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 that explains its inputs (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 data. 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 it is possible that the very same neural activity can be described as probabilistic inference by neural sampling in the *Encoding* framework while also forming a linear probabilistic population code (PPC) in the *Decoding* framework. This demonstrates that certain families of Encoding and Decoding models are compatible with each other rather than competing explanations of data. In sum, Bayesian Encoding and Bayesian Decoding are distinct, non-exclusive philosophies, and appreciating their similarities and differences will help organize future work and allow for stronger empirical tests about 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

*equal contribution

cortical activity (Berkes et al., 2011), neural variability (Hoyer and Hyvärinen, 2003; Orbán et al., 2016; Festa et al., 2021), 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., 2020). Meanwhile, parametric codes 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. Notably, contrast-invariant tuning and divisive normalization have also been replicated by sampling models (Orbán et al., 2016; Echeveste et al., 2020).

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 illustrate the complementary nature of these two philosophies using a simple model of 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 derive the implied *Decoding* model. We show 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* represents parameters (in the "Bayesian Decoding" sense), and depending on the encoding models' generative model, and the considered task, this parametric code may even be a linear PPC. We conclude with a discussion of distributional neural codes in general.

2 Bayesian Encoding vs Bayesian Decoding

We follow the seminal work of Zemel et al. (1998) in assuming that patterns of neural activity represent entire probability distributions over a variable, not just a point estimate of it, i.e. that they form a *distributional* code. The nature of this "variable" and its relationship to neural response is key to the distinction between the Bayesian Encoding and the Bayesian Decoding 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.

Bayesian Encoding requires a source for the reference distribution. In the context of the sensory system, this typically takes the form of an internal generative model of sensory inputs, and the distribution to be encoded is the posterior over latent variables in that model (Figure 1a-b). 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

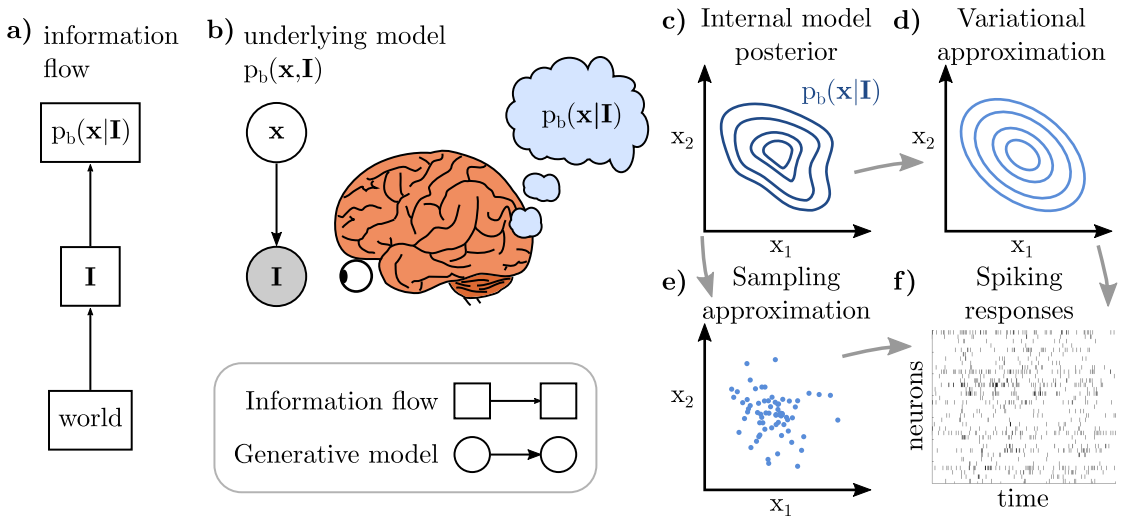

Figure 1: Visualization of Bayesian Encoding. **a)** Diagram of information flow: the world provides sensory inputs ($\mathbf{I}$), which then give rise to inferences about latent variables ($\mathbf{x}$). **b)** Bayesian Encoding typically assumes that the brain has an internal model of its inputs, and that perceptual inferences are about variables in this internal model, not necessarily corresponding to quantities in the external world *per se*. With Bayesian Encoding, it is also typical to assume that the internal model is *task-independent* and that the brain always computes a posterior over internal variables, $p_b(\mathbf{x}|\mathbf{I})$, regardless of whether $\mathbf{I}$ is a highly controlled stimulus in a task or encountered in the wild. **c-f)** The defining feature of Bayesian Encoding is the existence of a reference distribution (c), typically 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 (c→d) or sampling (c→e), and that this approximation is then realized in patterns of neural activity (f).

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 a particular configuration of the causes for generating a particular sensory observation. The encoded posterior distribution in this framework is defined over the latent 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 $\mathrm{p_b}(\cdot)$ 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. A prototypical case of Bayesian Encoding 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 1c), 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 1d-f), 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 1e), as well as approaches inspired by variational inference techniques, related to parametric neural codes (Figure 1d) (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 $\mathrm{p}(s|\mathbf{r})$ (Figure 2). The key question within the Bayesian Decoding framework is this: what conditions must the stimulus-driven neural activity ($\mathrm{p}(\mathbf{r}|s)$) 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).

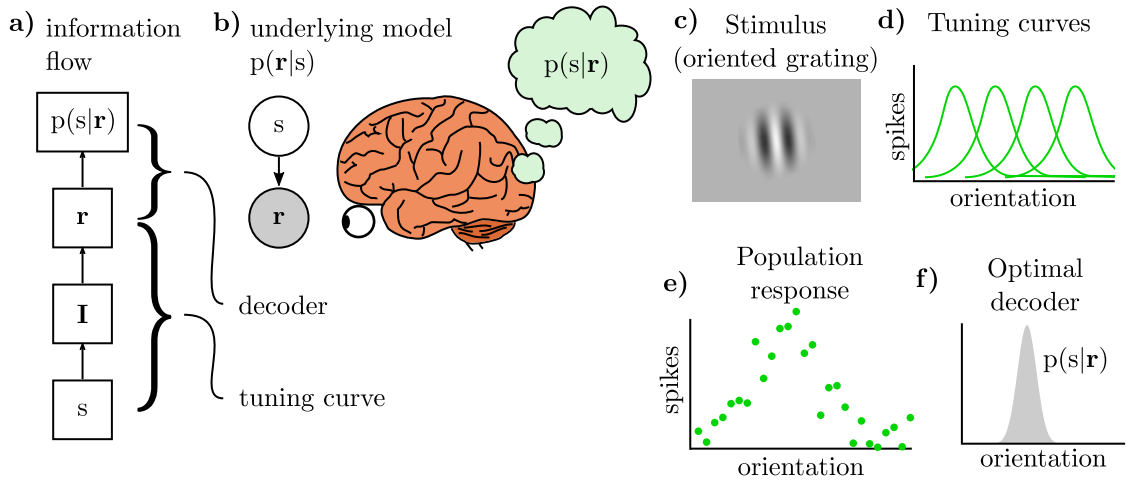

Figure 2: Visualization of Bayesian Decoding. **a)** Diagram of information flow: a quantity of interest (s) in the world elicits neural responses ($\mathbf{r}$), mediated by sensory inputs ($\mathbf{I}$). The decoding question is how the brain forms an internal estimate, $\hat{s}$, from $\mathbf{r}$. **b)** The underlying probabilistic model assumes that s generates $\mathbf{r}$, so inference is the problem of recovering s from $\mathbf{r}$. **c)** The decoding problem usually begins with a stimulus, such as the orientation of a grating of a given spatial frequency, size, location, and contrast. **d-f)** Given a population of neurons' tuning curves to s (d) and an observation of spikes on a single trial (e), an optimal decoder computes $\mathrm{p}(s|\mathbf{r})$ (f).

2.3 Contrasting Bayesian Encoding and Bayesian Decoding

There are four key differences between the Bayesian Encoding and Bayesian Decoding perspectives, which we will discuss in each of the following sections: (1) what they assume the brain is inferring, (2) what the terms "likelihood" and "posterior" refer to, (3) the role of neural responses in the theory, and (4) 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, especially in the case 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; Bányai et al., 2019). 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, 2022).

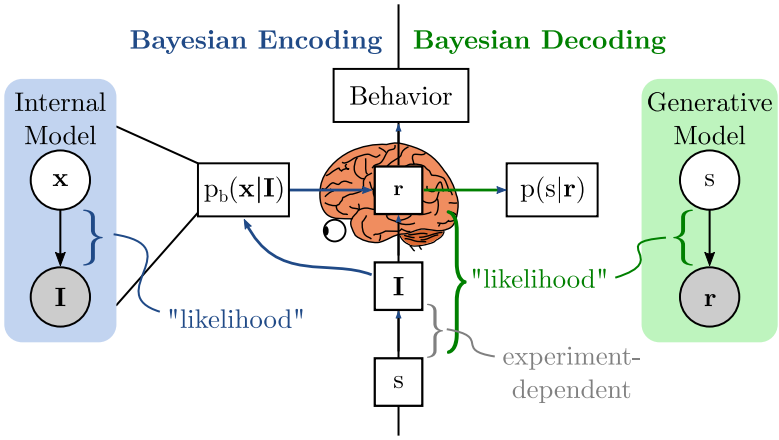

Figure 3: Side-by-side comparison of Bayesian Encoding and Bayesian Decoding. In both frameworks, it is understood that there exists a mechanistic, biophysical connection between stimuli ($\mathbf{I}$), sensory neural responses ($\mathbf{r}$), and behavior. In the Bayesian Decoding framework, emphasis 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; Lange and Haefner, 2022). 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 means that the "likelihood" in a Bayesian Decoding model depends on choices made by the experimenter (such as their choice of stimuli), but not in a Bayesian Encoding model.

In contrast, Bayesian Decoding models are typically applied in the context of estimating 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, nor does it require a true or correct reference distribution over 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, $\mathrm{p}(\mathbf{r}|s)$, making it observable experimentally. The difference between (typical realizations of) the Bayesian Encoding and Decoding perspectives is illustrated in Figure 3.

We emphasize that Bayesian Encoding *typically* but not *necessarily* involves a task-independent internal generative model, and Bayesian Decoding likewise is *typically* but not *necessarily* applied to task-specific variables. In fact, where the ideas of encoding and decoding distributions first appeared in Zemel et al (1998), the encoded distribution was over task quantities (such as $\mathbf{x}$ being motion or heading direction), without specifying an internal generative model, and the decoding problem was framed as the inverse to the encoding problem – that is, recovering the encoded $\mathrm{p}(\mathbf{x}|\ldots)$ from $\mathbf{r}$. This again emphasizes the complementary nature of these philosophies: we are free to apply the Bayesian Decoding framework to variables in a task-independent internal model (given $\mathbf{r}$, what do we know about $\mathbf{x}$ or $\mathrm{p_b}(\mathbf{x}|\mathbf{I})$?), or to apply the logic of Bayesian Encoding to task-specific quantities (construct $\mathbf{r}$ to encode a desired $\mathrm{p}(s|\mathbf{I})$), but such examples are rare. In the remainder of this paper, we will associate Bayesian Encoding with task-independent internal generative models and Bayesian Decoding with variables in a task, and in the Discussion we will return to the possibility that task variables are explicitly represented as part of the brain's internal model.

2.3.2 Differing notions of likelihood

Another difference in philosophy is evidenced by divergent usage of the term "likelihood" (Figure 3). In the typical Bayesian Encoding setting, 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 variables in a generative model of sensory data*. In the typical Bayesian Decoding setting, 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 Differences in the relationship between distributions and neural activity

Bayesian Encoding models require two distinct assumptions: first, what is the source of the reference distribution to be encoded (e.g. what is the brain's internal model $\mathrm{p_b}(\mathbf{x}, \mathbf{I})$); and second, what is the linking hypothesis that maps probability distributions to neural activity (Figure 1)? This approach of starting with the encoded distribution abstracts away from the details of neural circuits that must actually *implement* inference. Take the model of Orbán et al. (2016) for example. In this work, the authors assume that neurons in primary visual cortex implement a sampling algorithm to encode the posterior distribution over latent variables in a Gaussian Scale Mixture model. This specifies the reference distribution. It is then assumed that, by some *unspecified* mechanism, the trajectory of a set of neurons' membrane potentials over time traces out real-valued samples from the posterior, and that these membrane potentials elicit spikes through a nonlinear accumulation process. This specifies the linking hypothesis, or the map from the reference distribution to neural data. This model successfully reproduced a diverse set of known properties about V1 (Orbán et al., 2016), but it is not a mechanistic model. From a modeling standpoint, the way that an input image elicits neural activity is *mediated* by the reference posterior: an example of "downward causation" (Campbell, 1974; Yablo, 1992).

While for Encoding models there is a clear separation of computational model and neural link, they still of course beg the question of how inference in the computational model is implemented in neural circuits. Prior work has investigated the question of how biologically-plausible recurrent circuits could implement sampling (Bill et al., 2015; Probst et al., 2015; Aitchison and Lengyel, 2016; Petrovici et al., 2016; Dold et al., 2019; Echeveste et al., 2020) or message-passing (George and Hawkins, 2009; Beck et al., 2012; Raju and Pitkow, 2016; Grabska-Barwinska et al., 2013; Grabska-Barwińska et al., 2017; George et al., 2018) through their

dynamics. However, in these examples there is typically a cost to increased biological plausibility, either by degrading the quality of the encoded distribution, or by degrading the match to empirical neural data.

Bayesian Decoding models, in contrast, do not distinguish between the uncertainty in an underlying probabilistic model and the uncertainty of a downstream brain area applying a Bayesian decoder to some neural activity. As a result, Decoding models replace the assumption about the link to neural activity with a *constraint* on the relationship between stimuli (s) and neural activity ($\mathbf{r}$).

To illustrate this point, let us revisit one of the motivating examples for distributional codes of Zemel et al. and contrast the Encoding and Decoding approaches. Consider a rat who is placed into a water maze and must navigate to a hidden platform (Morris, 1984). Initially, the rat may be uncertain about which direction it is facing, e.g. if opposite walls of the maze look the same its correct belief about direction will be bimodal. Similar to the orientation of a grating, head direction is a scalar variable in $[0, 2\pi]$ that we will call s. In the Encoding approach, one might begin by asking what is $\mathrm{p_b}(s|\mathbf{I})$ according to an internal model of the environment, where $\mathbf{I}$ stands for the sensory cues the rat uses to orient itself. The distribution $\mathrm{p_b}(s|\mathbf{I})$ determines how uncertain the rat *ought* to be, according to the internal model. Continuing the Encoding approach, one would then adopt a linking hypothesis (sampling, parametric, etc.) whereby $\mathrm{p_b}(s|\mathbf{I})$ is encoded in neural activity $\mathbf{r}$. In an Encoding model, the encoding of a distribution may be imperfect and lossy, or it may contain more information about the distribution than is being used by downstream circuits. In either case, the way a downstream circuit *uses* the neural activity will generally differ from a Bayesian decoder.

Applying the Bayesian Decoding framework to the same problem, we would say that the uncertainty in $\mathrm{p}(s|\mathbf{r})$ is the primary kind of uncertainty we should be concerned with, and that there is no distinction between this and the rat's internal model. Crucially, this does not trivialize representations of uncertainty as "just" a matter of optimal decoding. In the Decoding approach, there may still be an ideal uncertainty that the rat *ought* to have when it is first placed into the maze; however the assumption is that this uncertainty is realized through the way $\mathbf{r}$ is tuned to its inputs $\mathbf{I}$. That is, it is left to the brain (evolution, learning) to have carefully constructed tuning functions $\mathrm{p}(\mathbf{r}|\mathbf{I})$, such that $\mathrm{p}(s|\mathbf{r})$ is equal to $\mathrm{p_b}(s|\mathbf{I})$ (Ma et al., 2006). One way that the Encoding and Decoding perspectives can become identical, then, is when the decoded distribution $\mathrm{p}(s|\mathbf{r})$ equals the reference distribution $\mathrm{p_b}(s|\mathbf{I})$. From the Encoding point of view, this requires that the encoding of $\mathrm{p_b}(s|\mathbf{I})$ into $\mathbf{r}$ is lossless (or "efficient" in the terminology of Beck et al. (2012)). From the Decoding point of view, they are identical by assumption.

Finally, the preceding discussion points to an important practical difference between Encoding and Decoding philosophies in terms of how neural responses are interpreted by downstream areas. In a Decoding model, a downstream area implicitly applies Bayes' rule to the neural responses arriving from an upstream area to extract information about a stimulus. 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. 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; Angelaki et al., 2009; 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 Decoding is further motivated by experimental data showing

	Bayesian Encoding	**Bayesian Decoding**
Sampling-based representation	Hoyer and Hyvärinen (2003) Pecevski et al. (2011) Berkes et al. (2011) Buesing et al. (2011) Gershman et al. (2012) Savin and Denève (2014) Probst et al. (2015) Orbán et al. (2016) Haefner et al. (2016) Aitchison and Lengyel (2016) Festa et al. (2021)	Moreno-Bote et al. (2011)†
Parametric representation	Zemel et al. (1998) Sahani and Dayan (2003) Friston (2005) George and Hawkins (2009) Beck et al. (2012) Raju and Pitkow (2016) Vertes and Sahani (2018) Tajima et al. (2016)*	Ma et al. (2006) Beck et al. (2008) Beck et al. (2011) Hou et al. (2019) 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. Tajima et al., marked with "*" contains elements of both encoding and decoding. Moreno-Bote et al., marked with "†", contains elements of both sampling-based and parametric decoding.

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 typical Bayesian Encoding theories (Rahnev, 2019; Koblinger et al., 2021). 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 early visual cortex, as evidenced by changes in neural responses in primary visual cortex over development (Berkes et al., 2011) and task-learning (Haefner et al., 2016; Lange and Haefner, 2022). 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., 2020). 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). 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; Koblinger et al., 2021).

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 to approximate the posterior over variables in a generative model, including Probabilistic Population Codes (PPCs) (Beck et al., 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 Case Study: primary visual cortex (V1)

We now focus on primary visual cortex (V1) to provide a concrete example illustrating and further elaborating on our general points above. Focusing on area V1 has the advantage that much neurophysiological data exists, and both encoding and decoding approaches have enjoyed some success. We will first briefly describe existing work from both perspectives, and then use a simple example to show how they can lead to very different conclusions about the neural code. To that end, we will assume a Bayesian Encoding model that encodes the posterior over internal variables by sampling and show analytically how to derive the corresponding Bayesian Decoding model, obtaining a parametric representation (PPC) (Shivkumar et al., 2018).

2.5.1 Bayesian Encoding models for V1

The starting point for the Bayesian Encoding approach, applied to V1, is an assumption about the brain's generative model $\mathrm{p_b}(\mathbf{x}, \mathbf{I})$. That is, we must specify what is $\mathbf{x}$, the variable assumed to be inferred and represented by V1 neurons, and how $\mathbf{x}$ is related to the sensory observations, $\mathbf{I}$. For simple cells in area V1, Olshausen and Field proposed a linear Gaussian likelihood $\mathbf{I} \sim \mathcal{N}(\mathbf{A}\mathbf{x}, \mathbf{\Sigma_I})$ with a sparse independent prior $\mathrm{p_b}(\mathbf{x})$ as the brain's internal model Olshausen and Field (1996, 1997). (We use the notation $\mathbf{I} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ to indicate a random variable drawn from a multivariate normal distribution, and $\mathcal{N}(\mathbf{I}; \boldsymbol{\mu}, \mathbf{\Sigma})$ to denote its density function). In this model, the observed retinal image, $\mathbf{I}$, is assumed to be a linear combination of "projective fields" ($\mathbf{PF}_i$) plus unexplained pixel noise $\mathbf{\Sigma_I}$; 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, $\mathbf{x} = (x_1, \ldots, x_n)^\top$. Intuitively, in this model, V1 activity is assumed to somehow represent beliefs about what values for $\mathbf{x}$ best explain a given retinal image, $\mathbf{I}$.

The next assumption in the Encoding framework is about the neural code, or how the posterior distribution, $\mathrm{p_b}(\mathbf{x}|\mathbf{I})$, is represented by neural responses, $\mathbf{r}$. Continuing the previous example, Olshausen and Field assumed that each x_i was represented by a single neuron whose firing rate was proportional to the most probable value for x_i given an image (maximum a posteriori, MAP): $r_i \propto \mathrm{argmax}_{x_i}\, \mathrm{p_b}(\mathbf{x}|\mathbf{I})$. In this model, a single neuron represents the most likely intensity with which a visual feature is present in the image. This is not a fully Bayesian Encoding model in the sense that only the MAP, but not the full posterior distribution $\mathrm{p_b}(\mathbf{x}|\mathbf{I})$ is encoded in neural responses. Empirical support for this model is based on the observation that learning (fitting) this model on natural images yields visual features ($\mathbf{PF}_i$) that are localized, oriented, and band-pass filtered, implying neural responses and receptive fields with similar properties – just as observed empirically (Olshausen and Field, 1996).

Subsequent work has both modified and extended this generative model, and combined it with different neural codes. Hoyer and Hyvärinen proposed that neural responses can be understood as samples from the posterior in the same generative model to qualitatively explain the variability and mean-variance relationship of neural responses. Schwartz and Simoncelli extended the generative model to a Gaussian scale mixture model to explain the empirically observed contrast normalization of V1 responses, and Orbán et al. (2016) found agreement between the predictions of a Gaussian scale mixture model combined with neural sampling and a wide range of observations related to the stimulus-dependence of neural variability. Bornschein et al. (2013) proposed a variation of the generative model of Olshausen and Field using a nonlinear Gaussian likelihood and/or binary as opposed to continuous latents $\mathbf{x}$, and Coen-Cagli et al. (2015) found that a

further extension to the generative model in the form of a Mixture of Gaussian Scale Mixture model could explain center-surround interactions in V1. Finally, Haefner et al. (2016) combined the generative model of Olshausen and Field with the ideal observer model of a discrimination task to explain choice probabilities and task-dependent noise correlations of V1 neurons.

The key shared element of all these models is an explicit assumption about the computational variable $\mathbf{x}$ that is being represented, and how this variable is related to the sensory observations $\mathbf{I}$. This model being adapted to natural inputs is an important constraint, and the model parameters are usually obtained by fitting the model to sets of natural images. These models are then general purpose and can be queried using natural inputs or images presented in a task. The fundamental framing is how V1 neurons *encode* the posterior over $\mathbf{x}$ given an arbitrary input, $\mathbf{I}$.

2.5.2 Bayesian Decoding models for V1

The starting point for the Bayesian Decoding approach, applied to V1, is a measurement of the conditional probability (or likelihood of s), $\mathrm{p}(\mathbf{r}|s)$, for some stimulus s to which V1 neurons are tuned, and that is hypothesized to be represented, such as orientation. Importantly, this means that the measured likelihood is to some extent under experimental control, since the experimenter chooses what images correspond to each value of s (e.g. the size, contrast, or spatial frequency of a grating). In general, for an arbitrary s, this likelihood will be very complicated reflecting the fact that s cannot easily be decoded from $\mathbf{r}$ (e.g. object identity from V1). However, for V1 responses it has empirically been found that the optimal Bayesian decoder for orientation is approximately linear in spike counts and invariant to contrast, a classic nuisance variable (Graf et al., 2011). This finding has been interpreted as meaning that V1 activity "represents" orientation. In conjunction with the Poisson-like neural response variability in V1, this implies that the beliefs of a Bayesian decoder of orientation applied to the neural responses are part of the exponential family. Furthermore, the sufficient statistics are linear in the neural responses. Such a neural representation of a belief has been called a Probabilistic Population Code (PPCs) as introduced by Ma et al..

The same logic applies to other candidates for s that modulate the responses of V1 neurons in a straightforward manner, such as spatial frequency or location. The key element of the Bayesian Decoding approach is taking the perspective of downstream circuits trying to extract information about s from V1 activity: how is the information about s formatted in V1 activity, and is $\mathrm{p}(s|\mathbf{r})$ "simple"? In contrast to the Bayesian Encoding perspective which justifies its choice of $\mathbf{x}$ by its fit to natural images and its ability to predict neural responses, the Bayesian Decoding perspective justifies its choice of s by desirable properties of an efficient decoder, e.g. linearity and invariance to nuisance variables.

2.5.3 Example model where decoding a stimulus s from encoded samples results in a PPC

Since our main points are conceptual in nature, we will develop the link between the Encoding and the Decoding approach for the simple case of a linear Gaussian model with a Gaussian prior, under the assumption of a sampling-based neural code. These simplifying assumptions make the difference between Encoding and Decoding clear and analytically tractable, but are not meant to maximize biological plausibility. For instance, the posterior variance in this model is independent of $\mathbf{I}$, whereas it is well-known that neural response variance is stimulus-dependent, and this effect is captured by neural sampling models with less trivial generative models (Orbán et al., 2016; Bányai et al., 2019; Festa et al., 2019). Beginning with a more complicated Encoding model would lead to a more complicated relationship to Decoding models (e.g. where the Decoder is more complex, e.g. a nonlinear PPC, or not even in the exponential family). Importantly, the core of our argument remains: that an Encoding model based on one type of neural code (e.g. sampling) and a Decoding model based on another type (e.g. parametric) need not be in contradiction with each other, and offer complementary perspectives on the same system.

Given an image, $\mathbf{I}$, we assume that V1 neurons *encode* the posterior $\mathrm{p_b}(\mathbf{x}|\mathbf{I})$ by sampling t values from 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 $\mathbf{x}$ (Hoyer and Hyvärinen, 2003). We assume that responses from a population of n neurons correspond to samples from the posterior over $\mathbf{x}$, so that at each instant, the population response, $\mathbf{r}^{(t)}$, equals the sample $\mathbf{x}^{(t)}$. Each sample of x_i (or r_i) represents the brain's instantaneous belief about the intensity of the feature $\mathbf{PF}_i$ in the image.

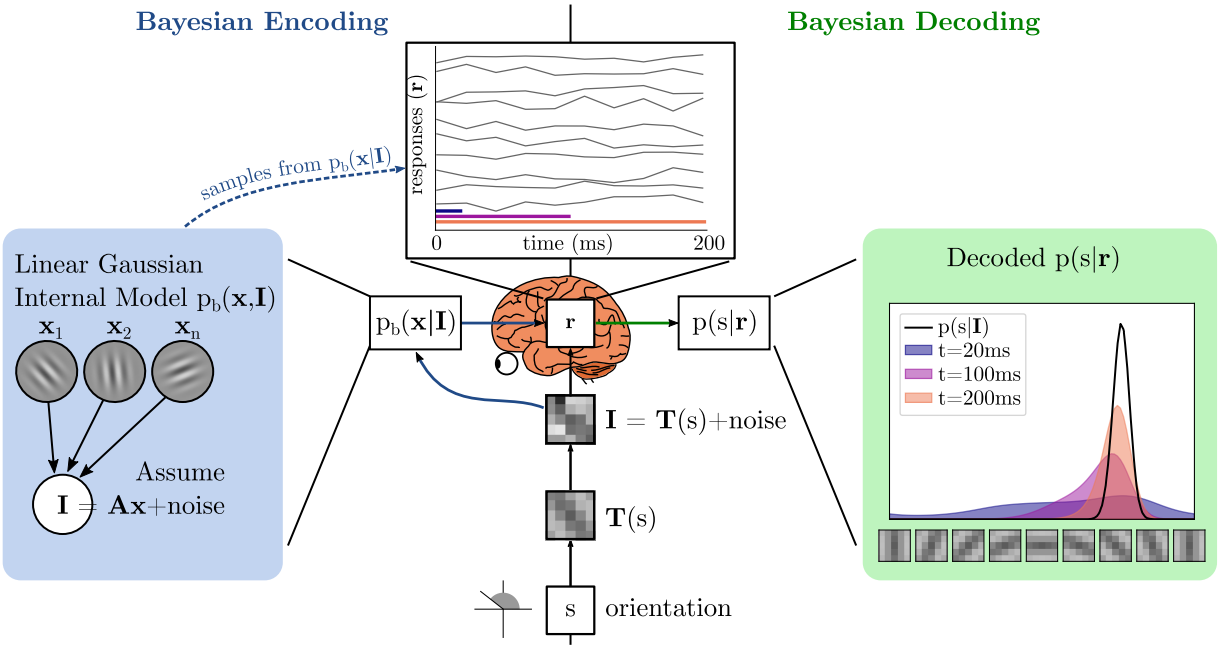

Figure 4: Encoding by sampling followed by decoding of orientation from the samples in a simplified model. As in Figure 3, Encoding elements are on the left, and Decoding elements are on the right. In our example model, the brain performs sampling-based inference over $\mathbf{x}$ in a probabilistic model of images, here a Linear Gaussian model. In a given experiment, the image is generated according to an experimenter-defined process that turns a scalar stimulus s, e.g. orientation, into an image observed by the brain. To simplify, neural responses $\mathbf{r}$ are assumed to reflect instantaneous real-valued samples of $\mathbf{x}$ drawn from the posterior $p_b(\mathbf{x}|\mathbf{I})$. In our simulation we drew 10 samples and assumed 20ms per sample. The samples drawn from the model are then probabilistically "decoded" to a probability distribution over s. This distribution sharpens as more samples are observed. The optimal decoder for any t is a linear PPC.

We will now apply the Bayesian Decoding approach to the sequence of samples produced by the sampling-based Encoding model described above. An ideal observer applies Bayes' rule to infer $p(s|\mathbf{r}^{(1)},\ldots,\mathbf{r}^{(t)})$ using knowledge of the probabilistic relationship between samples ($\mathbf{x}$ or $\mathbf{r}$) and s:

$$\begin{aligned} p(s|\mathbf{r}^{(1)},\ldots,\mathbf{r}^{(t)}) &\propto p(s)\, p_b(\mathbf{r}^{(1)},\ldots,\mathbf{r}^{(t)}|s) \\ &\propto p(s) \int p(\mathbf{I}|s)\, p_b(\mathbf{r}^{(1)},\ldots,\mathbf{r}^{(t)}|\mathbf{I})\, d\mathbf{I}\,. \end{aligned} \tag{2}$$

That is, the optimal decoder combines knowledge of (i) how likely an image $\mathbf{I}$ is to generate a set of samples of $\mathbf{x}$ (or $\mathbf{r}$), and (ii) how likely a stimulus value s is to generate an image $\mathbf{I}$. In general, this decoded distribution over s may be arbitrarily complex and intractable. One factor that is under experimental control is the"template" function $\mathbf{T}(s)$ which renders an image, such as a grating with orientation s. This provides the link between s and $\mathbf{I}$ in equation (2). In our model, we assume that the input the brain receives is a noisy version of that template (Figure 4).

The first simplification to the general form of the optimal decoder in (2) we can derive, under the assumption of a Gaussian likelihood, is to show that the posterior over s depends only on the mean rate of $\mathbf{r}$ (i.e. a rate code rather than temporal code):

$$p(s|\mathbf{r}^{(1)},\ldots,\mathbf{r}^{(t)}) = p(s|\bar{\mathbf{r}}) \tag{3}$$

where $\bar{\mathbf{r}} = \frac{1}{t}\sum_{i=1}^{t} \mathbf{r}^{(t)}$ is the mean response after t samples (Supplemental Text). Any decoder that obeys (3) can be seen as a kind of *parametric code* over s, where the rates $\bar{\mathbf{r}}$ are the parameters. A second convenient

property for a decoder to have is if the optimal decoder is in an exponential family, or

$$\mathrm{p}(s|\bar{\mathbf{r}}) \propto g(s) \exp\left(\mathbf{h}(s)^\top \mathbf{f}(\bar{\mathbf{r}})\right) . \tag{4}$$

Whenever (4) is true, then we would say that the neural activity forms a particular kind of parametric code called a *Nonlinear PPC* over s. A final convenient property for a decoder to have is if $\mathbf{f}(\bar{\mathbf{r}})$ is linear:

$$\mathrm{p}(s|\bar{\mathbf{r}}) \propto g(s) \exp\left(\mathbf{h}(s)^\top \bar{\mathbf{r}}\right) . \tag{5}$$

This is the definition of a *Linear PPC* over s (Ma et al., 2006)[1].

In our simplified Encoding model, we can analytically derive the optimal decoder of the experimenter-defined s conditioned on neural responses, where those neural responses are generated by sampling from the Linear Gaussian internal model described above (derivation in the Supplemental Text). We find that the optimal decoder is, in fact, a Linear PPC over s as defined in (5)![2] This sequence of steps from equation (2) through (5) suggests a general way to derive the Bayesian Decoding model implied by a given Bayesian Encoding model.

As we discussed earlier, Encoding models typically (but not necessarily) consider inference in a task-independent internal model, while Decoding models typically (but not necessarily) consider inference of low-dimensional task-specific quantities. The model we described in this section is "typical" in this sense: inference of $\mathbf{x}$ is constructed to be task-independent, while the decoder of s given $\mathbf{r}$ depends inextricably on the "template function" $\mathbf{T}(s)$, which is under the control of the experimenter. The kernels $\mathbf{h}(s)$ will be different for gratings of different size and spatial frequency, for plaids, or for different objects. This example shows how a Bayesian Decoding model for s, implied by a task-independent Bayesian Encoding model, can nonetheless be *experiment-dependent*. This points to a potentially empirically decidable question: do downstream areas such as V2 interpret V1 activity like a Bayesian Decoder, or do they interpret V1 activity as representing a posterior over a set of latents, $\mathbf{x}$?

3 Discussion

We have identified and characterized a previously unstated difference between 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 provides 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 but compatible models of the same neural circuit, underlining the point that Bayesian Encoding and Decoding 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 example model sheds light on 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, demonstrating that it is possible that the very same neural responses are compatible with both depending on perspective.

Our model is a constructive proof that Encoding and Decoding models *can* be compatible on the same data, but this is will not be true in general. For instance, non-Gaussian Encoding models will not generally form a linear PPC from the decoding perspective, or only for specific sets of stimuli, or they may be decodable only as a nonlinear PPC. Generalizing from our specific example, the key question is, which families of Encoding models, consisting of both $\mathrm{p_b}(\mathbf{I}, \mathbf{x})$ and an assumption about the link to neural responses, are compatible with which families of Decoding models, consisting of $\mathrm{p}(s|\mathbf{r})$ and $\mathrm{p}(\mathbf{I}|s)$? These will come in

[1] PPCs also place restrictions on nuisance variables which we have omitted here.

[2] Further discussion of the nature of this PPC and its relation to the parameters of the internal model can be found in Shivkumar et al. (2018).

pairs – each family of Encoding models defines a family of compatible Decoding models, and vice versa. Identifying these pairs of compatible model families is a theoretical question with important implications for the interpretation of empirical data: while Encoding and Decoding models traditionally have been supported (and falsified) by different kinds of empirical data (see section 2.3.4), understanding their link will allow us to bridge that divide. For instance, if an Encoding model implies a particular family of Decoding models, then data that falsifies that Decoding family will also falsify the Encoding family. Similarly, if a Decoding model is only compatible with a family of Encoding models that is too constrained to effectively model natural inputs, then that would pose a challenge for the Decoding model. As an example of this kind of argument, Orbán et al. note in supplemental analyses that their sampling model appears to be empirically consistent with a contrast-invariant linear PPC over orientation, but that "linear decoding of population responses will significantly fall short of being optimal" once more complex tasks are considered (Orbán et al., 2016).

More generally, our arguments also raise 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 evidence 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). This would be a poor Bayesian Encoding model in the sense that the full $\mathrm{p_b}(\mathbf{x}|\mathbf{I})$ distribution is not recoverable from $\mathbf{r}$. Yet, 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. Another example of non-Bayesian encoding but Bayesian decoding is given by Orhan and Ma. 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 (Rahnev, 2019; Koblinger et al., 2021). 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 seminal paper by Zemel et al. (1998) introduced the concept of encoding (and decoding) general probability distributions in (and from) neural activity. Most work over the following 20+ years typically focused on either Encoding or Decoding, as shown by Table 1, despite Zemel et al. considering both perspectives as tightly linked. This divergence was likely strengthened by the fact that Encoding studies almost exclusively considered internal latent variables ($\mathbf{x}$), while work taking the Decoding perspective considered distributions over task-defined variables (s). From today's perspective, the encoding formalism of Zemel et al. and its application in Sahani and Dayan (2003) maps naturally onto our Bayesian Encoding category. Furthermore, it is philosophically closely aligned with the other studies in this category, and shares with them the idea that implied decoders that are non-Bayesian (but note that "decoding is only an implicit operation that the system need never actually perform" Zemel et al. (1998)). Interestingly, while Zemel et al. (1998) discounted the possibility of optimally decoding the encoded distribution using Bayes' rule as too inflexible, almost all later studies that took the decoding approach were based on Bayes' rule, and now form our Bayesian Decoding category.

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 of the brain's internal model? One possibility is that "orientation" (or any other s in a task) 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 as the result of learning the present task (Haefner et al., 2016; Lange and Haefner, 2022). Echoing section 2.3.3 above, 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 possible other internal variables $\mathbf{x}$) and ask how the true posterior $\mathrm{p_b}(s, \mathbf{x}|\mathbf{I})$ is represented by neural responses $\mathbf{r}$. 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, $\mathrm{p}(s|\mathbf{r})$, 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).

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 issues raised in this paper for models of visual perception also have implications for Bayesian models of cognition, where ideas related to 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), variational inference (Hohwy et al., 2008; Daw et al., 2008; Sanborn and Silva, 2013), or both (Lange et al., 2021) 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 is typical in Bayesian Encoding) from experimenter-defined models of a particular task (as is typical 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 a detailed internal model, as in variational approximations. It is further conceivable that multiple "samples" correspond to a mixture of variational approximations over an internal model (Jaakkola and Jordan, 1998; Lange et al., 2022). Conversely, 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.

Koblinger et al. (2021) recently posed the question whether uncertainty in the brain is represented "constitutively", i.e. about many variables regardless of their relevance for a specific task, or "opportunistically," only about task-relevant variables. While this distinction appears related to the difference between Bayesian Encoding and Decoding with respect to their task-independence, there are important differences. While Encoding and Decoding models have so far mostly been applied in task-independent and task-dependent contexts, respectively, to what degree representations of uncertainty are task-specific is an empirical question that can be productively asked within both the Encoding and Decoding approaches. For instance, Bayesian Encoding models of object recognition may differ in whether they propose that the brain represents posteriors only over task-relevant object identities, or all possible objects. One can similarly imagine both "constitutive" and "opportunistic" Decoding models. For instance, the toy example model we presented above is an opportunistic Decoding model, where s is determined by the experimenter, and $\mathbf{r}$ is only said to represent a distribution over s in that task context. In a constitutive Decoding model, the representation of a distribution about one quantity, like $\mathrm{p}(\text{orientation}|\mathbf{r})$, would potentially interact with representations of other quantities, like $\mathrm{p}(\text{location}|\mathbf{r})$, regardless of the immediate task-relevance of each.

Walker et al. (2022) pointed out a distinction between "descriptive" versus "process" approaches to the study of neural representations of uncertainty. In their classification, the "descriptive" approach derives estimates about the observer's subjective uncertainty from either presented stimuli or recorded behavioral reports. The "process" approach, on the other hand, derives an estimate of uncertainty from neural responses. To what degree this classification is related to the Bayesian Encoding and Decoding approaches, respectively, is unclear, and likely depends on additional assumptions, e.g. about the relationship between behavioral reports and reference posterior in the Encoding approach, $\mathrm{p_b}(\mathbf{x}|\mathbf{I})$, and about the nature of the model used to infer uncertainty from neural responses.

To conclude, 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.

Code availability

Two panels in Figure 4 were generated by simulation. The code is available at https://github.com/haefnerlab/bayesian-encoding-decoding.

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.

Dora E. Angelaki, Yong Gu, and Gregory C. DeAngelis. Multisensory integration: psychophysics, neurophysiology, and computation. *Current opinion in neurobiology*, 19(4):452–8, aug 2009.

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.

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.

Johannes Bill, Lars Buesing, Stefan Habenschuss, Bernhard Nessler, Wolfgang Maass, and Robert Legenstein. Distributed Bayesian computation and self-organized learning in sheets of spiking neurons with local lateral inhibition. *PLoS ONE*, 10(8):1–51, 2015.

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.

Ruben Coen-Cagli, Adam Kohn, and Odelia Schwartz. Flexible gating of contextual influences in natural vision. *Nature Neuroscience*, 18(11):1648–1655, 2015.

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.

Dominik Dold, Ilja Bytschok, Akos F. Kungl, Andreas Baumbach, Oliver Breitwieser, Walter Senn, Johannes Schemmel, Karlheinz Meier, and Mihai A. Petrovici. Stochasticity from function — Why the Bayesian brain may need no noise. *Neural Networks*, 119:200–213, 2019.

Rodrigo Echeveste, Laurence Aitchison, Guillaume Hennequin, and Máté Lengyel. Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *Nature Neuroscience*, 23:1138–1149, 2020.

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

Dylan Festa, Amir Aschner, Adam Kohn, and Ruben Coen-Cagli. A Functional Model of Neuronal Response Variability in Primary Visual Cortex. In *Cognitive Computational Neuroscience*, 2019.

Dylan Festa, Amir Aschner, Aida Davila, Adam Kohn, and Ruben Coen-Cagli. Neuronal variability reflects probabilistic inference tuned to natural image statistics. *Nature Communications*, 12(1):1–11, 2021.

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.

Agnieszka Grabska-Barwińska, Simon Barthelmé, Jeff Beck, Zachary F. Mainen, Alexandre Pouget, and Peter E. Latham. A probabilistic approach to demixing odors. *Nature Neuroscience*, 20(1):98–106, 2017.

Arnulf B.A. Graf, Adam Kohn, Mehrdad Jazayeri, and J. Anthony Movshon. Decoding the activity of neuronal populations in macaque primary visual cortex. *Nature Neuroscience*, 14(2):239–247, 2011.

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.

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.

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.

Ádám Koblinger, József Fiser, and Máté Lengyel. Representations of uncertainty: where art thou? *Current Opinion in Behavioral Sciences*, 38:150–162, 2021.

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. *PLOS Computational Biology*, 18(3), 2022.

Richard D. Lange, Ankani Chattoraj, Jeffrey M. Beck, Jacob L. Yates, and Ralf M. Haefner. A confirmation bias in perceptual decisionmaking due to hierarchical approximate inference. *PLoS Computational Biology*, 17(11):1–30, 2021.

Richard D. Lange, Ari S. Benjamin, Ralf M. Haefner*, and Xaq Pitkow*. Interpolating between sampling and variational inference with infinite stochastic mixtures. *UAI*, August 2022.

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.

Richard Morris. Developments of a water-maze procedure for studying spatial learning in the rat. *Journal of Neuroscience Methods*, 11(1):47–60, 1984.

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. Emin Orhan and Wei Ji Ma. Efficient probabilistic inference in generic neural networks trained with non-probabilistic feedback. *Nature Communications*, 8(138):1–30, 2017.

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.

Mihai A. Petrovici, Johannes Bill, Ilja Bytschok, Johannes Schemmel, and Karlheinz Meier. Stochastic inference with spiking neurons in the high-conductance state. *Physical Review E*, 94, 2016.

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.

Edgar Y Walker, Stephan Pohl, Rachel N Denison, David L Barack, Jennifer Lee, Ned Block, Wei Ji Ma, and Florent Meyniel. Studying the neural representations of uncertainty. *arXiv*, pages 1–28, 2022.

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 decoded posterior for Gaussian prior over x

In this section we provide a brief derivation of the optimal posterior over an experimenter-defined s conditioned on samples of internal-model variables $\mathbf{x}$, where the brain's internal model $\mathrm{p_b}(\mathbf{x}, \mathbf{I})$ is assumed to be a linear Gaussian model with a Gaussian prior over $\mathbf{x}$. The use of a Gaussian prior over $\mathbf{x}$ is a further simplification of the derivation in Shivkumar et al. (2018). Formally, the setup is as follows:

1. Assume that the scalar s (such as orientation) gives rise to observed images $\mathbf{I}$ as

$$\mathbf{I} = \mathbf{T}(s) + \boldsymbol{\eta}\,,$$

   where $\mathbf{T}(s)$ is a "template" function (such as a grating image), and $\boldsymbol{\eta}$ is zero-mean Gaussian-distributed pixel noise with covariance $\boldsymbol{\Sigma}_{\mathrm{e-b}}$.

2. Assume that the brain's internal model, $\mathrm{p_b}(\mathbf{x}, \mathbf{I})$, factorizes as $\mathrm{p_b}(\mathbf{x})\mathrm{p_b}(\mathbf{I}|\mathbf{x})$, where the prior is Gaussian,

$$\mathrm{p_b}(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{\mathrm{p}}, \boldsymbol{\Sigma}_{\mathrm{p}})\,,$$

   and images are assumed to be generated as a linear combination of basis vectors,

$$\mathrm{p_b}(\mathbf{I}|\mathbf{x}) = \mathcal{N}(\mathbf{I}; \mathbf{A}\mathbf{x}, \boldsymbol{\Sigma}_{\mathbf{I}})\,.$$

3. **Bayesian Encoding Model:** Assume that, conditioned on $\mathbf{I}$, the brain samples $\{\mathbf{x}^{(i)}\} \sim \mathrm{p_b}(\mathbf{x}|\mathbf{I})$, and that each value of $\mathbf{x}$ corresponds to a neuron, so that $\mathbf{r}^{(i)} = \mathbf{x}^{(i)}$. (We will use "$\mathbf{r}^{(i)}$" to denote the vector of neural activity at time i, and $\mathbf{r} = \{\mathbf{r}^{(1)}, \ldots, \mathbf{r}^{(t)}\}$ to denote all neural activity in the relevant population up to time t.)

4. **Bayesian Decoding Model:** We will derive the Bayesian decoder of s given $\{\mathbf{r}^{(1)}, \ldots, \mathbf{r}^{(t)}\}$.

We are interested in the optimal decoder of s after t time has elapsed, or $\mathrm{p}(s|\mathbf{r})$. By Bayes' rule, this is proportional to $\mathrm{p}(\mathbf{r}|s)\mathrm{p}(s) = \mathrm{p}(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)}|s)\mathrm{p}(s)$. That is, the quantity we must compute in order to optimally decode $\mathrm{p}(s|\mathbf{r})$ is the probability of seeing a given set of samples, $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)}$, provided a value of s.

Since s affects $\mathbf{r}$ through $\mathbf{I}$, this likelihood function can be evaluated by marginalizing across all possible images

$$\mathrm{p}(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)}|s) = \int \mathrm{p}(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)}|\mathbf{I})\mathrm{p}(\mathbf{I}|s)d\mathbf{I} \tag{S1}$$

We know from our definition that

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

and the posterior probability of all t independent samples for a given $\mathbf{I}$ is

$$\mathrm{p}(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)}|\mathbf{I}) = \prod_{i=1}^{t} \mathrm{p}(\mathbf{x}^{(i)}|\mathbf{I})\,.$$

Under the simplifying assumption that both $\mathrm{p_b}(\mathbf{x})$ and $\mathrm{p_b}(\mathbf{I}|\mathbf{x})$ are Gaussian, the brain's internal model posterior, $\mathrm{p_b}(\mathbf{x}|\mathbf{I})$ is also Gaussian,

$$\mathrm{p}(\mathbf{x}^{(i)}|\mathbf{I}) = \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}', \boldsymbol{\Sigma}')\,, \tag{S2}$$

where

$$\boldsymbol{\mu}' = \boldsymbol{\Sigma}'(\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{\mu}_{\mathrm{p}} + \mathbf{A}^{\top}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{I}) \qquad \text{and}$$
$$\boldsymbol{\Sigma}' = (\boldsymbol{\Sigma}_{\mathrm{p}}^{-1} + \mathbf{A}^{\top}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{A})^{-1}\,.$$

Note that the only dependence on $\mathbf{I}$ (and therefore on s) is through $\boldsymbol{\mu}'$.

Equation (S2) gives the probability of seeing a single sample $\mathbf{x}^{(i)}$ given $\mathbf{I}$. The probability of all t samples is

$$\begin{aligned}\mathrm{p}(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}|\mathbf{I}) &= \prod_{i=1}^{t}\mathcal{N}(\mathbf{x}^{(i)};\boldsymbol{\mu}',\boldsymbol{\Sigma}') \\ &= \mathcal{N}(\bar{\mathbf{x}};\boldsymbol{\mu}',t^{-1}\boldsymbol{\Sigma}')c(\mathbf{x},\boldsymbol{\Sigma}')\end{aligned}$$

where $\bar{\mathbf{x}} = t^{-1}\sum_{i=1}^{t}\mathbf{x}^{(i)}$ is the average of samples up to time t, and $c(\mathbf{x},\boldsymbol{\Sigma}')$ is a term that depends on $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}$ and on $\boldsymbol{\Sigma}'$ but not on $\boldsymbol{\mu}'$ (and therefore not on $\mathbf{I}$, so it can be dropped later).

We can now evaluate the integral in (S1) to get the probability of t samples for a given s:

$$\begin{aligned}\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} \\ &= \int \mathcal{N}(\bar{\mathbf{x}};\boldsymbol{\mu}',t^{-1}\boldsymbol{\Sigma}')c(\mathbf{x},\boldsymbol{\Sigma}')\mathcal{N}(\mathbf{I};\mathbf{T}(s),\boldsymbol{\Sigma}_{\mathrm{e-b}})\mathrm{d}\mathbf{I} \\ &= c(\mathbf{x},\boldsymbol{\Sigma}')\int \mathcal{N}(\bar{\mathbf{x}};\underbrace{\boldsymbol{\Sigma}'(\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{\mu}_{\mathrm{p}}+\mathbf{A}^{\top}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{I})}_{\text{definition of }\boldsymbol{\mu}'},t^{-1}\boldsymbol{\Sigma}')\mathcal{N}(\mathbf{I};\mathbf{T}(s),\boldsymbol{\Sigma}_{\mathrm{e-b}})\mathrm{d}\mathbf{I} \\ &= c(\mathbf{x},\boldsymbol{\Sigma}')\int \mathcal{N}(\bar{\mathbf{x}}-\boldsymbol{\Sigma}'\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{\mu}_{\mathrm{p}};\quad \underbrace{\boldsymbol{\Sigma}'\mathbf{A}^{\top}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{I}}_{\text{Let }\mathbf{x}'\equiv\boldsymbol{\Sigma}'\mathbf{A}^{\top}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{I}},t^{-1}\boldsymbol{\Sigma}')\mathcal{N}(\mathbf{I};\mathbf{T}(s),\boldsymbol{\Sigma}_{\mathrm{e-b}})\mathrm{d}\mathbf{I} \\ (*) &\propto \int \mathcal{N}(\bar{\mathbf{x}}-\boldsymbol{\Sigma}'\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{\mu}_{\mathrm{p}};\mathbf{x}',t^{-1}\boldsymbol{\Sigma}')\mathcal{N}(\mathbf{x}';\boldsymbol{\Sigma}'\mathbf{A}^{\top}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{T}(s),\boldsymbol{\Sigma}'\mathbf{A}^{\top}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\Sigma}_{\mathrm{e-b}}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{A}\boldsymbol{\Sigma}')\mathrm{d}\mathbf{x}' \\ &= \mathcal{N}(\bar{\mathbf{x}}-\boldsymbol{\Sigma}'\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{\mu}_{\mathrm{p}};\boldsymbol{\Sigma}'\mathbf{A}^{\top}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{T}(s),t^{-1}\boldsymbol{\Sigma}'+\boldsymbol{\Sigma}'\mathbf{A}^{\top}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\Sigma}_{\mathrm{e-b}}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{A}\boldsymbol{\Sigma}') \\ &= \mathcal{N}(\bar{\mathbf{x}};\boldsymbol{\mu}'',\boldsymbol{\Sigma}'')\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{\mu}'' &= \boldsymbol{\Sigma}'\left(\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{\mu}_{\mathrm{p}}+\mathbf{A}^{\top}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{T}(s)\right) \\ \boldsymbol{\Sigma}'' &= t^{-1}\boldsymbol{\Sigma}'+\boldsymbol{\Sigma}'\mathbf{A}^{\top}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\Sigma}_{\mathrm{e-b}}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{A}\boldsymbol{\Sigma}'\end{aligned}$$

In the line marked $(*)$, we changed variables to switch from an integral over $\mathbf{I}$ to an integral over $\mathbf{x}'$. This line is a proportionality because we also dropped terms that do not depend on s, including the Jacobian term from the change of variables, since later we will use this expression as a likelihood function of s.

Expanding the definition of $\mathcal{N}(\ldots)$, we can now write the posterior over s given $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}$ as

$$\begin{aligned}\mathrm{p}(s|\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(t)}) &= \mathrm{p}(s|\bar{\mathbf{x}}) \\ &\propto \mathrm{p}(s)\exp\left(-\frac{1}{2}(\bar{\mathbf{x}}-\boldsymbol{\mu}'')^{\top}\boldsymbol{\Sigma}''^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}'')\right) \\ &\propto \mathrm{p}(s)\exp\left(\bar{\mathbf{x}}^{\top}\boldsymbol{\Sigma}''^{-1}\boldsymbol{\mu}''-\frac{1}{2}\boldsymbol{\mu}''^{\top}\boldsymbol{\Sigma}''^{-1}\boldsymbol{\mu}''\right).\end{aligned}$$

Substituting $\bar{\mathbf{r}}$ for $\bar{\mathbf{x}}$ and rewriting in terms of a Linear PPC, this is

$$\mathrm{p}(s|\mathbf{r}) \propto \mathrm{g}(s)\exp(\mathbf{h}(s)^{\top}\bar{\mathbf{r}}) \qquad \text{((5) restated)}$$

where

$$\begin{aligned}\mathrm{g}(s) &= \mathrm{p}(s)\exp\left(-\frac{1}{2}\boldsymbol{\mu}''(s)^{\top}\boldsymbol{\Sigma}''^{-1}\boldsymbol{\mu}''(s)\right) \\ \mathbf{h}(s) &= \boldsymbol{\Sigma}''^{-1}\boldsymbol{\Sigma}'\mathbf{A}^{\top}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{T}(s).\end{aligned}$$