the brain infers* is assumed to not be under the control of an experimenter. One exception to this rule is a family of models in which the *prior* over internal variables changes through extensive exposure to stimuli in a particular task (Haefner et al., 2016; Lange and Haefner, 2020).

In contrast, the Bayesian Decoding view usually deals directly with estimation of task-relevant variables. For instance, in an motion discrimination task, a Bayesian Decoding question would be how the brain represents uncertainty over directions of motion. Importantly, answering this question does not require a generative model of possible motion stimuli; it requires only a statistical model of the relation between scalar motion direction (and possibly nuisance variables like coherence) and neural responses, i.e. $p(\mathbf{r}|s)$. The difference between these perspectives is illustrated in Figure 3.

2.3.2 Differing notions of "likelihood"

Another major difference is evidenced by divergent usage of the term "likelihood" (Figure 3). In Bayesian Encoding, the term "likelihood" is reserved for the abstract relationship between internal model variables and sensory data. For instance, one could speak of the "likelihood that this configuration of variables in the brain's model generated the observed image," or $p_b(\mathbf{I}|\mathbf{x})$. This usage supports the idea that the quantity being computed is a posterior *over internal variables*. In Bayesian Decoding, on the other hand, the "likelihood" refers to a relationship between stimuli and neural responses, $p(\mathbf{r}|s)$. This usage supports the idea that the quantity of interest is the posterior *over external stimuli*.

2.3.3 Differing Empirical and Theoretical Motivations

Finally, distinguishing Bayesian Encoding and Bayesian Decoding allows one to be more precise on what data and what normative arguments motivate different theories. Bayesian Decoding can be motivated by the fact that humans and other species are empirically sensitive to uncertainty and prior experience, as in the classic psychophysics results on multi-modal cue combination (Ernst and Banks, 2002; Knill and Pouget, 2004; Alais and Burr, 2004; Körding, 2007; Pouget et al., 2013). The vast literature on optimal or near-optimal Bayesian perception in controlled tasks motivates the question of how neural circuits facilitate Bayesian computations *with respect to stimuli in a task*. Bayesian Decoding is further motivated by neural data which show a correspondence between neural noise, behavioral indications of uncertainty, and decoding weights in a psychophysics task (Fetsch et al., 2013; Hou et al., 2019; Walker et al., 2019). Importantly, none of these results constitute direct evidence for inference with respect to an internal model, as hypothesized in Bayesian Encoding theories.

There are three motivations for Bayesian Encoding which are independent of the above motivations for Bayesian Decoding. First, there is a constraint on *all* well-calibrated statistical models that the prior must equal the average posterior (Dayan and Abbott, 2001). There is some empirical evidence that this constraint is satisfied by neural responses in visual cortex (Berkes et al., 2011; Lange and Haefner, 2020). Second, one can test for signatures of particular inference algorithms and particular internal models trained on natural stimuli. This approach has been employed by a series of sampling-based inference models and has successfully reproduced a wide range of neural response properties in early visual cortex (Orbán et al., 2016; Aitchison and Lengyel, 2016; Echeveste et al., 2019). Third, Bayesian Encoding is often motivated by purely normative arguments. Any rational agent that faces uncertainty *ought to* compute posterior distributions over unobserved variables (Jaynes, 2003). However, we emphasize again that existing evidence for near-optimality in psychophysical tasks only constitutes weak evidence in favor of inference with respect to a task-independent internal model of the sort usually studied in the Bayesian Encoding literature.

While the Encoding and the Decoding perspectives are complementary, it is important to make this distinction explicit. Failure to do so can lead to confusion and apparently conflicting results on the nature of the neural code. To illustrate this point, we next construct a model that *encodes* the posterior over internal variables by sampling and show analytically that it can be exactly *decoded* in a manner consistent with PPCs. An earlier version of the following section has appeared previously as NeurIPS conference proceedings (Shivkumar et al., 2018).

2.4 Decoding Samples from a Linear Gaussian Model is Equivalent to a PPC

An earlier version of this example originally appeared in the 2018 NeurIPS conference proceedings (Shivkumar et al., 2018). At a high level, our example proceeds as follows: we begin with a linear Gaussian internal generative model and we assume that neurons in V1 approximately infer a posterior distribution over image features. Inference consists of stochastic samples encoded by spiking responses over time. Next, we expose this system to stimuli from a task, such as oriented gratings. We then analytically derive the optimal decoder of task stimuli (e.g. grating orientation) from neural responses, and find that it is a linear PPC. We discuss a variety of implications, including the connection between neural variability and uncertainty and the role of nuisance variables in this system.

2.4.1 Encoding: Neural Sampling in a Linear Gaussian Model

We follow previous work in assuming that neurons in primary visual cortex (V1) implement probabilistic inference in a linear Gaussian model of the input image (Olshausen and Field, 1996, 1997; Hoyer and Hyvärinen, 2003; Bornschein et al., 2013; Haefner et al., 2016):

$$\mathbf{I} \sim \mathcal{N}(\mathbf{A}\mathbf{x}, \mathbf{\Sigma_x}) \tag{2}$$

where $\mathbf{\Sigma_x}$ is the covariance of pixel noise in the brain's generative model. The observed image, $\mathbf{I}$, is assumed to be drawn from a Normal distribution whose mean is a linear combination of "projective fields" ($\mathbf{PF}_i$); the matrix $\mathbf{A}$ is a feature dictionary with projective fields as its columns: $\mathbf{A} = (\mathbf{PF}_1, \ldots, \mathbf{PF}_n)$. Each of the n projective fields is weighted by a single latent variable, $\mathbf{x} = (x_1, \ldots, x_n)^\top$, which will later each be

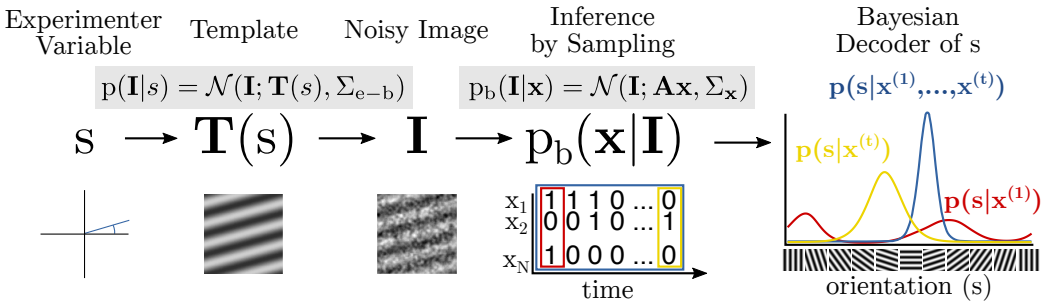

Figure 4: Encoding by sampling followed by decoding of orientation from the samples. Our model performs sampling-based inference over $\mathbf{x}$ in a probabilistic model of the image, $\mathbf{I}$. In a given experiment, the image is generated according to the experimenter's model that turns a scalar stimulus s, e.g. orientation, into an image observed by the brain. The samples drawn from the model are then probabilistically "decoded" in order to infer the implied probability distribution over s from the brain's perspective. While the samples shown here are binary, our derivation of the PPC is agnostic to whether they are binary or continuous, or to the nature of the brain's prior over $\mathbf{x}$.

associated with a single neuron. The main empirical justification for this model consists in the fact that under the assumption of a sparse independent prior over the $\mathbf{x}$, the model learns projective field parameters that resemble the localized, oriented, and bandpass features that characterize V1 neurons when trained on natural images (Olshausen and Field, 1996; Bornschein et al., 2013). Hoyer & Hyvarinen (2003) proposed that during inference neural responses can be interpreted as samples in such a model. Furthermore, Orban et al. (2016) showed that samples from a closely related generative model (Gaussian scale mixture model, (Schwartz and Simoncelli, 2001)) could explain many response properties of V1 neurons beyond receptive fields. Since our main points are conceptual in nature, we will develop them for the slightly simpler original model described above.

Given an image, $\mathbf{I}$, we assume that neural responses correspond to samples from the posterior distribution, $\mathbf{x}^{(t)} \sim p_b(\mathbf{x}|\mathbf{I}) \propto p_b(\mathbf{I}|\mathbf{x})p_b(\mathbf{x})$ where $p_b(\mathbf{x})$ is the brain's prior over $\mathbf{x}$. The exact form of $p_b(\mathbf{x})$ will not matter for the subsequent decoding arguments. We assume that spikes from a population of n neurons encode instantaneous values of samples from the posterior over $\mathbf{x}$, so that each instant, the population response, $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots, x_n^{(t)})^\top$, represents a sample from the brain's posterior belief about $\mathbf{x}|\mathbf{I}$. Each sample of x_i represents the brain' instantaneous belief about the intensity of the feature $\mathbf{PF}_i$ in the image. This interpretation is independent of any task demands or assumptions by the experimenter; as discussed above, $\mathbf{x} \to \mathbf{I}$ is the brain's *internal* model. In the next section we will show how these samples can also be interpreted as a population code over some experimenter-defined quantity like orientation.

2.4.2 Decoding: Inferring Task Stimuli from Samples Results in a PPC

In many classic neurophysiology experiments, an experimenter presents stimuli that only vary along a scalar dimension, such as the orientation of a grating or direction of dot motion (Parker and Newsome, 1998). We call this scalar quantity of interest "s." We then pose the following decoding question: assuming V1 implements sampling-based inference as defined in the previous section, what can downstream areas infer about s by observing the sequence of samples produced by V1? An ideal observer would apply Bayes' rule to infer $p(s|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)}) \propto p(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)}|s)p(s)$ using knowledge of the *likelihood of generating that set of samples* for each s. In the linear Gaussian image model, the optimal decoder can be computed analytically, which we do next.

We assume the image that is observed by the brain's sensory periphery (e.g. retinal ganglion cells) is defined by a template function $\mathbf{T}(s)$ plus noise. This template function could, for instance, represent a grating of a particular spatial frequency and contrast, or any other shape that is being varied along s in the course of the experiment (Figure 4). We further allow for Gaussian pixel noise around the template $\mathbf{T}(s)$ with covariance $\mathbf{\Sigma}_{e-b}$, which accounts for both (e)xternal pixel noise and noise internal to the (b)rain. This

means the likelihood that the brain observes the image $\mathbf{I}$ conditioned on s is

$$\mathrm{p}(\mathbf{I}|s) = \mathcal{N}(\mathbf{I}; \mathbf{T}(s), \mathbf{\Sigma}_{\mathrm{e-b}}) \,, \tag{3}$$

where $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma})$ denotes the probability density of a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance $\mathbf{\Sigma}$ evaluated at $\mathbf{x}$.

With these assumptions, we are able to analytically derive the optimal decoder of s conditioned on a sequence of t independent samples from the posterior, $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)}\}$. By Bayes' rule, the optimal decoder of s is simply the product of the prior $\mathrm{p}(s)$ with the likelihood of generating those t samples conditioned on s. This likelihood term is

$$\mathrm{p}(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)}|s) \propto \mathcal{N}\left(\mathbf{T}(s); \mathbf{A}\bar{\mathbf{x}}_t, \mathbf{\Sigma}_{\mathrm{e-b}} + \frac{1}{t}\mathbf{\Sigma}_{\mathbf{x}}\right) \int \frac{\kappa\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)}\right)}{\mathrm{p_b}(\mathbf{I})^t} \mathcal{N}\left(\mathbf{I}; \boldsymbol{\mu}_{\mathbf{I}}, \mathbf{\Sigma}_{\mathbf{I}}\right) \mathrm{d}\mathbf{I} \,, \tag{4}$$

where $\bar{\mathbf{x}}_t = \frac{1}{t}\sum_{i=1}^{t} \mathbf{x}^{(i)}$ is the average of all samples up to time t. A full derivation along, with the exact form of κ, $\boldsymbol{\mu}_{\mathbf{I}}$, and $\mathbf{\Sigma}_{\mathbf{I}}$ can be found in section S.1 or in Shivkumar et al. (2018). Importantly, as t gets large, $\boldsymbol{\mu}_{\mathbf{I}}$ goes to $\mathbf{A}\bar{\mathbf{x}}_t$, which means that none of the terms in the integral depend on s. In the limit of large t, then, the full decoder of s is given by the much simpler expression,

$$\lim_{t\to\infty} \mathrm{p}(s|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)}) \propto \mathrm{p}(s)\mathcal{N}\left(\mathbf{T}(s); \mathbf{A}\bar{\mathbf{x}}, \mathbf{\Sigma}_{\mathrm{e-b}}\right) \,. \tag{5}$$

Writing this expression in the canonical form for the exponential family gives

$$\lim_{t\to\infty} \mathrm{p}(s|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)}) \quad \propto \quad g(s)\exp(\mathbf{h}(s)^\top \bar{\mathbf{x}}) \qquad \text{where} \tag{6}$$

$$g(s) \quad = \quad \exp\left(-\frac{1}{2}\mathbf{T}(s)^\top \mathbf{\Sigma}_{\mathrm{e-b}}^{-1}\mathbf{T}(s)\right)\mathrm{p}(s) \qquad \text{and} \tag{7}$$

$$\mathbf{h}(s) \quad = \quad \mathbf{T}(s)^\top \mathbf{\Sigma}_{\mathrm{e-b}}^{-1}\mathbf{A} \,. \tag{8}$$

If samples of $\mathbf{x}$ are encoded by instantaneous neural responses, then firing rates $\mathbf{r}$ are proportional to $\bar{\mathbf{x}}$. We can then conclude that, in the limit of large t, this model is equivalent to a linear PPC over s as defined by Ma et al. (2006).

2.4.3 Simulations

We simulated this model system estimating the orientation of a grating image, where the generative model consisted of a mixture of uniformly spaced oriented Gabor patches in the columns of $\mathbf{A}$. Figure 5 shows a numerical simulation of decoded posteriors over s for different numbers of samples, using the large–t decoder of equations (6)-(8), to illustrate how drawing additional samples results in a sharper decoded posterior over s. When only a small number of samples of $\mathbf{x}$ are drawn, the decoded distributions over s are both wide and variable, but get sharper and less variable as the number of samples increases (Figure 5a-c), The black distribution shown in Figure 5d is both the optimal decoder of s in the limit of many samples as well as a PPC over orientation. The bottom row of Figure 5 shows the corresponding spike counts for each neurons on the $y-$axis sorted by the preferred stimulus of each neuron on the $\mathbf{x}-$axis.

2.4.4 The Decoded PPC is Task-Dependent

The relationships that we have derived for $g(s)$ and $\mathbf{h}(s)$ (equations (7) and (8))provide insights into the nature of the PPC that arises in a linear Gaussian model of the inputs. A classic stimulus to consider when probing and modeling neurons in area V1 is an oriented grating. If the images are identical up to rotation, and if the prior distribution over orientations is flat, then $g(s)$ will be constant. Equation (7) shows how $g(s)$ changes as either of those conditions does not apply, for instance when considering stimuli that vary along spatial frequency or binocular disparity, rather than orientation, for which the prior significantly deviates from constant. Further, we can read from equation (8) exactly how the kernels $\mathbf{h}(s)$, which characterize how each neuron contributes to the population code over s, depend both on the manifold of images defined by $\mathbf{T}(s)$, and on the projective fields contained in the columns of $\mathbf{A}$. For an intuition, consider the case of

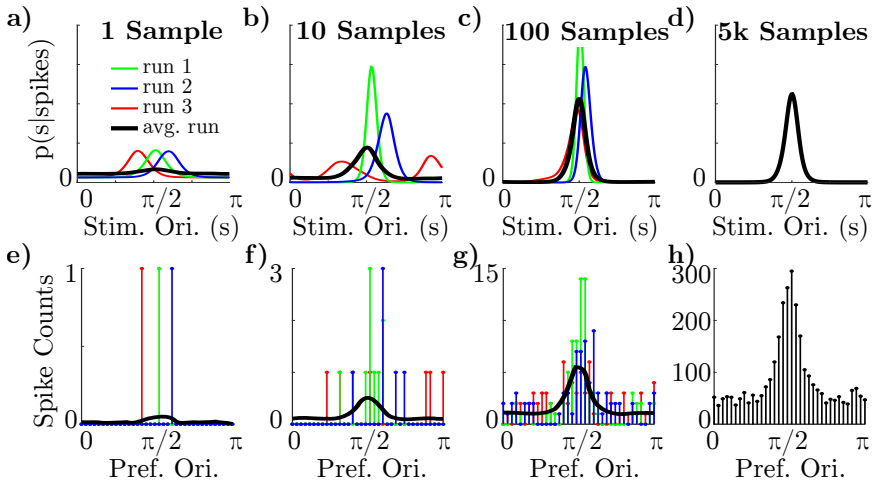

Figure 5: Visualization of the convergence of the decoder after more and more samples. **a-c)** Decoded posterior over s implied by equation (5) for 1, 10, and 100 samples, respectively. Colored lines are individual sampling runs. Black line is the average posterior over many runs. **d)** Decoded posterior over s after 5k samples in black, and with the mean $\bar{\mathbf{x}}$ estimated by Variational Bayes in orange. **e-h)** Population responses corresponding to each panel in (a-d) (note different scales from left to right). The three highlighted runs were selected for visualization post-hoc to ensure the first sample contained only a single spike (e), but this is not true in general for all runs.

isotropic pixel noise, that is $\mathbf{\Sigma}_{\text{e-b}} = \sigma^2_{\text{e-b}}\mathbb{I}$, in which case $\mathbf{h}(s)$ is simply the dot product between $\mathbf{T}(s)$ and $\mathbf{PF}_i$ for each neuron, scaled by $1/\sigma^2_{\text{e-b}}$. The more $\mathbf{T}(s)^\top \mathbf{PF}_i$ depends on s, the more informative neuron i's response is for the posterior over s.

Importantly, the PPC depends as much on the manifold of images defined for a particular experiment, $\mathbf{T}(s)$, as it does on the projective fields of the neurons, $\mathbf{A}$. The kernels $\mathbf{h}(s)$ will be different for gratings of different size and spatial frequency, for plaids, or for a house. This is what we mean when we say the code over s is *task-dependent*: $\mathbf{T}(s)$ is largely arbitrary and up to the experimenter. This means that a downstream area forming an estimate of s, or an area that is combining the information contained in the neural responses $\mathbf{x}$ with that contained in another population (e.g. in the context of cue integration) will need to learn the $\mathbf{h}(s)$ separately for each task.

2.4.5 Simultaneous Log- and Direct-Probability Codes

One way that questions about the nature of Bayesian inference in the brain has been posed is by considering a distinction between Log Probability Codes and Linear or Direct Probability Codes (Barlow, 1969; Pouget et al., 2013). Taking the log of equation (6) reveals that the neural responses in our model are linearly related to the logarithm of the posterior over s. By construction, neural responses in our simple model correspond to samples, i.e. neither probabilities nor log probabilities over $\mathbf{x}$. It is worth noting, however, that samples are proportional to probabilities in the special case where all latent variables are binary. In that case, on the time scale of a single sample, the response is either 0 or 1, making the firing rate of neuron i proportional to its marginal probability, $\text{p}_\text{b}(x_i|\mathbf{I})$. Such a binary image model has been shown to be as successful as the original continuous model of Olshausen & Field (1996) in explaining the properties of V1 receptive fields (Henniges et al., 2010; Bornschein et al., 2013), and is supported by studies on plausible implementations of sampling in spiking neurons (Buesing et al., 2011; Pecevski et al., 2011). This implies that for the special case of binary latents, our neural sampling model is simultaneously a direct probability code (over $\mathbf{x}_i$), and a log probability code (over s).

2.4.6 Dissociating Neural Variability and Uncertainty

It is important to appreciate the difference between the brain's posteriors over $\mathbf{x}$, and over s. The former represents a belief about an *internal* variable such as the intensity or absence/presence of individual image elements in the input. The latter represents knowledge about an external stimulus that caused the input given the neural responses. Neural variability, as modeled here, corresponds to variability in the samples $\mathbf{x}^{(i)}$ and is directly related to the uncertainty in the posterior over $\mathbf{x}$. The uncertainty over s encoded by the PPC, on the other hand, depends on the samples only through their *mean, rather than their variance*. Given sufficiently many samples, the uncertainty over s is only determined by the noise in the channel between experimenter and brain ($\mathbf{\Sigma}_{\mathrm{e-b}}$). This is is a sobering point for experiments that seek to determine whether the brain is sampling by testing the relationship between neural variability and "uncertainty" in broad terms: in our example model, only uncertainty over $\mathbf{x}$ but not over s manifests as neural variability, while s is the thing most commonly and naturally manipulated in an experiment.

3 Discussion

Although it is widely agreed that a primary function of sensory neural circuits is to infer *something*, it is not generally agreed *what* they infer. According to the Bayesian Decoding perspective, neurons represent distributions over external quantities such as stimuli in a task. According to the Bayesian Encoding perspective, neurons represent distributions over variables in an internal model which exists independently of a task. These are complementary perspectives, and the same system might be interpreted as a fundamentally different type of distributional code (sampling or a PPC) depending on what variables we assume the system represents (linear Gaussian features or task stimuli). The question of *how* the brain implements approximate inference is inextricable from the question of *what* it infers.

Historically, sampling-based neural models have taken the Bayesian Encoding approach, asking how neurons could sample from the posterior distribution over variables in an internal model, while PPCs have primarily been associated with Bayesian Decoding. However, this does not reflect any fundamental distinction between the two types of distributional codes. Parametric codes can and have been applied to Bayesian Encoding problems, including both PPCs and other types of parametric codes such as distributed distributional codes (DDCs) (Vertes and Sahani, 2018). Finally, one could consider cognitive sampling models as a kind of sampling-based decoding, which have been used to explain a wide variety of perceptual and cognitive phenomena from multi-stable perception (Gershman et al., 2012) to anchoring and availability biases (Lieder et al., 2013, 2017). Table 1 provides a list of examples in each of the four categories defined by the sampling versus parametric and the encoding versus decoding axes.

Although Bayesian Decoding is not a trivial problem, it is a weaker form of the Bayesian Brain hypothesis than Bayesian Encoding. One might call Bayesian Decoding the **Weak** Bayesian Brain Hypothesis, because it is more descriptive than prescriptive. That is, it describes properties that a neural code ought to have in order to make the job of downstream circuits "easy," and it is relatively tractable to ask whether populations of neurons have those properties – the challenge is to construct $\mathbf{r}|s$ to realize these properties (Zemel et al., 1998; Ma et al., 2006). Bayesian Encoding, on the other hand, might be called the **Strong** Bayesian Brain Hypothesis, because it requires committing to the potentially much harder to falsify idea that the brain contains an internal generative model of its sensory inputs so that the posterior $\mathrm{p_b}(\mathbf{x}|\mathbf{I})$ is unambiguously defined.

In section 2.3.3, we argued that Bayesian Encoding and Bayesian Decoding have largely disjoint empirical and theoretical support. Bayesian Decoding can motivated by the substantial psychophysics literature on near-optimal perception in the face of ambiguity (Knill and Richards, 1996). However, it would be a mistake to treat evidence for near-optimal or near-Bayesian behavior in a particular task alone as evidence that the brain represents probability distributions over variables in an internal model. One could imagine, for instance, extending our example above to the case where the image features, $\mathbf{x}$, are not represented by samples from their posterior, but by their MAP or mean posterior value. This would be a *point estimate* over internal variables and thus antithetical to the idea of Bayesian Encoding, but would nonetheless facilitate many forms of Bayesian Decoding; in fact, neurons encoding only the mean or MAP of $\mathbf{x}$ in our model would directly form a linear PPC over s! If point estimates of internal model variables are sufficient for Bayesian Decoding of task quantities, then Bayesian Encoding requires additional justification *outside* the usually-

	Encoding	Decoding
Sampling	Hoyer and Hyvärinen (2003) Berkes et al. (2011) Buesing et al. (2011) Orbán et al. (2016) Haefner et al. (2016) Aitchison and Lengyel (2016) Savin and Denève (2014)*	Lieder et al. (2013) Vul et al. (2014) Gershman et al. (2012) Sanborn and Chater (2016) Lieder et al. (2017)
Parametric	Friston (2005) Beck et al. (2012) Raju and Pitkow (2016) Vertes and Sahani (2018) Zemel et al. (1998) ? Sahani and Dayan (2003) ? Tajima et al. (2016)? Savin and Denève (2014)*	Ma et al. (2006) Beck et al. (2008) Beck et al. (2011) Hou et al. (2019) Zemel et al. (1998) ? Sahani and Dayan (2003) ? Tajima et al. (2016)?

Table 1: Dividing previous work along the lines of sampling versus parametric codes and encoding versus decoding. The fact that there is previous work in all four quadrants emphasizes that these are complementary distinctions. We marked three papers with "?" that are exceptions to the hard division between encoding and decoding. Savin and Denève (2014), marked with "*", can similarly be seen as an exception to the hard division between sampling-based and parametric encodings.

cited empirical psychophysics literature. The distinction between Bayesian Encoding and Bayesian Decoding might productively add to the open philosophical question: "if perception is probabilistic, why does it not seem probabilistic" (Block, 2018; Rahnev et al., 2020).

An important question for all Bayesian Encoding models is the extent to which they depend on assumptions about the brain's internal model or inference algorithm. As an example, Berkes et al (2011) compared the average stimulus-evoked neural activity in visual cortex to spontaneous activity, finding that they become more aligned over the course of development. This is argued to be evidence that the brain develops an internal statistical model of its sensory inputs in broad terms, since all *well-calibrated* statistical models have the property that the prior is equal to the average posterior (Dayan and Abbott, 2001). However, this link requires making crucial assumptions about the nature of the brain's internal model and its distributional code. First, Berkes et al assume that neural activity encoding the prior can be directly measured by recording spontaneous neural activity, i.e. by recording visual cortex in the dark. This assumption is motivated by the observation that the posterior in scale-mixture models reverts to the prior when contrast is zero, but is in general not true of other types of image models. As an alternative approach to assuming a particular type of internal model, one might instead assert that an internal model *exists* while also conceding that it is *unknown* to us as experimenters. This is the approach taken by Lange & Haefner (2020), who derived predictions for sensory neural activity from the same principle of learning a *well-calibrated* model, but without assuming that the brain's prior can be directly measured.

The key step in our example system above which allowed us to interpret samples of $\mathbf{x}$ as a PPC was to construct the PPC over a different variable – s. Still, the distinction between sampling and parametric codes may also be a false dichotomy *even when considering a single quantity to be inferred.* That is, the question of whether the brain samples or implements variational inference over its internal $\mathbf{x}$ may also lead to a false dichotomy. In principle, it is possible to interpret each sample as implying an entire distribution, and it is possible to improve variational inference by adding stochasticity to the parameters (Hoffman et al., 2013). Current proposals for how the brain could implement probabilistic inference are limited by inference algorithms known from statistics and machine learning, which also tend to divide cleanly into "sampling" or "variational" methods, but rarely both. One way to advance theories of neural inference, then, may be to develop statistical algorithms that trade-off the advantages and drawbacks of both sampling and variational inference (de Freitas et al., 2001; Gershman et al., 2012; Salimans et al., 2015).

References

Laurence Aitchison and Máté Lengyel. The Hamiltonian Brain: Efficient Probabilistic Inference with Excitatory-Inhibitory Neural Circuit Dynamics. *PLOS Computational Biology*, pages 1–24, 2016.

David Alais and David Burr. The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology*, 14(3):257–262, 2004.

Mihály Bányai, Andreea Lazar, Liane Klein, Johanna Klon-Lipok, Marcell Stippinger, Wolf Singer, and Gergő Orbán. Stimulus complexity shapes response correlations in primary visual cortex. *Proceedings of the National Academy of Sciences*, 116(7):2723–2732, 2019.

H. B. Barlow. Pattern Recognition and the Responses of Sensory Neurons. *Annals of the New York Academy of Sciences*, 156(2):872–881, 1969.

Jeffrey M. Beck, Wei Ji Ma, Roozbeh Kiani, Timothy D. Hanks, Anne K. Churchland, Jamie Roitman, Michael N. Shadlen, Peter E. Latham, and Alexandre Pouget. Probabilistic Population Codes for Bayesian Decision Making. *Neuron*, 36(6):1142–1152, 2008.

Jeffrey M. Beck, Peter E. Latham, and Alexandre Pouget. Marginalization in Neural Circuits with Divisive Normalization. *J. Neurosci.*, 31(43):15310–15319, 2011.

Jeffrey M. Beck, Katherine Heller, and Alexandre Pouget. Complex Inference in Neural Circuits with Probabilistic Population Codes and Topic Models. *Advances in Neural Information Processing Systems*, 25:3068–3076, 2012.

Pietro Berkes, Gergo Orbán, Máté Lengyel, and József Fiser. Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment. *Science*, 331(January):83–87, 2011.

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

Ned Block. If perception is probabilistic, why does it not seem probabilistic? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1755), 2018.

Jörg Bornschein, Marc Henniges, and Jörg Lücke. Are V1 Simple Cells Optimized for Visual Occlusions? A Comparative Study. *PLoS Computational Biology*, 9(6), 2013.

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

Peter Dayan and Larry F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press, London, 2001.

Nando de Freitas, Pedro Højen-Sørensen, Michael I. Jordan, and Stuart Russel. Variational MCMC. *UAI*, 2001.

Rodrigo Echeveste, Laurence Aitchison, Guillaume Hennequin, and Máté Lengyel. Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *bioRxiv*, page 696088, 2019.

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

Christopher R. Fetsch, Alexandre Pouget, Gregory C. DeAngelis, and Dora E. Angelaki. Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience*, 15(1):146–54, 2011.

Christopher R. Fetsch, Gregory C. DeAngelis, and Dora E. Angelaki. Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. *Nature Reviews Neuroscience*, 14(6): 429–442, 2013.

József Fiser, Pietro Berkes, Gergo Orbán, and Máté Lengyel. Statistically optimal perception and learning: from behavior to neural representations. *Trends in cognitive sciences*, 14(3):119–30, mar 2010.

Karl J. Friston. A theory of cortical responses. *Philosophical transactions of the Royal Society of London. Series B*, 360:815–836, 2005.

Samuel J. Gershman and Jeffrey M. Beck. Complex Probabilistic Inference: From Cognition to Neural Computation. In Ahmed Moustafa, editor, *Computational Models of Brain and Behavior*, chapter Complex Pr. Wiley-Blackwell, 2016.

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

Ralf M. Haefner, Pietro Berkes, and József Fiser. Perceptual Decision-Making as Probabilistic Inference by Neural Sampling. *Neuron*, 90:649–660, 2016.

Guillaume Hennequin, Yashar Ahmadian, Daniel B Rubin, Máté Lengyel, and Kenneth D Miller. The Dynamical Regime of Sensory Cortex: Stable Dynamics around a Single Stimulus-Tuned Attractor Account for Patterns of Noise Variability. *Neuron*, 98:846–860, 2018.

Marc Henniges, Gervasio Puertas, Jörg Bornschein, Julian Eggert, and Jörg Lücke. Binary Sparse Coding. In Vincent Vigneron, Vicente Zarzoso, Eric Moreau, Rémi Gribonval, and Emmanual Vincent, editors, *Latent Variable Analysis and Signal Separation*, pages 450–457, 2010.

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

Han Hou, Qihao Zheng, Yuchen Zhao, Alexandre Pouget, and Yong Gu. Neural Correlates of Optimal Multisensory Decision Making under Time-Varying Reliabilities with an Invariant Linear Probabilistic Population Code. *Neuron*, 104:1–12, 2019.

Patrik O. Hoyer and Aapo Hyvärinen. Interpreting neural response variability as monte carlo sampling of the posterior. *Advances in Neural Information Processing Systems*, 17(1):293–300, 2003.

E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, New York, 2003.

David C. Knill and Alexandre Pouget. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712–9, dec 2004.

David C. Knill and Whitman Richards, editors. *Perception as Bayesian Inference*. Cambridge University Press, New York, NY, 1996.

Konrad P Körding. Decision Theory: What "Should" the Nervous System Do? *Science Review*, 318, 2007.

Richard D. Lange and Ralf M. Haefner. Task-induced neural covariability as a signature of approximate Bayesian learning and inference. *bioRxiv*, 2020.

Falk Lieder, Thomas L. Griffiths, and Noah D. Goodman. Burn-in , bias , and the rationality of anchoring. *Advances in Neural Information Processing Systems*, 25:1–9, 2013.

Falk Lieder, Thomas L. Griffiths, Quentin J M Huys, and Noah D. Goodman. Empirical Evidence for Resource-Rational Anchoring and Adjustment. *Psychonomic Bulletin & Review*, 2017.

Wei Ji Ma, Jeffrey M. Beck, Peter E. Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438, 2006.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, Cambridge, MA, 2012.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images, 1996.

Bruno a Olshausen and David J. Field. Sparse coding with an incomplete basis set: a strategy employed by V1?, 1997.

Gergő Orbán, Pietro Berkes, József Fiser, and Máté Lengyel. Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex. *Neuron*, 92(2):530–543, 2016.

A J Parker and William T. Newsome. Sense and the Single Neuron: Probing the Physiology of Perception. *Annual Review of Neuroscience of neuroscience*, 21:227–277, 1998.

Dejan Pecevski, Lars Buesing, and Wolfgang Maass. Probabilistic inferences general graphical models through sampling in stochastic networks of spiking neurons. *PLOS Computational Biology*, 7(12), 2011.

Alexandre Pouget, Jeffrey M. Beck, Wei Ji Ma, and Peter E. Latham. Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, 16(9):1170–8, 2013.

Dobromir Rahnev, Ned Block, Janneke Jehee, and Rachel Denison. Is perception probabilistic? In *Cognitive Computational Neuroscience*, 2020.

Rajkumar V. Raju and Xaq Pitkow. Inference by Reparameterization in Neural Population Codes. *Advances in Neural Information Processing Systems*, 30, 2016.

Maneesh Sahani and Peter Dayan. Doubly Distributional Population Codes: Simultaneous Representation of Uncertainty and Multiplicity. *Neural Computation*, 15:2255–2279, 2003.

Tim Salimans, Diederik P. Kingma, and Max Welling. Markov Chain Monte Carlo and Variational Inference: Bridging the Gap. In *International Conference on Machine Learning*, pages 1218–1226, 2015.

Adam N Sanborn. Types of approximation for probabilistic cognition: Sampling and variational. *Brain and Cognition*, 2015.

Adam N Sanborn and Nick Chater. Bayesian Brains without Probabilities. *Trends in Cognitive Sciences*, 20 (12):883–893, 2016.

Cristina Savin and Sophie Denève. Spatio-temporal representations of uncertainty in spiking neural networks. *Advances in Neural Information Processing Systems*, 2014.

Odelia Schwartz and Eero P Simoncelli. Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8):819–825, 2001.

Sabyasachi Shivkumar, Richard D. Lange, Ankani Chattoraj, and Ralf M. Haefner. A probabilistic population code based on neural samples. *Advances in Neural Information Processing Systems*, 31:7070—-7079, 2018.

Chihiro I. Tajima, Satohiro Tajima, Kowa Koida, Hidehiko Komatsu, Kazuyuki Aihara, and Hideyuki Suzuki. Population code dynamics in categorical perception. *Nature Scientific Reports*, 6(22536):1–13, 2016.

Eszter Vertes and Maneesh Sahani. Flexible and accurate inference and learning for deep generative models. *Advances in Neural Information Processing Systems*, 31, 2018.

Hermann von Helmholtz. *Treatise on Physiological Optics*. The Optical Society of America, 1925.

Edward Vul, Noah D. Goodman, Thomas L. Griffiths, and Joshua B. Tenenbaum. One and done? Optimal decisions from very few samples. *Cognitive Science*, 38(4):599–637, 2014.

Martin J. Wainwright and Michael I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Edgar Y Walker, R. James Cotton, Wei Ji Ma, and Andreas S Tolias. A neural basis of probabilistic computation in visual cortex. *Nature Neuroscience*, 23:122–129, 2019.

Richard S. Zemel, Peter Dayan, and Alexandre Pouget. Probabilistic Interpretation of Population Codes. *Neural Computation*, 10(2):403–430, 1998.

S Supplemental Text

S.1 Derivation of the optimal decoder from samples

Here we derive a slightly more general result than is stated in the main text by considering arbitrary covariance matrices: we consider here the case where $\mathbf{I}$ is distributed with mean $\mathbf{T}(s)$ and covariance $\mathbf{\Sigma}_{\mathrm{e-b}}$, and the brain's internal model generates images with mean $\mathbf{Ax}$ and covariance $\mathbf{\Sigma}_{\mathbf{x}}$. The probability of drawing a single neural sample, $\mathbf{x}^{(i)}$, given an observed image is, by assumption, equal to the posterior probability of $\mathbf{x}$ in the brain's internal model. The probability of drawing a sequence of t independent samples of $\mathbf{x}$ is,[1]

$$\begin{aligned}
\mathrm{p}(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}|\mathbf{I}) &= \prod_{i=1}^{t}\mathrm{p}(\mathbf{x}^{(i)}|\mathbf{I}) \\
&= \prod_{i=1}^{t}\frac{\mathrm{p_b}(\mathbf{I}|\mathbf{x}^{(i)})\mathrm{p_b}(\mathbf{x}^{(i)})}{\mathrm{p_b}(\mathbf{I})} \\
&= \frac{1}{\mathrm{p_b}(\mathbf{I})^t}\prod_{i=1}^{t}\mathrm{p_b}(\mathbf{I}|\mathbf{x}^{(i)})\mathrm{p_b}(\mathbf{x}^{(i)})\,.
\end{aligned}$$

Our results primarily follow from this identity for the product of two multivariate normal distributions:

$$\begin{aligned}
\mathcal{N}(\mathbf{y};\mu_1,\mathbf{\Sigma}_1)\mathcal{N}(\mathbf{y};\mu_2,\mathbf{\Sigma}_2) &= \mathcal{N}(\mathbf{y};\mu_3,\mathbf{\Sigma}_3)\mathcal{N}(\mu_1;\mu_2,\mathbf{\Sigma}_1+\mathbf{\Sigma}_2) \qquad \text{(S1)}\\
\mathbf{\Sigma}_3 &= \left(\mathbf{\Sigma}_1^{-1}+\mathbf{\Sigma}_2^{-1}\right)^{-1} \\
\mu_3 &= \mathbf{\Sigma}_3\left(\mathbf{\Sigma}_1^{-1}\mu_1+\mathbf{\Sigma}_2^{-1}\mu_2\right)
\end{aligned}$$

Letting $\bar{\mathbf{x}}_{t'}=\frac{1}{t'}\sum_{i=1}^{t'}\mathbf{x}^{(i)}$ denote the running mean of the samples up to t', it follows from the above product identity that

$$\prod_{i=1}^{t}\underbrace{\mathcal{N}\left(\mathbf{I};\mathbf{Ax}^{(i)},\mathbf{\Sigma}_{\mathbf{x}}\right)}_{\mathrm{p_b}(\mathbf{I}|\mathbf{x}^{(i)})} = \mathcal{N}\left(\mathbf{I};\mathbf{A}\bar{\mathbf{x}}_t,\frac{1}{t}\mathbf{\Sigma}_{\mathbf{x}}\right)\prod_{t'=2}^{t}\mathcal{N}(\mathbf{Ax}^{(t')};\mathbf{A}\bar{\mathbf{x}}_{t'-1},\frac{t'}{t'-1}\mathbf{\Sigma}_{\mathbf{x}})\,. \qquad \text{(S2)}$$

We next absorb all terms that do not depend on s or $\mathbf{I}$ into $\kappa(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)})$. Specifically, let

$$\kappa(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}) = \prod_{i=1}^{t}\mathrm{p_b}(\mathbf{x}^{(i)})\prod_{t'=2}^{t}\mathcal{N}(\mathbf{Ax}^{(t')};\mathbf{A}\bar{\mathbf{x}}_{t'-1},\frac{t'}{t'-1}\mathbf{\Sigma}_{\mathbf{x}})\,.$$

After simplifying further, this can be written in terms of a ratio of Gaussian densities with mean zero, times the product of priors on each $\mathbf{x}$:

$$\kappa(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}) = \frac{\prod_{i=1}^{t}\mathcal{N}(\mathbf{Ax}^{(i)};\mathbf{0},\mathbf{\Sigma}_{\mathbf{x}})\mathrm{p_b}(\mathbf{x}^{(i)})}{\mathcal{N}(\mathbf{A}\bar{\mathbf{x}}_t;\mathbf{0},\frac{1}{t}\mathbf{\Sigma}_{\mathbf{x}})}\,.$$

Then, the likelihood of drawing a particular set of t independent samples of $\mathbf{x}$ conditioned on $\mathbf{I}$ is

$$\mathrm{p}(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}|\mathbf{I}) \;\propto\; \frac{\kappa\left(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}\right)}{\mathrm{p_b}(\mathbf{I})^t}\mathcal{N}\left(\mathbf{I};\mathbf{A}\bar{\mathbf{x}}_t,\frac{1}{t}\mathbf{\Sigma}_{\mathbf{x}}\right)\,. \qquad \text{(S3)}$$

.

Since a decoder looking only at samples of $\mathbf{x}$ has no direct access to the image, the likelihood for a full sequence of samples conditioned on s requires marginalizing over all possible images $\mathbf{I}$ that could be generated conditioned on a fixed s:

$$\mathrm{p}(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}|s) = \int \mathrm{p}(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}|\mathbf{I})\mathrm{p}(\mathbf{I}|s)\mathrm{d}\mathbf{I}\,.$$

[1] We write $\mathrm{p}(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}|\mathbf{I})$ rather than $\mathrm{p_b}(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}|\mathbf{I})$ because while $\mathbf{x}$ is part of the brain's internal model, the *samples* of $\mathbf{x}$ are not, but are viewed through the lens of an outside observer or optimal decoder.

Substituting in (S3), this is

$$\mathrm{p}(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}|s) = \int \frac{\kappa\left(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}\right)}{\mathrm{p_b}(\mathbf{I})^t}\,\mathcal{N}\left(\mathbf{I};\mathbf{A}\bar{\mathbf{x}},\frac{1}{t}\boldsymbol{\Sigma}_{\mathbf{x}}\right)\mathrm{p}(\mathbf{I}|s)\mathrm{d}\mathbf{I}\,.$$

Next, making use of the assumption that $\mathbf{I}|s$ is a multivariate normal centered on $\mathbf{T}(s)$ with pixel covariance $\boldsymbol{\Sigma}_{\mathrm{e-b}}$ and applying the multivariate normal product identity (S1), it follows that

$$\begin{aligned}\mathrm{p}(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}|s) &= \int \frac{\kappa\left(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}\right)}{\mathrm{p_b}(\mathbf{I})^t}\,\mathcal{N}\left(\mathbf{I};\mathbf{A}\bar{\mathbf{x}}_t,\frac{1}{t}\boldsymbol{\Sigma}_{\mathbf{x}}\right)\mathcal{N}(\mathbf{I};\mathbf{T}(s),\boldsymbol{\Sigma}_{\mathrm{e-b}})\mathrm{d}\mathbf{I}\\ &= \mathcal{N}\left(\mathbf{T}(s);\mathbf{A}\bar{\mathbf{x}}_t,\boldsymbol{\Sigma}_{\mathrm{e-b}}+\frac{1}{t}\boldsymbol{\Sigma}_{\mathbf{x}}\right)\int \frac{\kappa\left(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}\right)}{\mathrm{p_b}(\mathbf{I})^t}\,\mathcal{N}\left(\mathbf{I};\boldsymbol{\mu}_{\mathbf{I}},\boldsymbol{\Sigma}_{\mathbf{I}}\right)\mathrm{d}\mathbf{I}\,, \qquad \text{(4 restated)}\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{I}} &= \left(t\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}+\boldsymbol{\Sigma}_{\mathrm{e-b}}^{-1}\right)^{-1}\\ \boldsymbol{\mu}_{\mathbf{I}} &= \boldsymbol{\Sigma}_{\mathbf{I}}\left(\boldsymbol{\Sigma}_{\mathrm{e-b}}^{-1}\mathbf{T}(s)+t\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{A}\bar{\mathbf{x}}_t\right)\,.\end{aligned}$$

As we will show below, the first term in (4), $\mathcal{N}(\mathbf{T}(s);\mathbf{A}\bar{\mathbf{x}}_t,\ldots)$, implies that the decoder is a linear PPC. The integral in (4) requires further discussion. First, note that as the number of samples, t, increases, $\boldsymbol{\Sigma}_{\mathbf{I}}$ shrinks towards zero, and $\boldsymbol{\mu}_{\mathbf{I}}$ goes to $\mathbf{A}\bar{\mathbf{x}}_t$, which implies that $\mathcal{N}(\mathbf{I};\boldsymbol{\mu}_{\mathbf{I}},\boldsymbol{\Sigma}_{\mathbf{I}})$ goes to a delta distribution around $\mathbf{A}\bar{\mathbf{x}}$. This implies that for large t, the integral ceases to depend on s, and hence can be ignored by a decoder. Thus, for large t, we have

$$\mathrm{p}(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}|s) \propto \mathcal{N}\left(\mathbf{T}(s);\mathbf{A}\bar{\mathbf{x}}_t,\boldsymbol{\Sigma}_{\mathrm{e-b}}\right)\,, \qquad \text{(S4)}$$

where the proportionality should be understood in the context of decoding s, and is only approximate for finite t. Note that when t is small, it may still be the case that the integral in (4) does not depend strongly on s. This is the case, for instance, if the brain's internal model assigns equal probability to all $\mathbf{T}(s)$, in which case $\mathrm{p_b}(\mathbf{I})$ evaluated at $\boldsymbol{\mu}_{\mathbf{I}}$ does not depend on s.

Applying Bayes' rule to *decode* s from the samples of $\mathbf{x}$, and absorbing all terms that do not contain s into the proportionality constant, (S4) implies

$$\mathrm{p}(s|\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}) \;\propto\; \mathrm{p}(s)\mathcal{N}(\mathbf{T}(s);\mathbf{A}\bar{\mathbf{x}},\boldsymbol{\Sigma}_{\mathrm{e-b}})\,. \qquad \text{(S5)}$$

We can now rewrite this expression in the canonical form for the exponential family

$$\begin{aligned}\mathrm{p}(s|\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}) &\propto g(s)\exp(\mathbf{h}(s)^\top\bar{\mathbf{x}}) \quad \text{where} && \text{(6 restated)}\\ g(s) &= \exp\left(-\frac{1}{2}\mathbf{T}(s)^\top\boldsymbol{\Sigma}_{\mathrm{e-b}}^{-1}\mathbf{T}(s)\right)\mathrm{p}(s) \quad \text{and} && \text{(7 restated)}\\ \mathbf{h}(s) &= \mathbf{T}(s)^\top\boldsymbol{\Sigma}_{\mathrm{e-b}}^{-1}\mathbf{A}. && \text{(8 restated)}\end{aligned}$$

Equating samples of $\mathbf{x}$ with instantaneous neural responses, the firing rate $\mathbf{r}$ is proportional to $\bar{\mathbf{x}}$. We can then conclude that, in the limit of large t, this model is a linear PPC over s as defined by (Ma et al., 2006).