Bayesian Encoding and Decoding as Distinct Perspectives on Neural Coding

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

Abstract

One of the most influential, and controversial, ideas in neuroscience has been to understand the brain in terms of Bayesian computations. Unstated differences in how Bayesian ideas are operationalized across different models make it difficult to ascertain both which empirical data support which models, and how Bayesian computations might be implemented by neural circuits. In this paper, we make one such difference explicit by identifying two distinct philosophies that underlie existing neural models of Bayesian inference: one in which the brain recovers experimenter-defined structures in the world from sensory neural activity (Decoding), and another in which the brain represents latent quantities in an internal model (Encoding). These philosophies require profoundly different assumptions about the nature of inference in the brain, and lead to different interpretations of empirical data. Here, we characterize and contrast both philosophies in terms of motivations, empirical support, and relationship to neural responses. We also show that this implicit difference in philosophy underlies some of the debate on whether neural activity is better described as a sampling-based, or a parametric, distributional code. Using a simple model of primary visual cortex as an example, we show mathematically that probabilistic inference by neural sampling in the *encoding* framework is, in fact, indistinguishable from a linear probabilistic population code (PPC) in the *decoding* framework. In sum, Bayesian Encoding and Bayesian Decoding are distinct, non-exclusive philosophies, and appreciating their similarities and differences will help organize future work on the nature of inference in the brain.

1 Introduction

According to the Bayesian Brain hypothesis, neural circuits carry out statistical computations by combining prior knowledge with new evidence, combining multiple sources of information according to their reliability, and taking actions that account for uncertainty. 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 evolution. 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 and perceptual decision-making (Knill and Richards, 1996; Kersten et al., 2004; Fiser et al., 2010; 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 computational standpoint, sampling and parametric codes each 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 used to explain spontaneous cortical activity (Berkes et al., 2011), neural variability (Hoyer and Hyvärinen, 2003; Orbán et al., 2016), structure in noise correlations (Haefner et al., 2016; Bányai et al., 2019), and onset transients and oscillations (Aitchison and Lengyel, 2016; Hennequin et al., 2018; Echeveste et al., 2019). Meanwhile, parametric codes

*equal contribution

have been cited in explanations of contrast-invariant tuning curves (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 and mutually exclusive mathematical models of the same neural circuits, with no decisive evidence presented favoring one over the other model.

The primary goal of this paper is to characterize and contrast two distinct perspective on the Bayesian Brain hypothesis. which we call **Bayesian Encoding** and **Bayesian Decoding**. These are complementary perspectives that make different assumptions about the nature of the inference problems faced by the brain, and are supported or falsified by different kinds of empirical data. We argue that not making their differences explicit has led to confusion about how to interpret empirical data. In particular, we describe how the above debate on whether neural responses are better modeled as samples or parameters is complicated by the fact that sampling codes usually make assumptions consistent with Bayesian Encoding while parametric codes often make assumptions consistent with Bayesian Decoding. However, neither the connection between Bayesian Encoding and sampling, nor between Bayesian Decoding and parametric codes, is a necessary consequence of either theory. Indeed, there are Encoding models built on parametric codes, Decoding models based on sampling, and still other models that contain elements of both approaches (Section 2.4 and Table 1 below).

Finally, we demonstrate the complementary nature of these two philosophies using a model of simple cells in primary visual cortex (Shivkumar et al., 2018). In this example, we *construct* a sampling-based encoding model based on a linear Gaussian model of natural images (Olshausen and Field, 1996, 1997), and show analytically that firing rates in this model form a canonical kind of parametric code: a Probabilistic Population Code (PPC). 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 distributional neural codes in general.

2 Results

Both Bayesian Encoding models and Bayesian Decoding models make the assumption of a *distributional* neural code. This means that any given pattern of neural activity is interpreted as representing an entire probability distribution over a variable, not just a point estimate of it. The nature of this "variable" that neurons infer 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 potential 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 neurons are hypothesized to represent 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 this 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 goal of sensory areas of the brain is to learn a statistical model of its sensory inputs (Dayan et al., 1995; Dayan and Abbott, 2001; Fiser et al., 2010; Berkes et al., 2011) in which sensory observations, such as an image on the retina, are explained as the result of higher order causes. Whereas the information on the retina is highly mixed – objects, lights, textures, and optics interact in complex ways to create an image – the internal model aims to explain sensory data in terms of unobserved causes that are often assumed to be sparse and independent (von Helmholtz, 1925; Olshausen and Field, 1996; Bell and Sejnowski, 1997). A generative model makes this process explicit by assigning prior probabilities to the (co)occurrence of causes (represented by latent variables) and by quantifying the likelihood of generating a particular sensory observation from a particular configuration of the causes. The encoded posterior distribution in this framework is defined over the latent variables in this statistical model.

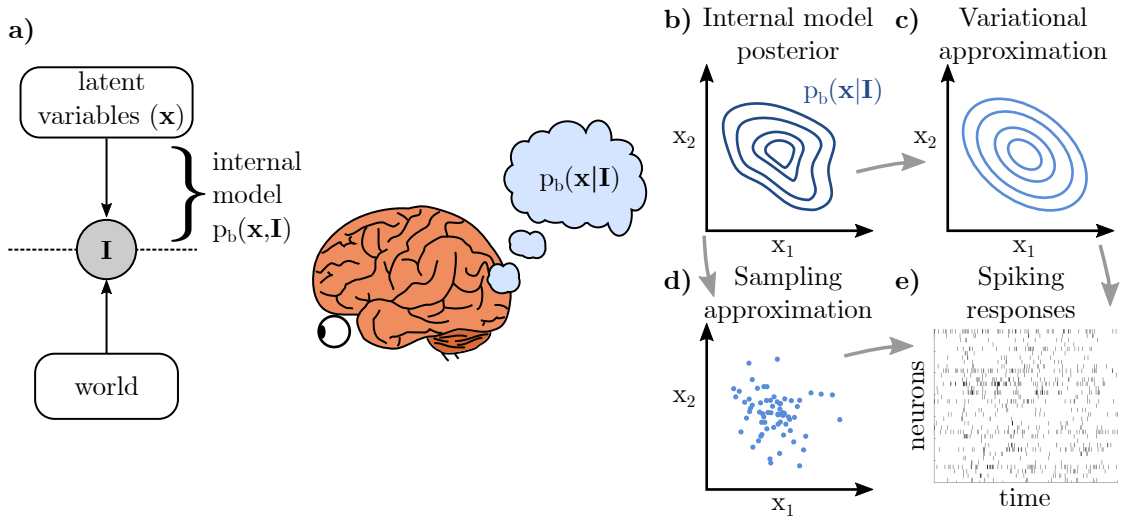

Figure 1: Visualization of Bayesian Encoding. **a)** Bayesian Encoding assumes that the brain has an internal model of its inputs, and that perceptual inferences are about variables in this internal model, not necessarily corresponding to quantities in the external world *per se*. This diagram emphasizes this distinction between the world and the internal model. Whether or not stimuli come from natural experience or from an artificial task, the brain always computes a posterior over internal variables, $p_b(\mathbf{x}|\mathbf{I})$. **b-e)** The defining feature of Bayesian Encoding is the existence of a reference distribution (b), often the posterior over a set of latent variables, $\mathbf{x}$, given a sensory measurement, $\mathbf{I}$. One then 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).

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

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

We use the subscript b in $p_b(\cdot)$ to refer to quantities in the brain's internal model, and 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 $p_b(\mathbf{x}|\mathbf{I})$ for any sensory $\mathbf{I}$, given the internal model that the brain has learned (Figure 1b), and how it can learn this internal model in the first place. In general, exact inference is an intractable problem (Murphy, 2012; Wainwright and Jordan, 2008; Bishop, 2006), leading to the question of how the brain could compute and represent an *approximation* to the true posterior (Figure 1c-e), and what the nature of this approximation is. 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, related to parametric neural codes (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 $p(s|\mathbf{r})$ (Figure 2). In general, this decoder may be prohibitively complex or it may be 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

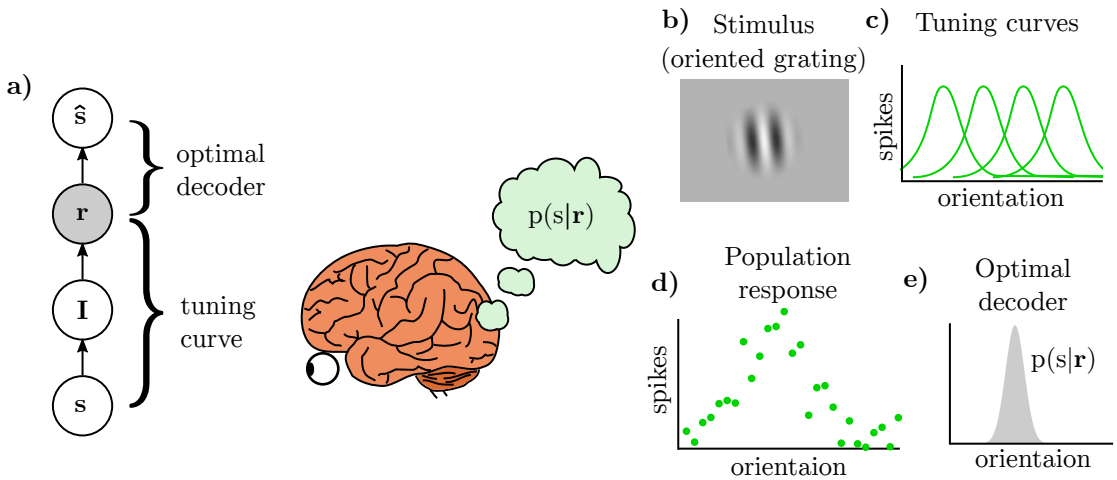

Figure 2: Visualization of Bayesian Decoding. **a)** Decoding is fundamentally a problem of estimating *external* quantities from internal (neural) representations. This diagram emphasizes the symmetry between stimuli that exist in the world, and quantities estimated or inferred in the brain. Here, a scalar stimulus, s, parameterizes the sensory input, $\mathbf{I}$, that elicits neural responses, $\mathbf{r}$. 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 orientation of a grating of a given spatial frequency, size, location, and contrast. **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).

($\mathrm{p}(\mathbf{r}|s)$) fulfill such that the decoder ($\mathrm{p}(s|\mathbf{r})$) is *simple* (e.g. linear and invariant to nuisance)? For instance, imposing linearity and invariance constraints on the decoder implies constraints on tuning curves and the distribution of neural noise (Zemel et al., 1998; Ma et al., 2006).

Bayesian Decoding is closely related to familiar notions of optimal neural decoding. 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, the emphasis is on how neural activity is interpreted by the rest of the brain and influences behavior, and how this depends on the brain's uncertainty about a behaviorally-relevant stimulus.

Probabilistic Population Codes (PPCs), as introduced by Ma et al (2006), exemplify the Bayesian Decoding approach. PPCs construct a Bayesian decoder that is both simple and invariant to nuisance: if a population of neurons tuned to s has "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 as required by cue-integration (Ma et al., 2006) and evidence accumulation Beck et al. (2008). Equally 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, under the assumption that the brain employs a computationally convenient neural code, linearity for cue combination and multiplicative gain by nuisance variables become *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 constrains the possible set of evoked response distributions, $\mathrm{p}(\mathbf{r}|s)$. These constraints have then been formulated as predictions and tested empirically (Fetsch et al., 2011, 2013; Pouget et al., 2013; Hou et al., 2019).

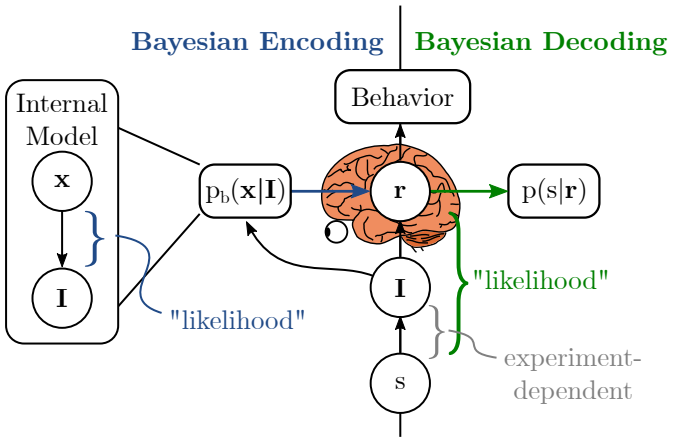

Figure 3: Side-by-side comparison of Bayesian Encoding and Bayesian Decoding. In both frameworks, it is understood that there exists a mechanistic, biophysical connection between stimuli ($\mathbf{I}$), sensory neural responses ($\mathbf{r}$), and behavior. In the Bayesian Decoding framework, emphasis is placed on the uncertainty of a decoder estimating a 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 $\mathbf{r}$ can be seen as an instance of *downward causation* between levels of abstraction, where changes to the posterior (at the algorithmic level) imply changes to neural responses (at the implementation level) (Campbell, 1974; Yablo, 1992). 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 $\mathbf{r}$. In any psychophysical task, the link between s and $\mathbf{I}$ depends on the experiment (gray bracket). Importantly, this places the "likelihood" in a Bayesian Decoding model indirectly under the control of the experimenter (through their choice of stimuli), but not in a Bayesian Encoding model.

2.3 Contrasting Bayesian Encoding and Bayesian Decoding

There are three key differences between the Bayesian Encoding and Bayesian Decoding perspectives: (1) what they assume the brain is inferring, (2) what the terms "likelihood" and "posterior" refer to, 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 could in principle be implemented *in silico* or in the brains of other individuals or other species. Deriving predictions for neural data 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 and spike times or spike counts. Bayesian Encoding thus decomposes the question of what sensory neurons compute into two parts: first, what is the internal model which defines optimal inference (the reference distribution), and second, how do neural circuits carry out approximate inference in that model (e.g. sampling or parametric)?

The brain's internal model is typically assumed to have been calibrated through exposure to natural stimuli (Dayan et al., 1995; Dayan and Abbott, 2001; Berkes et al., 2011) and to only change slowly with exposure to new stimuli in adult brains. For this reason, the generative model in Bayesian Encoding models of early sensory areas is often assumed to be independent of experimental context. For instance, if the brain's internal model comprises patches of local image features, then it is assumed that the brain infers and encodes the same set of image features, whether viewing natural scenes or artificial stimuli in a task (Haefner et al., 2016; Orbán et al., 2016; Shivkumar et al., 2018). The assumption of calibration in a Bayesian Encoding

framework also makes predictions for how the internal model should change in response to the statistics of sensory inputs during development (Berkes et al., 2011), and to extensive exposure to stimuli in a particular task (Lange and Haefner, 2020).

In contrast, Bayesian Decoding models usually directly estimate task-relevant variables. For instance, in a 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 all possible naturally-occurring motion stimuli; it requires only a statistical model of the relation between scalar motion direction in a particular task (and possibly nuisance variables like coherence) and neural responses, $p(\mathbf{r}|s)$. The difference between the Bayesian Encoding and Decoding 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 observations. For instance, one could speak of the "likelihood that this configuration of variables in the brain's model generated the observed image," or $\mathrm{p_b}(\mathbf{I}|\mathbf{x})$. 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, $\mathrm{p}(\mathbf{r}|s)$. This usage supports the idea that the quantity of interest is the posterior *over external stimuli in a task*.

2.3.3 Conceptual differences in the posterior

The posteriors resulting from the Encoding and the Decoding approach differ in the uncertainties they incorporate. The Encoding posterior, $\mathrm{p_b}(\mathbf{x}|\mathbf{I})$, is defined entirely by the brain's internal model and observations, independent of any noise in how this posterior is represented, e.g. by neurons and neural circuits that may suffer from biophysical noise. On the other hand, the Decoding posterior, $\mathrm{p}(s|\mathbf{r})$, by virtue of being conditioned on those potentially noisy responses, also depends on implementation noise. Consequently, as the temporal window over which neural responses are decoded is lengthened, the Decoding posterior sharpens even while the Encoded posterior stays constant (see section 2.5.4 for an example).

This conceptual difference points to an important practical difference in how neural responses are interpreted by downstream areas in both approaches. In a Decoding model, a downstream area applies Bayes' Rule to the neural responses arriving from an upstream area to extract information about a stimulus (e.g. Ma et al. (2006); Beck et al. (2008)). In an Encoding model, on the other hand, upstream neural activity represents samples or parameters that are then processed by the downstream area according to an underlying approximate inference algorithm, which generally will *not* apply Bayes' rule to the incoming activity directly (e.g. Haefner et al. (2016); Raju and Pitkow (2016)). To put it another way, if one assumes that upstream neural activity encodes samples or parameters in an approximate inference algorithm, then there is an important conceptual difference between a downstream area that interprets upstream activity *as samples* or *as parameters* (as in Encoding models), and a downstream area that *decodes* the activity it receives by applying Bayes' rule to the neural activity.

2.3.4 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 large 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*, which are often scalar or low-dimensional. With the additional assumption that the neural representation of task-relevant aspects of stimuli is formatted to be easily decoded (e.g. linear and invariant to nuisance (Ma et al., 2006)), this line of reasoning has given rise to predictions for neural data. These predictions have since been largely confirmed for the representation of self-motion in dorsal medial superior temporal area (MSTd) (Fetsch et al., 2011, 2013; Hou et al., 2019).

	Bayesian Encoding	**Bayesian Decoding**
Sampling-based representation	Hoyer and Hyvärinen (2003) Pecevski et al. (2011) Berkes et al. (2011) Buesing et al. (2011) Probst et al. (2015) Orbán et al. (2016) Haefner et al. (2016) Aitchison and Lengyel (2016) Savin and Denève (2014)	Moreno-Bote et al. (2011)† Gershman et al. (2012)
Parametric representation	Friston (2005) George and Hawkins (2009) Beck et al. (2012) Raju and Pitkow (2016) Vertes and Sahani (2018) Zemel et al. (1998) * Sahani and Dayan (2003) * Tajima et al. (2016)*	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)* Moreno-Bote et al. (2011)†

Table 1: Classifying previous work on Bayesian neural models according to whether they construct Bayesian Encoding or Decoding models, and whether they use a sampling-based or a parametric neural representation. Papers marked with "*" are exceptions to the hard division between encoding and decoding. Moreno-Bote et al., marked with "†", contains elements of both sampling-based and parametric decoding.

Bayesian Decoding is further motivated by neural data showing a correspondence between non-parametric likelihood functions, neural noise, and behavioral indications of uncertainty (Walker et al., 2019).

Importantly, none of these results constitute direct evidence for inference with respect to an (usually high-dimensional) internal model of natural stimuli, as hypothesized in Bayesian Encoding theories (Rahnev, 2019). The three lines of support for Bayesian Encoding models are largely independent of the above motivations for Bayesian Decoding. First, Bayesian Encoding can be motivated by the purely normative argument that any rational agent that faces uncertainty *ought to* compute probability distributions over unobserved variables, as long as those variables directly enter into calculations of expected utility (Jaynes, 2003). Second, there is some empirical evidence for a key prediction of Bayesian encoding models: a general constraint on *all* well-calibrated statistical models is that the prior must equal the average posterior (Dayan and Abbott, 2001). Existing observations suggest that this constraint is satisfied in primary visual cortex, as evidenced by changes in neural responses in primary visual cortex over development (Berkes et al., 2011) and task-learning (Lange and Haefner, 2020). Third, there is empirical evidence for signatures of particular inference algorithms and particular internal models fit to 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). A similar approach has also been taken by parametric models, where neural circuits have been hypothesized to implement the dynamics of a variational inference algorithm (Friston, 2005; George and Hawkins, 2009; Beck et al., 2012; Grabska-Barwinska et al., 2013; Raju and Pitkow, 2016; George et al., 2018; Lavin et al., 2018). Finally, we emphasize again that existing evidence for Bayesian-like behavior in psychophysical tasks only constitutes weak evidence in support of the idea that the brain computes distributions over variables in a task-independent internal model, as usually studied in the Bayesian Encoding literature (Rahnev, 2019).

2.4 Classification of existing models

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 studied in the context of inference of low-dimensional task-relevant quantities. However, this does not reflect a fundamental distinction between the two types of distributional codes. Parametric codes can and have been used in Bayesian Encoding models, including Probabilistic Population Codes (PPCs) (Beck et al.,

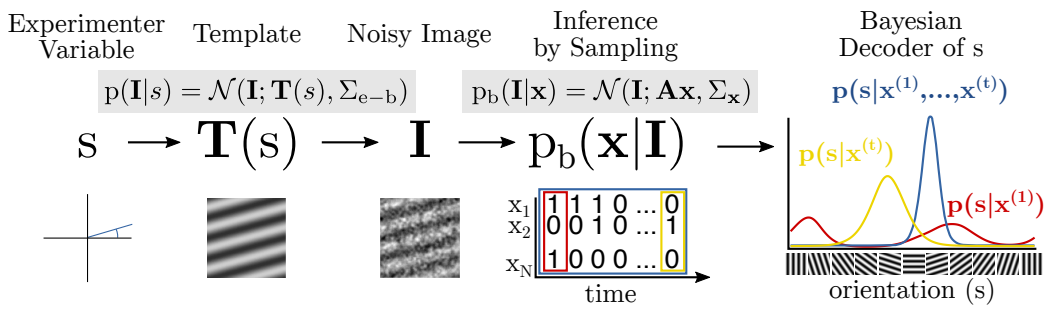

Figure 4: Encoding by sampling followed by decoding of orientation from the samples. Our model performs sampling-based inference over $\mathbf{x}$ in a probabilistic model of the image, $\mathbf{I}$. In a given experiment, the image is generated according to the experimenter's model that turns a scalar 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. While the samples shown here are binary, our derivation of the PPC applies whether they are binary or continuous, and regardless of the nature of the brain's prior over $\mathbf{x}$.

2012; Grabska-Barwinska et al., 2013; Raju and Pitkow, 2016), Distributed Distributional Codes (DDCs) (Vertes and Sahani, 2018), and others (Friston, 2005; George and Hawkins, 2009; Lavin et al., 2018; George et al., 2018). Markov Chain Monte Carlo (MCMC) sampling has been used to explain perceptual bistability (Moreno-Bote et al., 2011; Gershman et al., 2012), which could be seen as a form of sampling-based Bayesian Decoding (cf. Hohwy et al. (2008)). To summarize, 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. The fact that there is previous work in all four quadrants emphasizes that these are complementary distinctions.

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

Importantly, the Bayesian Encoding and the Bayesian Decoding perspectives are complementary and do not necessarily lead to mutually exclusive models. To demonstrate this point, we next construct a Bayesian Encoding model that encodes the posterior over internal variables by sampling and show analytically that it is equivalent to a Bayesian Decoding model based on a parametric representation (PPC). An earlier version of the following section has appeared previously as conference proceedings (Shivkumar et al., 2018).

2.5.1 Bayesian 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):

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

where $\mathbf{\Sigma}_{\mathbf{x}}$ 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 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, Orbán 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, $\mathbf{x}^{(t)} \sim \mathrm{p_b}(\mathbf{x}|\mathbf{I}) \propto \mathrm{p_b}(\mathbf{I}|\mathbf{x})\mathrm{p_b}(\mathbf{x})$ where $\mathrm{p_b}(\mathbf{x})$ is the brain's prior over $\mathbf{x}$. The exact form of $\mathrm{p_b}(\mathbf{x})$ 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's instantaneous belief about the intensity of the feature $\mathbf{PF}_i$ in the image. This interpretation is independent of the experimenter's choice of sensory inputs, and along what dimension to vary them; as discussed above, $\mathbf{x} \to \mathbf{I}$ is the brain's *internal* model. In the next section we will show how these samples form a (parametric) probabilistic population code (PPC) over some experimenter-defined quantity like orientation.

2.5.2 Bayesian Decoding: Inferring Task Stimuli from Samples Results in a PPC

In many classic neurophysiology experiments, an experimenter presents stimuli that vary along a scalar dimension, such as the orientation of a grating or direction of dot motion (Parker and Newsome, 1998). We denote this scalar quantity of interest by "s." Assuming V1 implements sampling-based inference as defined in the previous section, we now construct a Bayesian decoding model by asking the decoding question: what can downstream areas infer about s by observing the sequence of samples produced by V1? An ideal observer applies Bayes' rule to infer $\mathrm{p}(s|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)})$ using knowledge of the probabilistic relationship between samples and s. In the linear Gaussian image model, the optimal decoder can be computed analytically, which we do next.

We assume that the input at the brain's sensory periphery (e.g. the image on the retina) 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 $\mathbf{\Sigma}_{\mathrm{e-b}}$, 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 we obtain

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

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

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

If samples of $\mathbf{x}$ are encoded by instantaneous neural responses, then firing rates $\mathbf{r}$ are proportional to $\bar{\mathbf{x}}$. Replacing $\bar{\mathbf{x}}$ with $\mathbf{r}$ in equation (6), we see that in the limit of large t, the posterior over s is equivalent to a PPC over s as defined by Ma et al. (2006). In line with intuition, the kernels $\mathbf{h}(s)$ can be interpreted as quantifying the match between the stimulus template $\mathbf{T}(s)$ and the projective fields for the population, $\mathbf{A}$. If, on a given trial, a particular neuron with projective field $\mathbf{PF}_i$ is highly active, then sampling interpretation is that $\mathbf{PF}_i$ is likely a constituent feature of the image, and the PPC interpretation is that s is likely somewhere that $\mathbf{T}(s)$ "matches" $\mathbf{PF}_i$ in the sense of equation (8).

2.5.3 The Decoded PPC is Experiment-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 isotropic pixel noise, that is $\mathbf{\Sigma}_{\mathrm{e-b}} = \sigma^2_{\mathrm{e-b}}\mathbb{I}$, in which case $\mathbf{h}(s)$ is simply the dot product between $\mathbf{T}(s)$ and $\mathbf{PF}_i$ for each neuron, scaled by $1/\sigma^2_{\mathrm{e-b}}$. 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 structure of this PPC depends not only on the projective fields of the neurons, $\mathbf{A}$, but also on the manifold of images defined for a particular experiment, $\mathbf{T}(s)$. The kernels $\mathbf{h}(s)$ will be different for gratings of different size and spatial frequency, for plaids, or for a house. This demonstrates how this code for s, implied by the Bayesian encoding model, is *experiment-dependent*: $\mathbf{T}(s)$ is largely arbitrary and defined by 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 experiment/task.

2.5.4 Time dependence of the posterior in the Bayesian Decoding model

Our example illustrates the difference in time-dependence between the posteriors implied by the Bayesian Encoding and the Bayesian Decoding approaches (section 2.3.3): while the Bayesian Encoding posterior over $\mathbf{x}$ is constant as a function of time, the Bayesian Decoding posterior over s changes dramatically, sharpening as a function of time. 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, i.e. the counting window over which spikes are decoded is small, the decoded distributions over s are both wide and variable from run to run. As the counting window increases in size and spike counts increase (5e-h), the decoded posterior gets sharper and less variable (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 linear PPC over orientation.

3 Discussion

Despite wide agreement that a primary function of sensory neural circuits is to infer *something*, it is not generally agreed *what* they infer. To clarify this question, we have identified and characterized a new distinction between two approaches to constructing Bayesian neural models: Bayesian Encoding and Bayesian Decoding. This distinction is orthogonal to existing and much-debated distinctions like whether neural responses reflect parameters or samples of the inferred distribution. Making the distinction between Bayesian Encoding and Bayesian Decoding explicit has provided us with new insights into the long-standing debate about the nature of the neural code. Importantly, we have demonstrated that these two approaches can give rise to different models of the same neural circuit, underlining the point that Bayesian Encoding and Decoding

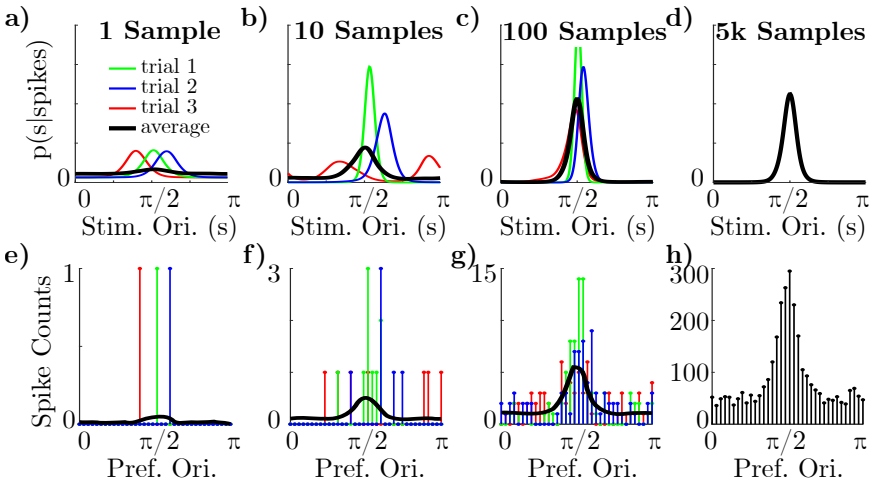

Figure 5: Visualization of the convergence of the decoded posterior as a function of time (samples). **a-c)** Decoded posterior over s implied by equation (5) for 1, 10, and 100 samples, respectively. Colored lines are individual trials (separate sampling runs). Black line is the average posterior over many trials. **d)** Decoded posterior over s after 5k samples in black. **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.

models are complementary, and not mutually exclusive. The complementary nature of these approaches has direct implications for both theoretical debates and the correct interpretation of empirical data.

Our results also provide insights into the much-debated question of whether neural responses are more closely related to parameters of the encoded probability distribution, as in probabilistic population codes (PPC; Ma et al. (2006)) and in distributed distributional codes (DDC; Vertes and Sahani (2018)), or to samples from the distribution as in neural sampling (reviewed in Fiser et al. (2010); Pouget et al. (2013); Sanborn (2015); Gershman and Beck (2016)). In our example, the Bayesian Decoding model implies a (parametric) PPC while, by construction, neural responses in the Bayesian Encoding model represent samples. In our example system, it is *in principle* impossible to distinguish between these two models on the basis of empirical data, since they are not mutually exclusive. A productive direction forward will be to contrast models that take the same Encoding or Decoding perspective, and that can therefore be considered to be "competing." For instance, parametric models that take a Bayesian Encoding approach (Friston, 2005; George and Hawkins, 2009; Beck et al., 2012; Grabska-Barwinska et al., 2013; Raju and Pitkow, 2016; Vertes and Sahani, 2018) are comparable to Encoding-based neural sampling models.

Our results also address one of the oldest questions about neural representations of probability: whether neural responses are more closely related to probability (direct probability codes, (Barlow, 1969; Anderson and Van Essen, 1994)) or its logarithm (log probability codes, (Jazayeri and Movshon, 2006; Ma et al., 2006)). Our results demonstrate that both can be true at the same time, provided that one is implemented in a Bayesian Encoding model, and the other in a Decoding model: in our example, the Bayesian Decoding model is a linear PPC, which means that neural responses are linearly related to the logarithm of the distribution over s. At the same time, by construction, neural responses in our example correspond to samples, i.e. neither probabilities nor log probabilities over $\mathbf{x}$. However, for the special case where all latent variables are binary, with each sample corresponding to either 1 or 0 (spike or no spike), the firing *rate* of neuron i becomes proportional to the marginal probability, $\mathrm{p_b}(x_i|\mathbf{I})$. Thus, for binary $\mathbf{x}$, our example model is simultaneously a direct probability code over $\mathbf{x}$ and a log probability code over s. 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).

More generally, our derivation also raises questions about what makes a neural code "distributional", i.e. representing a whole distribution rather than just a point estimate, and what would constitute empirical ev-

idence for it. While the Bayesian Encoding model in our example assumed a sampling-based representation of the posterior over $\mathbf{x}$, consider a reduced, non-distributional version in which neural responses are proportional to a point estimate of $\mathbf{x}$ such as its mean or mode (Olshausen and Field, 1996). Even this reduced model gives rise to a probabilistic code (PPC) over s. Such a *point estimate* code over variables in the brain's internal model would still enable many of the apparently Bayesian behaviors observed in low-dimensional psychophysics tasks and used to motivate Bayesian Decoding theories, as discussed in section 2.3.4 above. It would therefore be a mistake to treat empirical 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 of sensory inputs. The distinction between Bayesian Encoding and Bayesian Decoding might thus productively add to the open philosophical question: "if perception is probabilistic, why does it not seem probabilistic?" (Block, 2018; Rahnev et al., 2020).

The difference between Bayesian Encoding and Bayesian Decoding models and, more generally, the choice of variable which is assumed to be inferred, also impacts the interpretation of neural variability. In our example above, neural variability is directly related to the uncertainty in the posterior over $\mathbf{x}$. In contrast, the uncertainty over s encoded by the Bayesian decoding model is unrelated to the neural variability, depending 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 an important point for experiments that seek to test the neural sampling hypothesis by relating neural variability and "uncertainty": in our example model, only uncertainty over $\mathbf{x}$ but not over s manifests as neural variability, while s is the variable most commonly and naturally manipulated in an experiment.

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. This raises the question: what if s is part of $\mathbf{x}$? In other words, could the brain infer and *encode* a distribution over the stimulus in a task? We offer two remarks on this question. First, even if s is part of the brain's internal model, Bayesian Encoding and Decoding models would nonetheless differ in their approach to the question of how neural responses, $\mathbf{r}$, relate to the distribution on s. Bayesian Encoding models would begin with a generative model of sensory input $\mathbf{I}$ from s and ask how the true posterior $\mathrm{p_b}(s, \mathbf{x}|\mathbf{I})$ is represented by neural responses $\mathbf{r}$. Importantly, s would be inferred together with all the other internal variables, both in controlled tasks and with natural stimuli. Bayesian Decoding models, on the other hand, would investigate the relationship between s in the world and evoked neural responses, $\mathrm{p}(\mathbf{r}|s)$, and study a different kind of posterior, $p(s|\mathbf{r})$, which takes the perspective of the experimenter, or possibly the rest of the brain trying to read out s from $\mathbf{r}$. If the *decoded* distribution, $\mathrm{p}(s|\mathbf{r})$, matches the ideal *encoded* distribution, $\mathrm{p_b}(s|\mathbf{I})$, then the code for s is said to be *efficient* (Beck et al., 2012).

Second, s being part of $\mathbf{x}$ would beg the question: how can an experimenter-defined stimulus (s) enter into the brain's internal model? One possibility is that "orientation" (or any other s) is a useful abstraction of natural stimuli, in which case it may have been learned (or evolved) and may permanently be a part of the brain's internal model. Another possibility is that orientation (or any other s) is part of the brain's internal model because the brain changes its internal model on the fly to accommodate the demands of the present task. Such flexibility may be present in later stages of sensory processing. Earlier in sensory processing, however, what a neuron computes is usually assumed to be constant over time; if a population of neurons encodes $\mathbf{x} = 1$ for one input, then it might encode a different *value* like $\mathbf{x} = 2$ for another input, but does not switch to encoding a different *variable* like $\mathbf{y} = 1$. The assumption that the brain's *internal model* does not change depending on the task is essential to generalize from the laboratory to natural settings, and is at the heart of the Bayesian Encoding philosophy.

The presented insights related to the Bayesian Encoding and Decoding perspective also have implications for Bayesian models of cognition, where ideas related to both sampling (Vul and Rich, 2010; Sanborn et al., 2010; Lieder et al., 2014; Vul et al., 2014; Sanborn and Chater, 2016; Lieder et al., 2017; Zhu et al., 2020) and variational inference (Hohwy et al., 2008; Daw et al., 2008; Sanborn and Silva, 2013) have been invoked to explain a wide variety of heuristics and biases (reviewed in Sanborn (2015); Griffiths et al. (2012b)). Here, too, it is important to distinguish between probabilistic models of the world that are posited to exist in a subject's mind (as in Bayesian Encoding) from experimenter-defined models of a particular task (as in Bayesian Decoding). Closely related is the distinction drawn by Knill and Richards between the "inference problem" (what the brain infers in the internal model it assumes) and the "information problem" (what information is available in the world) (Knill and Richards, 1996, Chapter 1). For example, Vul et al. argue

that certain deviations from Bayes-optimal behavior can be explained as the result of basing decisions on a single Monte-Carlo sample. However, it is conceivable that what appears to be a single point-estimate sample over a quantity relevant to a task may, in fact, be a local, perhaps unimodal distribution over detailed internal model, as in mean-field approximations. It is further conceivable that multiple "samples" correspond to a mixture of mean-field approximations over an internal model (Jaakkola and Jordan, 1998). Conversely, we described earlier (2.5.4) how a single high dimensional point estimate of an internal model may be sufficient to facilitate apparently Bayesian behavior with respect to a low-dimensional task. Changing our "reference frame" from internal models to experimenter-defined tasks may make samples appear as variational parameters, or vice versa.

The Bayesian Brain Hypothesis is not a single idea, but a collection of computational models, philosophical ideas, and explanations for a variety of empirical data. It is a *framework* rather than a *theory* (Griffiths et al., 2012a). Bayesian Encoding and Bayesian Decoding are complementary approaches to constructing concrete models within the Bayesian Brain framework. These two approaches have been a major source of variation among models, and their complementary nature has previously gone unnoticed. We hope that these insights will lead to clearer and more productive discussions on the nature of inference in the brain, both in terms of neural representations of probability and in terms of behavior.

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.

Charles H Anderson and D C Van Essen. Neurobiological computational systems. *IEEE World Congress on Computational Intelligence*, pages 1–11, 1994.

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.

Anthony J Bell and Terrence J Sejnowski. The "Independent Components" of Scenes are Edge Filters. *Vision Research*, 37(23):3327–3338, 1997.

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.

Donald T. Campbell. Downward causation in Hierarchically organized biological systems. In F J Ayala, editor, *Studies in the philosophy of biology*, chapter 11, pages 179–186. Macmillan Publishers Limited, 1974.

Nathaniel D Daw, Aaron C Courville, and Peter Dayan. Semi-rational models of conditioning. In Nick Chater and Mike Oaksford, editors, *The Probabilistic Mind:: Prospects for Bayesian cognitive science*. Oxford Scholarship Online, 2008.

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

Peter Dayan, Geoffrey E. Hinton, RM Neal, and RS Zemel. The Helmholtz machine. *Neural Computation*, 7(5):1–16, 1995.

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.

Dileep George and Jeff Hawkins. Towards a mathematical theory of cortical micro-circuits. *PLoS computational biology*, 5(10):e1000532, oct 2009.

Dileep George, Alexander Lavin, J. Swaroop Guntupalli, David Mely, Nick Hay, and Miguel Lázaro-Gredilla. Cortical Microcircuits from a Generative Vision Model. In *Cognitive Computational Neuroscience*, 2018.

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.

Agnieszka Grabska-Barwinska, Jeffrey M. Beck, Alexandre Pouget, and Peter E. Latham. Demixing odors — fast inference in olfaction. *Advances in Neural Information Processing Systems*, 26, 2013.

Thomas L. Griffiths, Nick Chater, Dennis Norris, and Alexandre Pouget. How the Bayesians Got Their Beliefs (and What Those Beliefs Actually Are): Comment on Bowers and Davis (2012). *Psychological Bulletin*, 138(3):415–422, 2012a.

Thomas L. Griffiths, Edward Vul, and a. N. Sanborn. Bridging Levels of Analysis for Probabilistic Models of Cognition. *Current Directions in Psychological Science*, 21(4):263–268, 2012b.

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.

Jakob Hohwy, Andreas Roepstorff, and Karl J. Friston. Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, 108(3):687–701, 2008.

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.

Tommi S. Jaakkola and Michael I. Jordan. Improving the Mean Field Approximation via the Use of Mixture Distributions. In Michael I. Jordan, editor, *Learning in Graphical Models*. Kluwer Academic Publishers, 1998.

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

Mehrdad Jazayeri and J. Anthony Movshon. Optimal representation of sensory information by neural populations. *Nature Neuroscience*, 9(5):690–696, 2006.

Daniel Kersten, Pascal Mamassian, and Alan Yuille. Object perception as bayesian inference. *Annual Review of Psychology*, pages 271–304, 2004.

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.

Alexander Lavin, J. Swaroop Guntupalli, Miguel Lázaro-gredilla, Wolfgang Lehrach, and Dileep George. Explaining Visual Cortex Phenomena using Recursive Cortical Network. In *Cognitive Computational Neuroscience*, 2018.

Falk Lieder, Ming Hsu, and Thomas L. Griffiths. The high availability of extreme events serves resource-rational decision-making. In *Cognitive Science Society*, pages 2567–2572, 2014.

Falk Lieder, Thomas L. Griffiths, Quentin J M Huys, and Noah D. Goodman. The anchoring bias reflects rational use of cognitive resources. *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.

R. Moreno-Bote, David C. Knill, and A. Pouget. Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, 108(30):12491–12496, 2011.

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.

Dimitri Probst, Mihai A. Petrovici, Ilja Bytschok, Johannes Bill, Dejan Pecevski, Johannes Schemmel, and Karlheinz Meier. Probabilistic inference in discrete spaces can be implemented into networks of LIF neurons. *Frontiers in computational neuroscience*, 9(13):1–11, 2015.

Dobromir Rahnev. The bayesian brain: What is it and do humans have it? *Behavioral and Brain Sciences*, 42:e238, 2019.

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.

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.

Adam N. Sanborn and Ricardo Silva. Constraining bridges between levels of analysis: A computational justification for locally Bayesian learning. *Journal of Mathematical Psychology*, 57(3-4):94–106, 2013.

Adam N Sanborn, Thomas L Griffiths, and Daniel J Navarro. Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review*, 117(4):1144–67, oct 2010.

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 and Anina N. Rich. Independent Sampling of Features Enables Conscious Perception of Bound Objects. *Psychological Science*, 21(8):1168–1175, 2010.

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.

Stephen Yablo. Mental Causation. *The Philosophical Review*, 101(2):245–280, 1992.

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

Jian-Qiao Zhu, Adam N Sanborn, and Nick Chater. The Bayesian Sampler: Generic Bayesian Inference Causes Incoherence in Human Probability Judgments. *Psychological Review*, 127(5):719–748, 2020.

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:

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

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

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

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

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

.

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

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

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

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

where

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

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, $\mathbf{\Sigma_I}$ shrinks towards zero, and $\boldsymbol{\mu}_{\mathbf{I}}$ goes to $\mathbf{A}\bar{\mathbf{x}}_t$, which implies that $\mathcal{N}(\mathbf{I};\boldsymbol{\mu}_{\mathbf{I}},\mathbf{\Sigma_I})$ 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

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

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

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

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

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

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

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

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).