HOW DOES THE BRAIN COMPUTE WITH PROBABILITIES?

A PREPRINT

Ralf M. Haefner*
Department of Brain and Cognitive Sciences
University of Rochester
Rochester, NY
ralf.haefner@rochester.edu

Jeff Beck*
Department of Neurobiology
Duke University
Durham, NC
jeff.beck@duke.edu

Cristina Savin*
Departments of Neural Science and Data Science
New York University
New York, NY
csavin@nyu.edu

Mehrdad Salmasi
Gatsby Computational Neuroscience Unit
Max Planck UCL Centre for Computational
Psychiatry and Ageing Research
University College London
m.salmasi@ucl.ac.uk

Xaq Pitkow*
Neuroscience Institute
Department of Machine Learning
Carnegie Mellon University
Pittsburgh, PA

Department of Neuroscience
Center for Neuroscience and Artificial Intelligence
Baylor College of Medicine
Houston, TX

Department of Electrical and Computer Engineering
Department of Computer Science
Rice University
Houston, TX
xaq@cmu.edu

***equal contribution**

July 27, 2024

ABSTRACT

This perspective piece is the result of a Generative Adversarial Collaboration (GAC) tackling the question ‘How does neural activity represent probability distributions?’. We have addressed three major obstacles to progress on answering this question: first, we provide a unified language for defining competing hypotheses. Second, we explain the fundamentals of three prominent proposals for probabilistic computations – Probabilistic Population Codes (PPCs), Distributed Distributional Codes (DDCs), and Neural Sampling Codes (NSCs) – and describe similarities and differences in that common language. Third, we review key empirical data previously taken as evidence for at least one of these proposal, and describe how it may or may not be explainable by alternative proposals. Finally, we describe some key challenges in resolving the debate, and propose potential directions to address them through a combination of theory and experiments.

Keywords Bayesian brain, probabilistic inference, computation, representation, probabilistic population code, distributed distributional code, neural sampling

1 Introduction

Helmholtz observed that the sensory inputs to the brain are insufficient to give rise to the rich perceptual world that we experience, and that perception should be conceptualized as an active inference process which combines prior experiences with sensory inputs to form beliefs about behaviorally relevant features of the external world [Von Helmholtz, 1867, Kersten and Yuille, 2003]. Over the past few decades, a large body of research has supported this view, demonstrating that humans and animals display behaviors that are sensitive to the relative uncertainties in their inputs and prior knowledge [Ma, 2012]. Despite this extensive body of task specific research, there is still no consensus regarding either the means by which the computations that underlie this process are performed by neural circuits or the means by which neural activity represents uncertain beliefs[Fiser et al., 2010, Pouget et al., 2013].

In the absence of computational limitations, the optimal way to deal with uncertainty is via Bayesian inference, which provides normative means by which subjective probabilities should be updated and utilized [Laplace, 1812, Jaynes, 2003]. However, debates continue about whether inference in the brain is actually probabilistic [Rahnev et al., 2020, Orhan and Ma, 2017], and the jury is still out whether it is close enough to optimal to make Bayesian inference a useful mathematical framework for understanding the brain [Ma, 2012]. Even within the context of Bayesian inference it is unclear whether probabilistic beliefs about inferred (latent) variables are computed 'constitutively' across all latents, or are constructed 'opportunistically' in response to task demands [Koblinger et al., 2021]. For the purposes of this paper we will assume that there is some population of sensory neurons whose activity can be interpreted as a probabilistic belief about some latent variable, and we will focus on the relationship between this belief and neural activity.

A series of prior studies has addressed this question and presented models falling into three broad categories of 'neural codes': probabilistic population codes (PPCs) [Ma et al., 2006, Jazayeri and Movshon, 2006, Deneve, 2008a,b, Beck et al., 2008, 2012], distributed distributional codes (DDCs) [Sahani and Dayan, 2003, Vértes and Sahani, 2018, Pitkow, 2012], and neural sampling codes (NSCs) [Hoyer and Hyvärinen, 2003, Fiser et al., 2010, Savin and Deneve, 2014, Orbán et al., 2016, Haefner et al., 2016, Aitchison and Lengyel, 2016]. While each of these models is supported by empirical evidence, it is often unclear how well the presented data exclude alternative models, and studies that directly compare multiple coding schemes are rare [Grabska-Barwinska et al., 2013, Ujfalussy and Orbán, 2022]. Furthermore, comparisons across multiple papers are complicated by the fact that notation often differs, and differing assumptions are at times left implicit. In fact, recent work has identified some differences in assumptions and close relationships between models previously seen as mutually exclusive [Shivkumar et al., 2018, Lange et al., 2023], pointing to a need for a systematic comparison of approaches, standard notation, and shared metrics of success. This review is an attempt to develop a common language and notation by proponents of different theories of probabilistic representations. After we present a unified and consistent language for representations (Section 2) and computations with them (Section 3), we then use that language to review the basics of the three major classes of theories, as well as formal connections between them and a case study of how they compare on a simple inference task (Section 4). In Section 5 we systematically describe how these models each interpret a common set of empirical observations. Finally, in Section 6 we will note some of the inherent difficulties in comparing probabilistic coding schemes, and provide guidance for theoretical and empirical research designed to distinguish between the different theories.

1.1 Why probability in the brain?

The benefits of probabilistic computation are uncontroversial. It has been known since Laplace [Laplace, 1820] and from the Dutch Book Theorem [Ramsey, 1926] that probabilities provide the optimal way to empirically reason about the world and form decisions in the presence of uncertainty. Furthermore, ample behavioral evidence has shown that perceptual and sensorimotor decisions are sensitive to changing uncertainty in a manner approximately consistent with probabilistic inference [Knill and Richards, 1996, Knill and Pouget, 2004, Ma et al., 2006, Fiser et al., 2010, Pouget et al., 2013]. This implies that the brain represents uncertainty (if not entire probability distributions) over task relevant stimuli, implements the operations of probabilistic reasoning, and generates decisions based on that representation. We would like to understand the neural basis of these computations, and in particular whether there is a unifying theory that can explain how the brain computes with probabilities.

The existence of such a unifying theory is not a foregone conclusion. The brain may have learned to represent and manipulate probabilities in different ways for different tasks and variables it encounters. Probabilistic computations could arise in a highly flexible neural networks simply by extensive experience with naturally structured tasks [Orhan and Ma, 2017], as optimizing performance requires taking into account trial by trial fluctuations in uncertainty. Just because uncertainty must be represented and incorporated into the calculus of decision making does not mean that it is represented in the same generalizable manner for every task and latent variable, although there is some evidence in support of that claim [Houlsby et al., 2013]. That said, several theories posit that there *are* general purpose, recurring motifs for representing and computing with probabilities. These motifs should arise with similar properties across a range of variables and tasks. If such a structure were to exist, then it would provide the brain with a powerful inductive bias that would generalize efficiently to new tasks. An inductive bias favoring learning and using probabilities could be embodied in large-scale architecture, microcircuit structure, and savvy plasticity rules [Sinz et al., 2019].

1.2 What makes a 'good' neural representation?

Of course probabilistic information is already present at the retina, because one can always apply Bayes rule. That is not sufficient to count as a brain representation. Instead, a representation needs to be *used* in some way [Baker et al., 2021a].

Representations need to be evaluated in terms of how well they explain empirical observations. However, it is also important to consider how well representations can be implemented and used by the brain. Three common desiderata for a 'good' neural representation are: efficiency, representational simplicity, and computational convenience (also see Pohl et al. [2024]). Efficiency is typically measured in terms of bits per spike. It makes use of the notion that one goal of the brain is to represent as much behaviorally relevant information as it can with minimal energy expenditure. 'Representational simplicity' refers to ease of decoding of a neural representation by downstream circuits constrained by computational complexity, time, and data.

In statistical terms, a simple representation is one in which knowledge of low-order statistics like the mean and variance allow for efficient decoding. In contrast, a complex representation would relegate the encoding of objects and their poses, textures, and other properties, as well as the uncertainties about these properties, to complex, high-order statistics. A related coding principle is computational convenience. Certain representations make some computations easier to implement using the operations available to neural circuits. It is typically assumed that linear operations are easy, even though individual biological neurons are capable of more complex computations [Poirazi et al., 2003, Beniaguev et al., 2021, Jones and Kording, 2021, Gidon et al., 2020]. Two example fundamental computations of particular interest in probabilistic inference are the sum rule and product rule of probability. Later we will see that different theories of probabilistic representations give these operations different complexities.

Table 1: Glossary of notation.

symbol	meaning	symbol	meaning
$\boldsymbol{s}$	World state		
$\boldsymbol{o}$	Observation		
$\mathbf{z}$	Latent variable in brain	$p(\boldsymbol{o}\vert\mathbf{z})p(\mathbf{z})$	Generative model in brain
t	Time	$T(\mathbf{z})$	Statistic
$q(\cdot)$	Approximate posterior	η	Natural parameter
$\mathbb{E}_{q(x\vert y)}[\cdot]$	Expectation over $q(x\vert y)$	μ	Expectation parameter
$\mathbf{r}$	Neural responses	$U(s,a)$	Utility
ν	Nuisance variable (external noise)	ξ	Internal neural variability (internal noise)
$\boldsymbol{w}$	Synaptic weight	a	Action

2 What does it mean for the brain to represent probabilities?

We say that a neural activity pattern represents a probability distribution if there is a mapping between neural activity and probability distributions *and* if subsequent neural computations are consistent with the rules of probability and the proposed mapping [Luce et al., 1990, Zemel et al., 1998, Baker et al., 2021b, Lange and Haefner, 2022]. For further discussion and nuance on the nature of representations, see [Baker et al., 2021a].

In the Bayesian perspective, probabilities are subjective constructs. However, these subjective probabilities are grounded in a model of the data generation process and computations based upon that model. We assume that the brain's model of the world is based on a generative model of its sensory inputs (for a discussion of the differences between modeling the brain in terms of discriminative vs. generative models see [Peters et al., 2024]). This comprises a set of assumptions about the latent variables that generate or cause the animal's sensory inputs. A good generative model is useful because it can allow the brain to explain the sensory data by drawing inferences about those latent causes (analysis by synthesis [Kersten and Yuille, 2003]). Exact inference is intractable in general, and there will be algorithmic shortcuts or implementation constraints that lead to inferences that are only approximations to posterior distributions obtained through Bayes' rule. Theories about probabilistic brain computations often include such approximations.

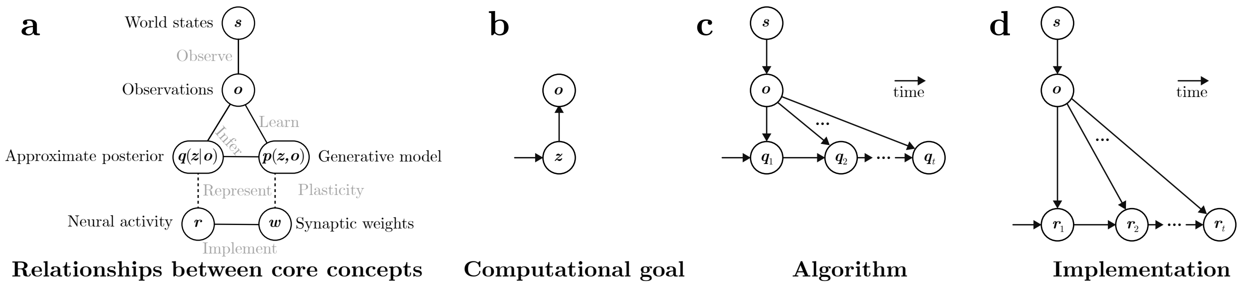

Figure 1: Relationships between key quantities for probabilistic inference. **A**: Schematic of the different elements of Bayesian inference and its neural implementation. Not to be interpreted as a graphical model! **B**: The computational goal of a Bayesian brain is to infer the brain's latent variables $\mathbf{z}$ from observations $\boldsymbol{o}$. **C**: Inferential dynamics at the algorithmic level, for a static problem. Latent causes in the world generate observations, which the brain interprets through approximate inference dynamics in terms of its own latents $\mathbf{z}$, eventually producing an action or decision. In theoretical models of inference, neural activity is a consequence of these algorithmic dynamics. **D**: In reality, the physical mechanism or implementation of this process has a different causal diagram without the interpretable approximate posteriors, and the inferential dynamics are merely an abstract interpretation of the activity in the biophysical system.

To construct a testable, quantitative theory of how the brain computes with probabilities, we need to relate probabilities and brain signals, defined based on several considerations: First, what latent variables or events $\mathbf{z}$ are the probabilities about? Second, what generative model $p(\boldsymbol{o}|\mathbf{z})p(\mathbf{z})$ are the probabilities based on? Third, what aspects of neural activity $\mathbf{r}$ represent the information about $\mathbf{z}$? Fourth, given the latent variables $\mathbf{z}$ and observations $\boldsymbol{o}$, what aspects of a posterior probability $p(\mathbf{z}|\boldsymbol{o})$, or its approximation $q(\mathbf{z}|\boldsymbol{o})$, are captured by $\mathbf{r}$? Given answers to those four questions, we can construct a model for how $\mathbf{r}$ represents $q(\mathbf{z}|\boldsymbol{o})$ (Figure 1).

2.1 What is $\mathbf{z}$? What variables are the probabilities about?

A crucial question for neural theories of probabilistic computation is, what are the intermediate latent variables $\mathbf{z}$ whose probabilities the brain may represent? These intermediate variables may have some causal status, *e.g.* the depth of different objects in a scene affects occlusion and thus the visual input, or they may be pragmatic constructs that make it easier for a brain to summarize sensory sensory input in a way that is useful for predicting the effects of actions on future inputs. For example, a set of oriented edges can be summed to create a compressed representation of an image while preserving the information needed to identify the pictured object [Olshausen and Field, 1997]. Either way, the brain has no access to an objective truth about the identities or values of intermediate latent variables and must rely on observations and the generative model assumptions about the relationships between latent variables to reason about them.

One possibility is that the brain should represent posteriors only over the variables that one can act upon. This is based on the idea that the goal of perception is, ultimately, to guide actions [Gibson, 1979, Shadlen et al., 2008]. To select good actions you need predictions about what those actions will do. Moreover, since the future is unknown and different outcomes have different consequences, it is useful to represent probabilities of those outcomes in order to evaluate the benefits and risks of different options as is done in Bayesian reinforcement learning [Dayan and Daw, 2008, Maloney and Mamassian, 2009] and active inference [Sajid et al., 2021]. In this setting, variables that have no predictive power in the relevant action space do not influence behavior and thus need not be encoded. We call task-irrelevant variables 'nuisance variables', $\boldsymbol{\nu}$, and a major computational goal of the Bayesian brain is to construct a representation of probabilities that are invariant to these nuisance variables.

An alternative possibility is that the brain constructs a task-independent model of the world that accommodates many situations, including those never seen before, and that the brain is always performing inference unconsciously, even when it is not performing a specific task ([Von Helmholtz, 1867], for a review see [Koblinger et al., 2021]).

However, even if inferring actionable variables is the eventual goal of the brain, there is often a complicated causal path to sensory observations from these latent variables. To perform inferences about the target variables, it may help to construct probability distributions over intermediate latent variables on that causal path [Peters et al., 2017]. Additionally, when tasks are amorphous and changing, it may help to represent latent variables that could later become actionable [Flesch et al., 2018], i.e. variables that lead to good generalization. Thirdly, the brain may benefit from constructing representations that facilitate subsequent learning or compress previous data. In each of these cases, representing a joint probability distribution about intermediate latent variables leads to better inferences.

2.2 What is o? What evidence are the probabilities conditioned on?

Probabilistic computations are based on evidence, whether it is the most immediate sensory observations or the evolutionary history of our ancestors. We usually assume that information from our evolutionary history is summarized in the architecture of the brain and modeled by prior belief about how the world works. Information about what we have learned from previous experience is summarized in our synapses, while the neural activity (and perhaps short-term synaptic state [Mongillo et al., 2008]) is responsible for encoding information about recent sensory evidence and relevant latent variables. This recent evidence is what we will call observations $\boldsymbol{o}$, and they determine the posterior probabilities $p(\mathbf{z}|\boldsymbol{o})$ over latent variables that are of immediate interest. Below we discuss how the brain may approximate this ideal posterior by some other distribution q. A model of probabilistic computation should therefore specify what observations $\boldsymbol{o}$ these probabilities are conditioned on.

What we count as an observation depends on the system we consider. Patterns of light are observations for the visual system, and patterns of sound are observations for the auditory system. But even within one modality, different subsystems receive different inputs: we might consider light as an observation for the retina, while the retinal ganglion cells' outputs are observations for the brain. In a broad, colloquial sense, we can consider an observation to be any input to a designated system. At the same time, there is a narrower, more technical definition of observation when we are considering probabilistic computation: an observation $\boldsymbol{o}$ is whatever the posterior $q(\mathbf{z}|\boldsymbol{o})$ is conditioned on. This requires making explicit modeling choices.

In vision, for example, one might consider $\boldsymbol{o}$ to denote the image, or the photoreceptor activations which provide the only evidence about the image, or the retinal output: none of these receive cortical feedback and thus can be treated strictly as inputs to downstream computations. Any computations performed by the retina itself, between image and retinal output, could be either modeled as part of the generative model using intermediate latent variables or, alternatively, as a potentially dynamic sensor that is not necessarily Bayesian in any meaningful way, and whose output is modeled as the observation from the perspective of the rest of the brain.

Ultimately, building a Bayesian model of some system requires defining the boundary of the system. In a Bayesian framework, system inputs constitute the observations, and the output constitute "actions." When modeling a cortical circuit, actions could simply consist of the transmission of the represented posterior. More generally, actions influence future observations and can be treated as either latents (represented by corollary discharge) or as part of subsequent observations. For example, in an active inference or Bayesian reinforcement learning setting, the goal is often to compute a posterior distribution over actions that maximize rewards. At the behavioral level, only one action is actually selected which, if directly observed, becomes part of the subsequent observations.

Other considerations that determine where observation ends and inference begins include timescales of the relevant behavior, feedback between sensory and cortical areas, and the effects of actions such as eye movement on future observations. For example, one might formalize intended actions as latents that affect future observations (e.g. as corollary discharge), or include actions themselves as observations. Because of this complexity, it is often best to adopt a systems view in which observations consist of the complete set of signals that drive the system of interest.

Where does sensory observation end and inference begin? Here, for simplicity of definition, we'll assume that the observation is constant over one inference step (whatever that means). However, this is usually not the case and things get messy. This is a tricky question, and one we do not claim to answer. There are questions of timescale, feedback, interaction with the environment, or an inability to make a clear separation between changing $\boldsymbol{o}$ and the inference given a single $\boldsymbol{o}$.

One crucial distinction between observations $\boldsymbol{o}$ on which probabilities are conditioned, and the neural activity $\mathbf{r}$ that represent those probabilities, is that $\mathbf{r}$ should summarize the relevant aspects of the recent past, whereas $\boldsymbol{o}$ provides information only from the current moment in time. Finally, the probabilities could in general incorporate all available evidence, including the animal's own actions $\boldsymbol{a}_{1:t}$, which influence the world and thus the probability through $p(\mathbf{z}|\boldsymbol{o}, \boldsymbol{a})$. One might formalize intended actions either as latents or as part of the observations via feedback from the motor plant (e.g. as corollary discharge).

2.3 What happened to s? What about the experimentally defined task stimulus?

The term "stimulus" often refers to either specific properties of observable inputs, as for gratings or random dot kinematograms [Hubel and Wiesel, 1962, Parker and Newsome, 1998], or the entire observable stimulus $\boldsymbol{o}$, as in studies

Box 1. The importance of the distinction between s and $\mathbf{z}$. The distinction between $\mathbf{z}$ and s is crucial and consequential when empirically testing the predictions of different probabilistic encoding schemes. Here, we provide three illustrations of the importance of this distinction:

Primary visual cortex (V1): Due to the strong orientation-selectivity of V1 neurons [Hubel and Wiesel, 1962], a dominant account of area V1 has been that it 'represents orientation' (among other things, like spatial frequency, binocular disparity, etc.) Equating s (here orientation) with $\mathbf{z}$ leads to the conclusion that V1 responses are incompatible with a sampling-based representation, since that would imply that higher stimulus contrast, i.e. higher certainty about the stimulus orientation, leads to narrower tuning curves. However, the tuning curves in V1 appear to be approximately contrast-invariant, i.e. scale multiplicatively with contrast, suggesting a falsification of the neural sampling hypothesis. If, on the other hand, one assumes that $\mathbf{z}$ represents the intensity, or absence and or presences, of (e.g. Gabor-shaped) image patches at the receptive field location [Olshausen and Field, 1997, Bornschein et al., 2013], then this conclusion changes, and even a sampling-based representation predicts an approximately multiplicative scaling of orientation tuning curves with contrast. In particular, [Shivkumar et al., 2018] showed that implementing neural sampling in such a sparse model appears like a probabilistic population code (PPC) when interpreted as a code over orientation (s).

Medial temporal area (MT): Hoyer and Hyvärinen were the first ones to suggest that neural responses could be interpreted as samples from a posterior distribution over latent variables (in their case intensity of localized patches learnt from natural images), and that their variability might reflect the uncertainty in the brain's beliefs [Hoyer and Hyvärinen, 2003]. However, an analysis of the responses of a large number of neurons in area MT did not find a higher response variability for stimuli whose velocity was more ambiguous — an apparent contradiction of the neural sampling hypothesis (unpublished, private communication by Eero Simoncelli). The assumption underlying these analyses was that since MT neurons are strongly tuned to motion direction, $\boldsymbol{s}$ (in addition to other variables), the represented beliefs would also be over s, such that higher uncertainty about $\boldsymbol{s}$ should be reflected in higher response variability. However, it is not clear whether MT responses in fact represent beliefs over motion direction or velocity, not some other variable $\mathbf{z}$, e.g. the absence or presence of motion primitives, which might make different predictions (in analogy to V1, see paragraph above).

Hippocampus (CA1): In a recent study, Ujfalussy and Orbán directly compared a set of neural predictions from three neural codes, PPC, DDC, and neural sampling, using simultaneous recordings from dorsal CA1 in the hippocampus. The analysis exploits the fact that neural responses during different phases of the theta cycle are believed to represent an animal's location at different times in past and future [Ego-Stengel and Wilson, 2007], with predictions increasing in uncertainty the further one goes into the future. This allows for strong qualitative predictions while avoiding the need to commit to a particular generative model about how location beliefs are updated in the light of sensory evidence. However, the analysis still relies on the assumption that the $\mathbf{z}$ about which activity in area CA1 represents beliefs is the animal's trajectory. While this assumption is at least as plausible as the assumption that V1 represents orientation, or MT represents motion direction, there is ample evidence that CA1 represents more a general variable than location [Eichenbaum, 2018, Jarzebowski et al., 2022, Sugar and Moser, 2019], and it is unclear whether the conclusions of [Ujfalussy and Orbán, 2022] would generalize for a broader definition of the latents.

with naturalistic images or sounds. Neither of these quantities necessarily directly correspond to the latent variables $\mathbf{z}$ that neurons represent, as we describe below.

Experimental stimuli $\boldsymbol{s}$ are often parameters of the observable sensory input provided to the brain $\boldsymbol{o}$. For example, videos may be built from oriented gratings or dots moving with some coherence in some direction. These variables are only a subset of the variables that define the sensory input, $\boldsymbol{o}$.

Many neuroscience studies take a philosophically distinct approach to asking questions about representations. They ask not about latent variables $\mathbf{z}$ obtained from a computation model, but instead focus on experimentally controlled or measured stimuli, $\boldsymbol{s}$, like orientation [Hubel and Wiesel, 1962] or spatial location [Sherrington, 1906, O'Keefe and Dostrovsky, 1971] that strongly modulate neural activity.

Importantly, the latent causes $\mathbf{z}$ in the brain's model of the world need not agree with the experimentally-defined stimuli $\boldsymbol{s}$. Typically, the experimentally controlled $\boldsymbol{s}$ only have some correlation with the brain's latent $\mathbf{z}$. For this reason it's important that we search for latents $\mathbf{z}$ for which neural activity forms computationally useful representations. This latent variable approach eschews intuitive definitions of stimuli, like orientation or objects, and provides a framework for discovering sophisticated latents that better describe neural activity that drives behavior (see Box 1).

2.4 What is p What is the posterior probability that the brain would like to infer?

Bayesian inference computes beliefs about unobserved latent variables z given a set of observations o and a generative model, $p(\boldsymbol{o}|\mathbf{z})p(\mathbf{z})$, by applying Bayes' rule:

$$p(\mathbf{z}|\boldsymbol{o}) \propto p(\boldsymbol{o}|\mathbf{z})p(\mathbf{z}) \tag{1}$$

for a likelihood $p(\boldsymbol{o}|\mathbf{z})$ and a prior $p(\mathbf{z})$. The likelihood captures the brain's assumptions about how observations $\boldsymbol{o}$ arise from latent causes $\mathbf{z}$. The prior captures the knowledge about the frequencies or values of causes and their dependencies on each other. This distribution is highly structured for natural inputs, and is often modeled as a hierarchical graphical model that efficiently expresses the conditional dependencies in $p(\mathbf{z})$. Together, these terms define a generative model of the world — a way of explaining how the observations were generated by latent causes in the environment. It is important to note that there need not be a direct correspondence between the brain's latents and quantities in the external world. As such, $p(\boldsymbol{o}, \mathbf{z})$, denotes the brain's *subjective* generative model of the world and $p(\mathbf{z}|\boldsymbol{o})$ denotes the posterior consistent with that generative model, and not some unknown (and unknowable) probability that describes how the world actually works, e.g. in terms of physics.

2.5 What is $q(\mathbf{z}|\boldsymbol{o})$? What is the brain's approximate inference?

Inference in the brain is necessarily approximate. We denote by $q(\mathbf{z}|\boldsymbol{o})$ as the brain's approximation to the exact posterior $p(\mathbf{z}|\boldsymbol{o})$ given its own subjective generative model (here, assumed to be fixed after learning). The general underlying assumption is that the inference dynamics try to compute a q that best approximates the desired p by some measure. While the nature of q depends on the specific approximate inference algorithm, it should be a well-defined probability distribution and not a point estimate.

Note that there is some flexibility about which approximations define the inference q versus the generative model p: one might define an altered generative model such that the q is an exact posterior according to that model. Making assumptions explicit and testing how well they generalize is the key to discriminating between such model components.

2.6 What is $\mathbf{r}$? Which neural properties do the representing?

In this paper, we will primarily consider neural activity as the seat of probabilistic computation. Previously proposed candidates in this context are membrane potentials [Orbán et al., 2016], spikes [Buesing et al., 2011, Pecevski et al., 2011, Legenstein and Maass, 2014, Savin and Deneve, 2014], and spike rates [Hoyer and Hyvärinen, 2003, Ma et al., 2006, Vasudeva Raju and Pitkow, 2016]. Some theories posit that probabilities are represented as a spatial code of spike counts in a long temporal window, manifested across neurons [Ma et al., 2006, Vértes and Sahani, 2018]. Other theories including temporal sampling and timing codes [Hoyer and Hyvärinen, 2003, Berkes et al., 2011, Orbán et al., 2016, Savin and Deneve, 2014] assert that the time series is the locus of probabilistic representations. Most generally, patterns of neural activity across both space and time may represent probabilistic information [Savin and Deneve, 2014].

It is an open empirical question which of these types of neural response properties provide the most parsimonious description of probabilistic neural computations. Since spike times can be seen as a summary statistic of the underlying membrane potentials, and spike rates of the underlying spike times, a key question will be whether the respective lower level of description will have predictive power beyond that provided by the higher level description [Hoel et al., 2013].

2.7 What is the relationship between $\mathbf{r}$ and q? How does neural activity represent probabilities?

Neural activity evolves according to biophysical mechanisms. Probabilistic models propose that these mechanisms can be interpreted *as if* they are implementing meaningful computations. The hypothesized link between $\mathbf{r}$ and q specifies the relationship between some biophysical properties and computationally meaningful ones (e.g. parameters of q, or samples from it, Figure 2). This link determines whether neural representations are mixed (multiple parameters or samples or statistics of q contributing to each neuron's responses) [Rigotti et al., 2013] or 'pure' (with only one aspect of q contributing to each neuron) [Hoyer and Hyvärinen, 2003, Fiser et al., 2010]. And it determines whether q is represented by individual neurons, or distributed across populations. Specifying this link is crucial for making testable neurophysiological prediction from any computational theory. Indeed, the mathematical link between $\mathbf{r}$ and q constitutes a primary distinction between the various probabilistic representational schemes.

Some coding schemes assume that the map between $\mathbf{r}$ and q is stochastic, even when the observation $\boldsymbol{o}$ is fixed. Trivially, this occurs when neural dynamics are noisy or not fully observed (nuisance variability), but can also be a product of a inference algorithm that that relies on optimization via stochastic gradient descent. These kinds of nuisance and computational variability are expected to be present even when parameteric schemes are utilized. Sampling based schemes, on the other hand, assume that neural variability is part of the approximate q itself, i.e. neural variability

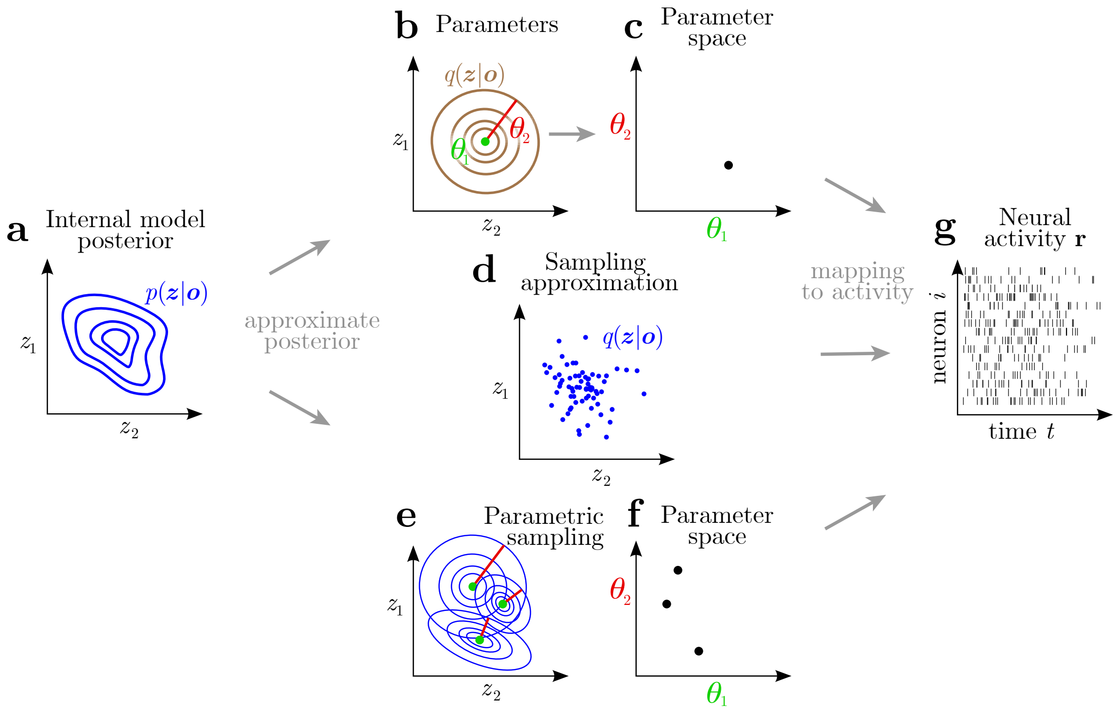

Figure 2: Examples of how a posterior distribution $q(\mathbf{z}|\boldsymbol{o})$ (**a**) could be mapped to neural activity. The distribution could be parameterized (**b**) and these parameters (**c**) could be mapped to neural activity (**e**). Alternatively, samples from the posterior (**d**) could determine neural responses. It is also possible to interpolate between these options, sampling the parameters [Lange et al., 2022] (Adapted from [Lange and Haefner, 2022].)

is proportional to posterior's uncertainty [Lange and Haefner, 2022]. From an encoding perspective, the posterior is "located" in the circuit and sensory input that together generates stochastic responses. On the other hand, downstream circuits do not have access to this process, and only can access the stochastic neural activity. From a decoding perspective, therefore, we can say there may be a *distribution* over posterior distributions, and write any realized posterior as a sample from it, $q(\mathbf{z}|\boldsymbol{o}) \sim P[q(\mathbf{z}|\boldsymbol{o})]$. This stochasticity has important implications for how we interpret neural variability, a topic we will return to in section 5.2.

3 What are the dynamics of probabilistic computations?

To understand how the brain computes with probabilities, we need to relate the dynamics of the neural activities $d\mathbf{r}/dt$ to the dynamics of the posterior probabilities dq/dt they represent. The specific relationship between these two quantities depends both on the format of the probabilistic representation and on the inference algorithm used to update approximate posteriors.

Additionally, on a slower timescale, we assume that the circuits' parameters changes over time as $d\theta/dt$ as the circuit learns an improved generative model through dp/dt and an improved approximate inference model. The next sections describe crucial properties of each of these computational dynamics.

3.1 What are the relevant timescales?

Computations unfold over time. For a Bayesian brain, there are multiple temporal processes that are helpful to distinguish. *Learning:* At the slowest timescale, the brain learns the rules of the world. This corresponds to changes in the generative model, $\frac{d}{dt}p(\mathbf{z}, \boldsymbol{o})$. *Inference:* This generative model describes a world of dynamic latent variables $\mathbf{z}_t$ that can be inferred using a time series of observations $\boldsymbol{o}_t$. These observations yield dynamic posteriors, $\frac{d}{dt}p(\mathbf{z}_t|\boldsymbol{o}_{-\infty:t})$, changing over time either as new evidence comes in, or as the brain anticipates future states. These posteriors may be dynamic even for an ideal Bayesian computation with unbounded computational abilities, since the dynamics are imposed by the time-dependent observations. *Algorithm:* At a faster timescale, an approximate Bayesian brain will use computations that unfold over time. Typically, those computations are iterative, and converge to the desired endpoint,

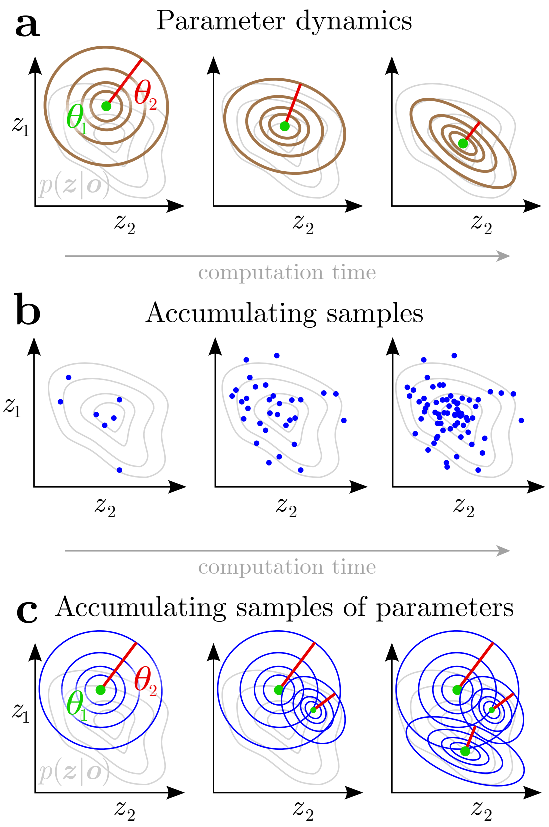

Figure 3: For a static posterior, $p(\mathbf{z}|\boldsymbol{o})$, the approximate posteriors, $q_t(\mathbf{z}|\boldsymbol{o})$ generally change over time. **a**: Parametric codes allow parameters $\boldsymbol{\theta}_t$ to depend on time. **b**: Dynamics can produce sequences of samples, which when accumulated gradually fill out the posterior. **c**: Dynamics can also produce sequences of sampled parameters [Lange et al., 2022] rather than samples of latent variables.

through dynamics $\frac{d}{dt}q_t(\mathbf{z}|\boldsymbol{o})$. [1] *Implementation:* Since a Bayesian brain requires a mapping between approximate posterior q and neural activity $\mathbf{r}$, a dynamic algorithm will manifest as neural dynamics $\frac{d}{dt}\mathbf{r}$ even for a constant q.

For simplicity, we will assume that the targeted posterior $p(\mathbf{z})$ is constant over time, using evidence $\boldsymbol{o}$ that is also constant over time, so that neural dynamics correspond only to the algorithm and its implementation. In general, this is an unrealistic assumption as natural tasks invariably involve dynamic latent variables $\mathbf{z}_t$ and series of observations $\boldsymbol{o}_t$. Nonetheless, the restriction to static cases is useful here because it allows us to more easily distinguish different model types. The concepts we describe in this paper can be extended to inference over a dynamic world.

3.2 What is $\dot{q}$? What are the inference dynamics?

For a fixed observation and a fixed posterior $p(\mathbf{z}|\boldsymbol{o})$, the computation of an approximate q unfolds over time, hopefully bringing q closer to p (Fig.3). In temporal or spatiotemporal codes, the approximate posterior q only manifests as a time series, so subsequent computations need to synthesize information across time. In contrast, spatial codes represent the q completely within each relevant time window.

For parametric codes and computations, the changes in the approximate posterior are captured by changes in the parameters, $q_t(\mathbf{z}|\boldsymbol{o}) = q(\mathbf{z}|\boldsymbol{o}, \eta_t)$, while for neural sampling codes they are the result of the changing set of samples.

[1] We define $q \equiv q_t$ as the distribution implied by the current approximate computations, e.g. by the finite number of samples that have been generated, or by the current values of the parameters of q even if they keep changing as part of an iterative algorithm.

3.3 What is $\dot{p}$? How does the generative model change as it learns?

As the brain gains experience in its environment, it can improve its model of the world, and use these changes to improve its inferences. Since we describe the brain's generative model of the world by $p(\mathbf{z}, \boldsymbol{o})$, we refer to the learning-induced changes in that model as $\dot{p}$. Changes in an internal model are often attributed to synaptic plasticity $\dot{\boldsymbol{w}}$, although there may be other physical mechanisms that may contribute, such as changes in bias or local nonlinearities, or even changes in the dynamics of short-term depression or facilitation.

3.4 What is $\dot{r}$? How do neural dynamics relate to computational dynamics?

Biological mechanisms cause neural activities to evolve over time, determining the dynamics $\dot{\mathbf{r}}$. In a Bayesian framework, these changes in $\mathbf{r}$ are interpreted as changes in q as it both incorporates new data and evolves toward the approximate posterior via the action of the inference algorithm employed.

Note that there may be some neural dynamics that are not relevant to the probabilistic representation, just as there may be some aspects of neural activity that do not encode relevant probabilities.

4 Models of probabilistic representations

Now that we've introduced the core ingredients of probabilistic models of the brain, we will consider these ingredients for three major model classes: Probabilistic Population Codes (PPCs), Distributed Distributional Codes (DDCs), and Neural Sampling Codes (NSCs). Each has its own computational advantages and disadvantages, and the brain may incorporate elements of more than one model. In Section 5 we will review the empirical evidence in support of each.

4.1 PPCs

Key idea: Probabilistic Population Codes (PPCs) assume that linear functions of neural activity represent *natural parameters* of a distribution, a core concept in probability we explain below. A direct consequence is that neural activity represents log probabilities. This makes it easy to multiply probabilities [Ma et al., 2006, Rao, 2004], as needed for cue combination and evidence integration. However, the other main probabilistic operation, marginalization, is more difficult.

What is q? In PPCs, neural activity represents exponential family probability distributions over latent variables $\mathbf{z}$ by using simple encodings of natural parameters, $\boldsymbol{\eta}$. Natural parameters are a mathematically convenient way to parameterize a probability distribution: in $q(\mathbf{z}) \propto \exp[\boldsymbol{\phi}(\mathbf{z})^\top \boldsymbol{\eta}]$, the natural parameters are coefficients of sufficient statistics $\boldsymbol{\phi}(\mathbf{z})$ for that distribution. For any given parameterized distribution, there is a unique relationship between natural parameters and the expectations of the sufficient statistic. For example, in a Gaussian distribution, the natural parameters are the the inverse variance and the mean divided by the variance.

What is r? As the name implies, PPCs are population codes, and authors typically assume that the relevant aspect of neural activity is firing rate or spike counts in large populations in some small time window.

Mapping between r and q: A PPC assumes that the natural parameters, $\boldsymbol{\eta}$, are linear functions of neural activity $\mathbf{r}$: $\boldsymbol{\eta}(\boldsymbol{o}) = M\mathbf{r}(\boldsymbol{o})$. This allows us to write the posterior distribution as

$$q(\mathbf{z}|\mathbf{r}(\boldsymbol{o})) = \exp\left[\boldsymbol{\phi}(\mathbf{z})^\top M\mathbf{r}(\boldsymbol{o}) + \mathrm{const}(\boldsymbol{o})\right] \tag{2}$$

where $(\phi(\mathbf{z})^\top M)_i$ represents the contribution of neural response r_i to the posterior log probability over $\mathbf{z}$. The basis functions $\boldsymbol{\phi}(\mathbf{z})$ determine the sufficient statistics of the associated exponential family distribution. For example, if $\boldsymbol{\phi}(\mathbf{z})$ is restricted to quadratic functions, then the associated posterior is a multivariate Gaussian.

PPCs only describe the dimensions of neural activity that are relevant to encoding the posterior. Other orthogonal dimensions of $\mathbf{r}$ are free to vary. (For downstream computations, these other dimensions serve as internal nuisance variables.) However, additional hypotheses about the neural activity may further constrain the connection between the task-relevant and -irrelevant aspects. For example, if $\mathbf{r}$ represents spike counts, as for independent Poisson neurons, then these responses must be integers. This can constrain the particular posteriors that can be represented, and may also constrain otherwise task-irrelevant variations that ensure that spike counts are integers. These additional assumptions are critical for making detailed neural predictions, to which we will return in Section 5.

What is $\dot{q}$? In this paper, we focus on static inference problems, which can in principle be solved as a static nonlinear transformation of the sensory input, such as through a simple feedforward neural network with no dynamics. However, these problems may also be solved through an iterative algorithm [Vasudeva Raju and Pitkow, 2016]. If the code remains consistent over time, then the neural dynamics would then correspond to iterative updates of the posterior q. Conversely, approximate inference schemes such as variational inference produce updates to natural parameters for the posterior, and therefore imply specific dynamics for the neural activity in the probabilistic coding dimensions [Beck et al., 2012].

What is $\dot{\mathbf{r}}$? In PPCs, neural activity is linearly related to natural parameters, so when probabilities are multiplied, neural activity is added. One consequence is that the amplitude of the neural response encodes confidence, with higher amplitudes corresponding to narrower posteriors. In contrast, marginalization is more difficult to perform on the neural activity, requiring non-linear operations such as coincidence detection and divisive normalization [Beck et al., 2011].

What is $\mathbf{z}$? The PPC literature has largely focused on task-relevant latent variables in laboratory experiments performed by overtrained animals, such as orientation and contrast, or direction of motion and coherence. The initial focus on task relevant latents and decision variables has led to the misguided criticism that PPCs only applicable to simple tasks or are not fully Bayesian. However, subsequent work showed how more general latent variables could be represented by PPCs [Beck et al., 2011, 2012, Vasudeva Raju and Pitkow, 2016] with network implementations that allow a variety of flexible computations on multivariate generative models.

What is $\boldsymbol{o}$? When investigating specific computations, $\boldsymbol{o}$ is typically assumed to be either the sensory input or a pattern of neural activity that arises from the sensory periphery. For example, in vision, $\boldsymbol{o}$ could be either the image itself, the photoreceptor absorptions, or the output of retinal ganglion cells. In an odor discrimination task, $\boldsymbol{o}$ could be the activity of olfactory receptor neurons while the probabilistically encoding $\mathbf{r}$ is the activity of downstream neurons in the olfactory bulb and piriform cortex.

4.2 DDCs

Key idea: In the absence of uncertainty, the brain can represent the deterministic value of the latent variable $\mathbf{z}$ through a set of neuronal encoding functions $\{\phi_i(\mathbf{z})\}_{i=1}^{K}$ (Fig.4). Aligned with the conventional notion of tuning functions, the average firing rate of the neuron for the unknown value $\mathbf{z}_0$ is given by $\mathbb{E}[\mathbf{r}_i] = \phi_i(\mathbf{z}_0)$, and some noise model (e.g. Poisson noise) captures the variability of the firing rate around the mean. However, due to the noise in the sensory system and intrinsic epistemic uncertainty, the brain generally needs to deal with a distributional belief over the latent variable $\mathbf{z}$, i.e. $p(\mathbf{z}|\boldsymbol{o})$. Here, $p(\mathbf{z}|\boldsymbol{o})$ refers to the exact posterior distribution in the generative model. A natural extension of the notion of tuning function is to assume that the firing rate of the neuron is determined by the weighted sum of the values of its tuning function at potential instances of the random variable $\mathbf{z}$, where the weights correspond to the probability of the instances [Zemel et al., 1998].

The firing rate of the neuron $i \in \{1, ..., K\}$, is determined by

$$\mathbf{r}_i = \int_{\mathbf{z}} \phi_i(\mathbf{z}) p(\mathbf{z}|\boldsymbol{o}) d\mathbf{z}. \tag{3}$$

The distributional belief $p(\mathbf{z}|\boldsymbol{o})$ is therefore represented by K expected values of the encoding functions, abbreviated here by the vector $\mathbf{r}$ (Fig.4).

The codes of this scheme are referred to as Distributed Distributional Codes (DDC), as they provide a representation for "distributional" beliefs, and the representation is actualized through the "distributed" activity of neurons. DDC is a natural extension of tuning functions—in the absence of uncertainty, the posterior distribution is a Dirac delta function, $p(\mathbf{z}|\boldsymbol{o}) = \delta(\mathbf{z} - \mathbf{z}_{\boldsymbol{o}})$, and the DDC values are $\mathbf{r}_i = \phi_i(\mathbf{z}_{\boldsymbol{o}})$, $i = 1, 2, ..., K$, which align with the definition of tuning functions.

Distributed distributional coding was first introduced in [Zemel et al., 1998] as "extended Poisson model" to provide an encoding scheme for distributions. In addition to the encoding scheme, the decoding of the distribution is discussed, and it is argued that one can actually find a distribution over all the distributions that are consistent with the vector of firing rates $\mathbf{r}$. Nevertheless, to simplify computations, an algorithm is suggested to approximate MAP estimate from the distribution over the distributions. The model was later renamed to "distributional population codes" in [Zemel and Dayan, 1998], where it was used to explain the single-cell recordings and behavioral data in a multiple-motion task. The framework was subsequently extended to "doubly distributional population codes" to capture both uncertainty and multiplicity simultaneously [Sahani and Dayan, 2003].

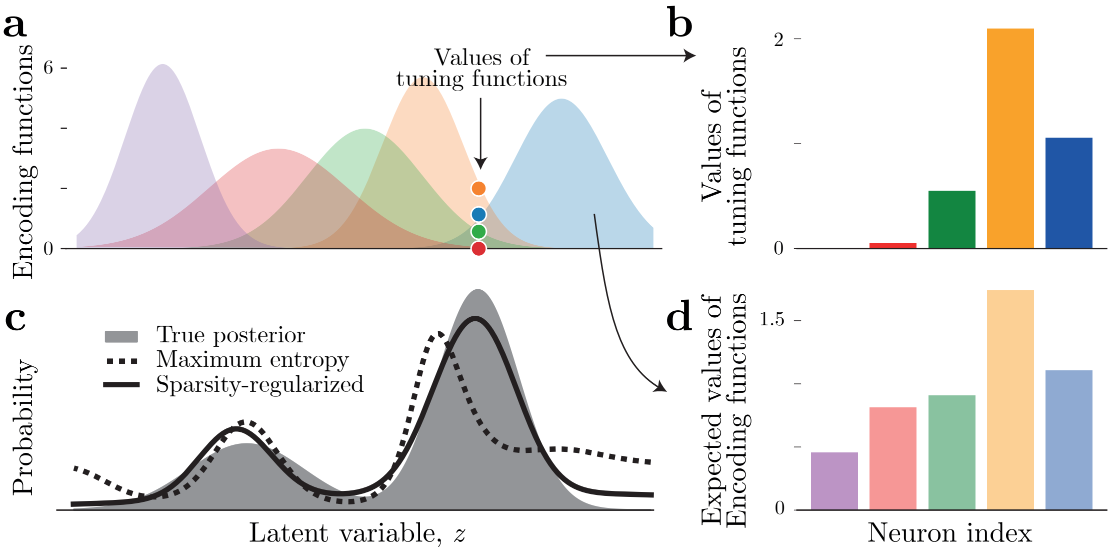

Figure 4: Distributed Distributional Codes (DDC): representation and decoding. (**a**) Five DDC encoding functions are assumed to represent the distribution. In the absence of uncertainty, e.g. $p(\mathbf{z}|\boldsymbol{o}) = \delta(\mathbf{z} - \mathbf{z}_{\boldsymbol{o}})$, the values of the encoding functions at $\mathbf{z}_{\boldsymbol{o}}$ represent the deterministic value of the latent variable (filled circles and the bar plot in **b**). (**c**) Under uncertainty, we illustrate the exact posterior distribution, $p(\mathbf{z}|\boldsymbol{o})$, as a mixture of two Gaussian distributions (gray). The DDC representation is based on the expected values of encoding functions under the full posterior distribution (**d**). The approximate posterior, $q(\mathbf{z}|\boldsymbol{o})$, that is decoded from the representation depends on additional decoding choices, two of which are shown here: Dashed black line: the maximum entropy distribution derived from the DDC values in **d**. Solid black line: the sparsity-regularized decoding of the belief.

It should be noted that prior to the linear encoding approach in [Zemel et al., 1998], a linear decoding had been suggested in [Anderson, 1994, Anderson and Van Essen, 1994]. Instead of defining the encoding process, one can assume that the distributional belief $p(\mathbf{z}|\boldsymbol{o})$ is a weighted sum of some basis functions $\psi_i(\mathbf{z})$. Therefore, the distributional belief can be computed through $p(\mathbf{z}|\boldsymbol{o}) = \sum_i \rho_i \psi_i(\mathbf{z})$, where ρ_i is the neuronal activity used to decode the distribution. In this approach, which is similar to kernel density estimation, the basis functions do not have direct relations with the tuning functions of the neurons. To find the decoding coefficients, ρ_i, different methods have been suggested including projection methods and EM algorithms. In the projection method, coefficients are computed by projecting the probability distribution on the basis functions, leading to a representation for ρ_i which is similar to the encoding in [Zemel et al., 1998].

How it works: Probabilistic computation and inference frequently involves deriving the expected value of some function $f(\mathbf{z})$. One basic approach is to calculate $\mathbb{E}_{p(\mathbf{z}|\boldsymbol{o})}[f(\mathbf{z})]$ is to decode the posterior distribution given the DDC values $\mathbf{r}$, find an estimate of $p(\mathbf{z}|\boldsymbol{o})$, denoted by $q(\mathbf{z}|\boldsymbol{o})$, and then compute the expected value through $\mathbb{E}_{q(\mathbf{z}|\boldsymbol{o})}[f(\mathbf{z})]$. However, there is a simpler way. The primary characteristic of the DDC framework is its ability to compute expected values without needing to decode the probability distributions explicitly. Let $f(\mathbf{z}) = \sum_i c_i \phi_i(\mathbf{z})$ represent a linear expansion of $f(\mathbf{z})$ with the encoding functions as the basis set. Then

$$\begin{aligned} \mathbb{E}_{p(\mathbf{z}|\boldsymbol{o})}[f(\mathbf{z})] &= \sum_i c_i \mathbb{E}_{p(\mathbf{z}|\boldsymbol{o})}[\phi_i(\mathbf{z})] \qquad (4) \\ &= \sum_i c_i \mathbf{r}_i \qquad (5) \end{aligned}$$

Therefore, within the DDC framework, the expected value of an arbitrary function can be estimated by a weighted sum of the DDC values, significantly facilitating inference and learning procedures [Vértes and Sahani, 2018, 2019, Wenliang and Sahani, 2020].

What is $\boldsymbol{o}$? In the DDC framework, the observation $\boldsymbol{o}$ denotes the set of signals provided by the sensory system; alternatively, it signifies the output of the preceding processing stages that act as the "observation" for the subsequent layer. For example, in the early stages of visual processing, the retinal activity constructs the observation $\boldsymbol{o}$, and is employed for inferring the DDC of posterior distribution over visual features, such as orientation and color. These inferred visual features can then be considered as "observations" in a generative model for higher-order processing, enabling another level of inference, like assessing the value of a shape in a decision-making task. It is worth considering that in theory, one can envision a scenario where all inferences occur within an extensive hierarchical generative model. The sole observations in this context are the sensory signals, while the remaining variables constitute the latent variables requiring inference. Nevertheless, it appears that in order to maintain computational feasibility, the brain might employ separate generative models and propagate essential distributional beliefs among them.

What is $\mathbf{z}$? The latent variable $\mathbf{z}$ encompasses any "unobserved" variable in the generative model that needs to be inferred. Within DDC, each latent variable is associated with a set of encoding functions whose expectations, with respect to the posterior distribution, determine the firing rates of the corresponding neurons. Latent variables in V1, for instance, may be linked to the orientation of image patches, whereas in CA1, they might denote the spatial location. The DDC encoding functions would then encode orientation in V1 and location in CA1. It is important to highlight that the concepts of orientation, location, and the like, as perceived by the observer, may not align precisely with the encoded information within the latent variables. The interpretation of the latent variables and establishing their connections with features in the external world poses a significant challenge. It is not yet fully understood how the brain dissects the stimuli to implant efficient features into the latent variables, and how an external observer can decrypt these features.

What is $\mathbf{r}$? The neuronal firing rate is primarily defined by the expected value of the neuron's encoding function with respect to the posterior distribution. However, there are nuances that give rise to diverse extensions of this definition. One might incorporate a noise model to account for the variability in the firing rate that is unexplainable by the posterior dynamics. For example, it can be a simple additive white Gaussian noise ϵ,

$$\mathbf{r} = \mathbb{E}_{p(\mathbf{z}|\boldsymbol{o})}[\phi(\mathbf{z})] + \epsilon, \tag{6}$$

or alternatively, the neuronal noise may manifest as Poisson noise, with an average given by $\mathbb{E}[\mathbf{r}] = \mathbb{E}_{p(\mathbf{z}|\boldsymbol{o})}[\phi(\mathbf{z})]$.

The other extension deals with the mapping between $p(\mathbf{z}|\boldsymbol{o})$ and $\mathbf{r}$ which can be non-linear,

$$\mathbf{r} = h\left(\mathbb{E}_{p(\mathbf{z}|\boldsymbol{o})}[\phi(\mathbf{z})]\right) \tag{7}$$

where h is a non-linear function. Indeed, the combination of a non-linear mapping and a noise model would be a feasible approach for further extension.

What are p and q? In a generative model with observation $\boldsymbol{o}$ and the latent variable $\mathbf{z}$, conducting exact inference yields the distributional belief $p(z|o)$. The DDC framework operates under the assumption that the brain's access is restricted to the DDC values $\mathbf{r}$, with the flexibility for the approximate posterior $q(\mathbf{z})$ to be any probability distribution consistent with these DDC values. To decode the probability distribution, one needs to find the probability distributions $q(z)$ that satisfy the constraints $\mathbf{r}_i = \int q(\mathbf{z})\phi_i(\mathbf{z})d\mathbf{z}$, for $i = 1, 2, ..., K$. This problem is known as generalized moment problem and have been studied extensively, initially for polynomial moments and later for generalized moments [Schmüdgen et al., 2017, Kemperman, 1968]. A given DDC vector $\mathbf{r}$ can correspond to none, one, or more than one distribution. When multiple distributions are associated with $\mathbf{r}$, various optimization functionals can be utilized to find the "optimal" posterior $q^*(\mathbf{z}|\boldsymbol{o})$. For a given utility functional U, the optimal posterior is derived from

$$q^*(\mathbf{z}|\boldsymbol{o}) = \underset{q}{\operatorname{argmax}}\, U(q) \quad \text{subject to:} \tag{8}$$

$$\int q(\mathbf{z})\phi_i(\mathbf{z})d\mathbf{z} = \mathbf{r}_i, \quad \text{for } i \in \{1, 2, ..., K\} \tag{9}$$

$$\int q(\mathbf{z})d\mathbf{z} = 1 \tag{10}$$

$$q(\mathbf{z}) \geq 0 \tag{11}$$

A well-known choice for the utility functional is the entropy of the probability distribution, $U(q) = -\int_{\mathbf{z}} q(\mathbf{z}) \log q(\mathbf{z})d\mathbf{z}$. The maximum entropy solution of the generalized moment problem is an exponential family distribution [Wainwright and Jordan, 2008].

$$q(\mathbf{z}|\boldsymbol{o}) = \exp\left(\sum_{i=1}^{K} \phi_i(\mathbf{z})\eta_i - A(\eta_1, \eta_2, ..., \eta_K)\right), \tag{12}$$

where $A(.)$ is the log-partition function. This approach yields the probability distribution with the highest uncertainty while satisfying the expectation constraints imposed by the DDC values. The natural parameters, η, correspond to the Lagrange multipliers used for solving the optimization problem.

Alternative utility functionals include measures such as the sparsity of distribution in a given basis, the smoothness of the distribution, and others. In Fig. 4, the maximum entropy distribution and sparsity-regularized distribution have been derived for the given DDC values. Another proposition involves considering the set of distributions $q(\mathbf{z}|\boldsymbol{o})$ that satisfy the expectation constraints and deriving a posterior distribution over all these distributions [Zemel et al., 1998]. In fact, by assuming a sparsity measure for optimization and additive noise for the DDC values, one can use an empirical Bayes method [Ji et al., 2008] to derive the posterior distribution over all the beliefs that satisfy the DDC values [Salmasi and Sahani, 2022].

As previously mentioned, DDC endeavors to bypass the decoding of the distributions $q(\mathbf{z}|\boldsymbol{o})$—instead, it transforms probabilistic inferences/computations into the determination of expected values of certain functions. These expectations are computed straightforwardly through weighted sums of the DDC values.

Mapping between q and $\mathbf{r}$: Any approximate posterior distribution $q(\mathbf{z}|\boldsymbol{o})$ should satisfy the generalized moment constraints $\int q(\mathbf{z})\phi_i(\mathbf{z})d\mathbf{z} = \mathbf{r}_i$, for $i = 1, 2, ..., K$, though the noise variance will allow some deviations from equality. The mapping from $\mathbf{r}$ to q is related to the generalized moment problem that was discussed in the previous part. To derive the maximum entropy distribution for the given DDC $\mathbf{r}$, the natural parameters $\eta(\mathbf{r})$ are calculated from the set of non-linear equations,

$$\int_z \phi_i(\mathbf{z}) \exp\left(\sum_{i=1}^{K} \phi_i(\mathbf{z})\eta_i - A(\eta_1, \eta_2, ..., \eta_K)\right) d\mathbf{z} = \mathbf{r}_i, \quad i \in \{1, 2, ..., K\} \tag{13}$$

where

$$A(\eta_1, \eta_2, ..., \eta_K) = \log \int \exp\left(\sum_{i=1}^{K} \phi_i(\mathbf{z})\eta_i\right) d\mathbf{z}, \tag{14}$$

and a closed-form distribution is obtained for $q(\mathbf{z}|\boldsymbol{o})$. However, as stated earlier, maximum entropy is not the sole approach to derive $q(\mathbf{z}|\boldsymbol{o})$ from $\mathbf{r}$.

What is $\dot{r}$: Various algorithms have been suggested for conducting inference and learning in the DDC framework. Helmholtz machines propose an elegant method for joint inference and learning—the wake-sleep algorithm iteratively refines the generative model and recognition network of the Helmholtz machine, and concurrently, learns the generative model of the world and acquires the capacity to infer the latent variables.[Dayan et al., 1995].

DDC has been integrated into the Helmholtz machine to provide a biologically plausible mechanism for inference and learning [Vértes and Sahani, 2018]. It is assumed that the generative model belongs to deep exponential family models—the conditional probabilities and priors are exponential families each characterized by specific sufficient statistics. The recognition network has a similar hierarchical structure and each layer is associated with a set of DDC encoding functions. The recognition network performs inference by mapping the observations to the DDC values of each layer, and the recognition outputs are interpreted as representing a posterior distribution with maximum entropy, meaning that the approximate posterior corresponds to an exponential family.

During the sleep phase, the generative model is used to generate samples of latent variables and sensory observations (dream sequence). The goal of the recognition network is to minimize the Kullback-Leibler divergence between the deep exponential family distribution of the generative model and the approximate maximum entropy distribution of the recognition network. Since both probability distributions are from exponential family, the parameters of the recognition network are modified to minimize the difference between the DDC values of the recognition network and the expectations of the sufficient statistics of the generative model.

During the wake phase, sensory observations are collected and utilized by the recognition network for inferring the DDC values of the posterior distributions over the latent variables. Subsequently, the sensory observations and the DDC values are used to update the parameters of the generative model in order to increase the variational free energy. It is shown that the gradient of the free energy can be derived by calculating the expected value of some functions of the latent variables. This means that by using a linear expansion for these functions, the gradient of free energy can be approximated by weighted sums of the DDC values. The remaining issue is the learning of expansion coefficient that can be conducted using the generated samples in the sleep phase [Vértes and Sahani, 2018].

What is $\dot{q}$: The dynamics of the approximate posterior q is intricately linked to the dynamic evolution of the DDC values $\mathbf{r}$. In the Helmholtz machine, for example, the sensory observations and the current weights of the recognition network determine the DDC values of each layer. By adopting the maximum entropy distribution, the conditional distribution of each layer is mapped to an exponential family, whose sufficient statistics are the DDC encoding functions of that layer and the natural parameters are calculated by the given DDC values. It is worth emphasizing that in the realm of DDC computations, there is a possibility of encountering DDC values that lack feasibility, meaning there is no corresponding distribution for them. Nevertheless, with a rich set of encoding functions, the probabilistic computations can still maintain a high degree of precision.

What is $\dot{p}$: Depending on the generative model employed, various algorithms can be used for learning the model's parameters. In the case of a deep exponential family model, the parameters of the generative model can be learnt by the wake-sleep algorithm, as discussed previously [Vértes and Sahani, 2018]. The natural parameters of each layer in the deep exponential family are calculated by a parametrized function of the parent variable. The parameters of these functions are updated by calculating the gradient of the variational free energy. The generative model is affected both in wake and sleep phases. The variational free energy is calculated through the expected values of some linear functions of sufficient statistics. The expectations are calculated by the weighted sums of the DDC values during the wake phase, and the weights of the linear functions (expansion coefficients) are learnt through the samples of the generative model in the course of the sleep phase.

4.3 Neural sampling

Key idea: A probability distribution can be approximated by a collection of samples from it, rather than by its parameters. The key idea underlying 'neural sampling' is that the neural activity in small time bins can be interpreted as one or more samples from the brain's posterior, and that over time or space, the distribution of neural activity reflects the posterior [Hoyer and Hyvärinen, 2003, Fiser et al., 2010]. To implement this, stochastic recurrent dynamics explore the state space of latent variables, occupying states in proportion to their posterior probability. Given these samples, the brain can directly estimate expectations of any function of the latent variables, which is helpful for choosing actions. For example, expectations can compute the posterior mean to generate a single estimate, a posterior variance to quantify uncertainty, or expected reward to compare states and actions.

What are p and q? Samples may be drawn either from the exact posterior p, or from an approximate posterior distribution, $q(\mathbf{z}|o)$, as in stochastic variational inference [Savin et al., 2011, Hoffman et al., 2013].

What is r? Neural sampling dynamics come in three main flavors, differing by which aspect of neural activity encodes the samples: (1) membrane potential (continuous $\mathbf{z}$)[Orbán et al., 2016, Bányai et al., 2019], (2) spike/no spike (binary $\mathbf{z}$) [Buesing et al., 2011, Pecevski et al., 2011, Haefner et al., 2016, Shivkumar et al., 2018], or (3) firing rate (continuous latent $\mathbf{z}$) [Hoyer and Hyvärinen, 2003, Haefner et al., 2016, Echeveste et al., 2020].

What is the mapping between q and r? Most proposals assume a one-to-one map between latent dimensions and responses of individual neurons, although the two can also be related less directly via a linear map, $\boldsymbol{M}$ [Savin and Deneve, 2014].

$$q(\mathbf{z}|\boldsymbol{o}) = \frac{1}{n}\sum_{k=1}^{n}\delta(\boldsymbol{M}\mathbf{r}^{(k)} - \mathbf{z}) \tag{15}$$

where n is the number of samples.

What is $\dot{q}$? Most neural sampling proposals consider static inference problems, in which a posterior is inferred for a given stimulus. On that (shortest) timescale, $\dot{q}$ simply reflects the additional samples generated over time, successively refining the posterior approximation. The sampling idea can be expanded to time-varying inference problems in which the posterior evolves over time on the time scale of the stimulus dynamics, for instance by neural dynamics analogue of particle filtering [Lee and Mumford, 2003, Kutschireiter and Pfister, 2018].

What is z? Prior work on neural sampling has primarily focused on generative models of natural images, including linear Gaussian models and sparse variants [Olshausen and Field, 1996, 1997, Hoyer and Hyvärinen, 2003, Haefner et al., 2016, Shivkumar et al., 2018] or Gaussian scale mixtures [Schwartz and Simoncelli, 2001, Wainwright et al., 2002, Orbán et al., 2016, Bányai et al., 2019]. Other examples include a hierarchical extensions of these [Haefner et al., 2016, Bányai et al., 2019, Csikor et al., 2023], a sparse linear Poisson model of olfactory inputs [Grabska-Barwinska et al., 2013, Grabska-Barwińska et al., 2017], or a probabilistic formalization of memory retrieval where the latents are possible items retrieved from memory [Savin et al., 2011, 2014]. In these works, latents are either continuous [Hoyer and Hyvärinen, 2003, Grabska-Barwinska et al., 2013, Savin and Deneve, 2014, Orbán et al., 2016, Haefner et al., 2016, Bányai et al., 2019], e.g. representing the intensity of an odorant or the amplitude of a Gabor feature in the visual input, or discrete [Buesing et al., 2011, Savin et al., 2011, 2014, Haefner et al., 2016, Shivkumar et al., 2018], e.g. binary variables for different task contexts.

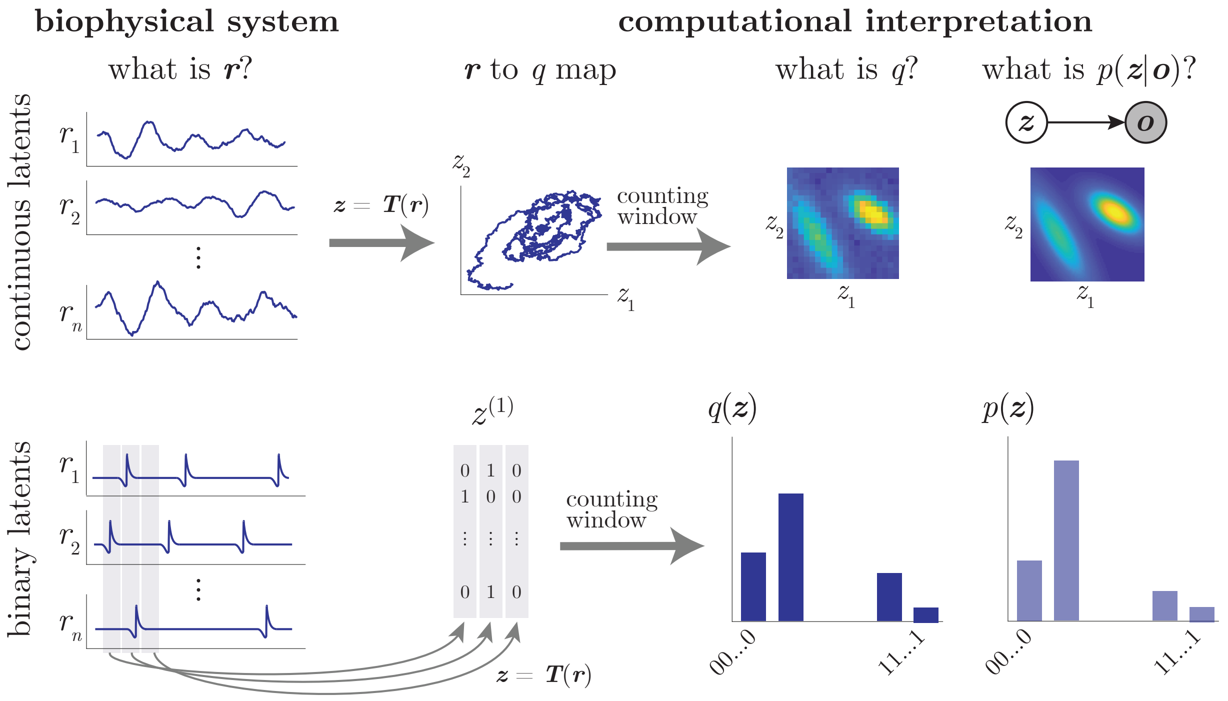

Figure 5: Illustration of neural sampling codes using both continuous and binary latents.

What is $\boldsymbol{o}$? The nature of the relevant observations depends on the generative model. They could be a retinal image or the whitened output of retina in most vision related studies, receptor activity in olfaction [Grabska-Barwinska et al., 2013], spikes in other brain regions [Beck et al., 2012], strength of synapses storing information about past experience in [Savin et al., 2011, 2014].

How it works: The details of the circuits implementing neural sampling can differ across proposals, but they generally take the form of stochastic dynamics that map a current sample into a new one, via a transition probability. For continuous latent variables, the simplest example is Langevin sampling where the dynamics perform gradient descent on an 'energy' (given by the negative log posterior), with additive gaussian noise:

$$\mathbf{z}^{(t+1)} = \mathbf{z}^{(t)} + \alpha \nabla \log p(\mathbf{z}|\boldsymbol{o}) + \boldsymbol{\epsilon}^{(t)} \tag{16}$$

where ν determines the step size and ϵ is zero-mean Gaussian noise whose variance only depends on α (Figure 5A). Another classic example algorithm is Gibbs sampling in which new samples are drawn one dimension of $\mathbf{z}$ at a time: $z_k^{(t+1)} \sim p(z_k|\mathbf{z}_{\neg k} = \mathbf{z}_{\neg k}^{(t)})$ where $\mathbf{z}_{\neg k}^{(t)})$ denotes the last sample, including all dimensions of $\mathbf{z}$ other than the kth (Figure 5B). Compellingly, the circuits resulting from Gibbs sampling only rely on local connectivity [Buesing et al., 2011, Haefner et al., 2016], matching classic anatomical results [Felleman and Van Essen, 1991]. The accuracy of sampling-based approximate inference critically depends on time, which makes sampling speed a key consideration when assessing the computational efficiency of different sampling schemes. Many of the recent theoretical efforts have been motivated by the conjecture that sampling in the brain is likely accelerated compared to classic sampling algorithms like Gibbs sampling, for instance through the use of balanced amplification [Hennequin et al., 2014, Echeveste et al., 2020], network oscillations [Savin et al., 2014, Aitchison and Lengyel, 2016], or other biophysical features [Buesing et al., 2011].

4.4 Formal connections between the different proposals

DDC and PPC are both categorized as parametric schemes, however, DDC is defined as an encoding framework, while PPC is primarily introduced as a decoding scheme [Lange et al., 2023]. Within the DDC framework, the posterior distribution $p(\mathbf{z}|\boldsymbol{o})$ is "encoded" by the expected values of the DDC encoding functions, determining the neuronal activity $\mathbf{r}$. In contrast, PPCs depart from defining the encoding scheme, and instead, assume that the logarithm of the posterior distribution can be approximately "decoded" through a weighted sum of a set of basis functions, where the weights correspond to the neuronal firing rates $\mathbf{r}$. This decoding-based scheme can be compared to the linear decoding approach in Anderson [1994], Anderson and Van Essen [1994], where it is assumed that the posterior distribution can be approximated through a weighted sum of given basis functions (similar to kernel density estimation). The distinguishing factor in PPC is the application of linear decoding to the "logarithm" of the probability distribution.

By narrowing our scope to a specific decoding approach for DDC, namely the maximum-entropy decoding, we can strengthen the connection between PPC and DDCs. In this case, the approximate posteriors in both DDC and PPC are members of the exponential family—PPC represents the natural parameters of the exponential family by a linear mapping of neural activity, while DDC encodes the expectations of sufficient statistics (mean parameters). Therefore, under the premise of maximum-entropy decoding, PPCs and DDCs can be transformed into one another through the one-to-one correspondence between natural parameters and mean parameters in exponential family distributions, though the conversion can be computationally expensive.

For the remainder of this subsection, we will assume that the underlying posterior (or approximate posterior) is a member of the exponential family of distributions with finite dimensional sufficient statistic $\mathbf{T}(\mathbf{z})$ and natural parameters $\boldsymbol{\eta}$. The distinction is that PPCs represent the natural parameters $\boldsymbol{\eta}$ linearly in neural activity, whereas DDCs represent the expectation parameters $\boldsymbol{\mu}$ linearly in neural activity. Because these two types of parameters can be uniquely related to each other, PPCs and DDCs are both equally expressive and can, in principle, be mapped onto each other. In fact, this conversion is a primary goal of many inference problems [Wainwright and Jordan, 2008], such as inferring marginals (directly related to expectation parameters) from a joint distribution specified by natural parameters. In practice, however, exact conversion is often difficult or even intractable, so probabilistic reasoning requires sophisticated approximation schemes or recognition networks that learn to approximate the relationship between $\boldsymbol{\mu}_{\mathbf{T}}$ and $\boldsymbol{\eta}$. Curiously, the principle means by which the relationship between natural parameters and expectations are discovered is via generating samples conditioned on the natural parameters and then using those samples to approximate the expectations.

It is worth noting, however, that many inference algorithms function by iteratively updating expectations and natural parameters, suggesting that both DDCs and PPCs can interact fruitfully to perform fundamental computations. For example, consider belief updating using variational inference with a factorized posterior, $q(\mathbf{z}_1, \mathbf{z}_2) = q(\mathbf{z}_1|\boldsymbol{\eta}_1)q(\mathbf{z}_2|\boldsymbol{\eta}_2)$. This algorithm uses iterative posterior updates that obey

$$\boldsymbol{\eta}_1 \cdot \mathbf{T}(\mathbf{z}_1) = \langle \log p(\mathbf{z}_1, \mathbf{z}_2) \rangle_{q(\mathbf{z}_2|\eta_2)} + \text{const}$$

where $p(\mathbf{z}_1, \mathbf{z}_2)$ is the target posterior to be approximated by q. When $\mathbf{T}(\mathbf{z})$ is expressed as a set of orthonormal basis functions, the joint distribution of observations and latents can be written as a linear combination of outer products and thus

$$\boldsymbol{\eta}_{1i} = \sum_j a_{ij} \langle T_j(\mathbf{z}_2) \rangle$$
$$\boldsymbol{\eta}_{2i} = \sum_j b_{ij} \langle T_j(\mathbf{z}_1) \rangle$$

A linear PPC assumes that the left hand sides of the above equations are linear in neural activity while a DDC assumes that the expectations on the right hand side are linear in neural activity. Thus, in this setting, a linear PPC representation for $q(\mathbf{z}_1)$ corresponds to a DDC representation of $q(\mathbf{z}_2)$ and vice versa since

$$M_1^{\text{PPC}} \mathbf{r}_1^{\text{PPC}} = \mathbf{A} \boldsymbol{M}_2^{\text{DDC}} \mathbf{r}_2^{\text{DDC}}$$
$$M_2^{\text{PPC}} \mathbf{r}_2^{\text{PPC}} = \mathbf{B} \boldsymbol{M}_1^{\text{DDC}} \mathbf{r}_1^{\text{DDC}}$$

More generally it can be shown that in a multilayer generative model implemented using a DDC, a PPC for latent variables in layer j, can be constructed from a quadratic combination of neural activity representing a DDC in layers $j-1$ and $j+1$. Indeed, the same quadratic combination of DDC representations is what drives learning in the DDC framework precisely because the quantities being learned parameters that are linearly related to natural parameters. As a result, learning signals in a network in which neural activity forms a DDC, the learning signals form are PPC representations of the corresponding posteriors.

Because they are both based upon exponential family distributions and associated approximate inference schemes, the difference between PPCs and DDCs comes down to the empirical question of whether the brain uses activity to linearly represent natural parameters (PPCs) or expectation parameters (DDCs). Computationally, these representational schemes differ in computational convienence and efficiency. For example, the product rule (e.g. evidence integration) is linear in log-probability or in natural parameters, which makes them easy to implement for linear PPCs, but is nonlinear in probability. Similarly, marginalization is a linear operation on probability and hence also linear for expectations so this operation is 'easy' for a DDC while the product rule requires a quadratic operation. Of course, this argument assumes the linear operations are in some sense preferred by neural circuits.

Inference generally requires both computations, and previous work has shown that a network capable of implementing a quadratic non-linearity and divisive normalization is sufficiently computationally expressive to implement both evidence integration and marginalization with a PPC [Beck et al., 2011]. In the Kalman filter, for example, all of the complexity of the standard equations are explained simply by the need to switch back and forth between a natural parameter representation to use the product rule of probability when updating posteriors with new evidence, and an expectation parameter representation to use the sum rule of probability when marginalizing over states. This switching suggests that, in the brain, one might expect to find signatures of both PPCs and DDCs at different stages or in different functional subpopulations.

Similar links exist between sampling and DDCs. For instance, for generative models based upon exponential family distributions, NSCs require the evaluation of the sufficient statistic $\mathbf{T}(\mathbf{z})$ or its gradient for each sample $\mathbf{z}$. Since a DDCs is present when the average of this quantity is available, a simple average of neural activity associated with an NSC leads to a DDC.

4.4.1 Special cases and alternative proposals

It is worth noting that the Free Energy Principle [Friston, 2010] is a special case of a parametric code in which neural activity represents the parameters of q. In its instantiation as predictive coding, it further assumes a mean-field approximation to the full posterior, i.e. $q(\mathbf{z}) = \prod_i q_i(z_i)$ in which each of the $q(z_i)$ is Gaussian [Gershman, 2019]. Such a representation is extremely limited in its expressive power compared to more general PPCs, DDCs, or NSCs, since it cannot represent any dependencies in the posterior p due the factorization.

Furthermore, PPCs, DDCs, and NSCs are not the only possibilities for probabilistic representations. For example, recent work in distributional reinforcement learning has proposed that the brain may use expectile codes [Dabney et al., 2020]. Such representations could themselves be constructed based on other probabilistic codes, such as sampling [Rullán Buxó and Savin, 2021].

It is also possible to interpolate between the model classes. For example, it is possible to sample the parameters of a distribution, rather than sampling the latent variables directly, like a sampled mixture of PPCs or DDCs [Lange et al., 2022] (Fig. 3**d–e**, Fig. 2**c**).

4.5 Case study: cue integration

Figure 7 shows a simple case study we will use to showcase computations needed for all three classes of theories. Here we use a probabilistic graphical model in which one top-level variable z_3 affects two lower-level latents z_1 and z_2, each generating its own observations o_1 and o_2 (Figure 7). We assume that tasks based on this model depend only on single latent variables, so the goal is to calculate marginal probabilities conditioned on all of the observations. In other words, the goal for inference in this model is to compute a representation of the high-level marginal posterior $p(\mathbf{z}_3|\boldsymbol{o}_1, \boldsymbol{o}_2)$, as well as the low-level posterior $p(\mathbf{z}_1|\boldsymbol{o}_1, \boldsymbol{o}_2)$ that accounts for both direct evidence from $\boldsymbol{o}_1 \to \mathbf{z}_1$ *and* indirect evidence $\boldsymbol{o}_2 \to \mathbf{z}_1$. Inference in this model is a nice case study because it requires use of both the product rule of probability, when combining the direct and indirect evidence, and the sum rule, when marginalizing over latent variables.

4.5.1 PPC

Integrating independent cues is straightforward in PPCs: this operation corresponds to the product rule for probabilities, so if neural activity is proportional to log probability, this means simply adding neural activity. We could do this directly if we had separate PPCs $\mathbf{r}^{(3|1)}$ and $\mathbf{r}^{(3|2)}$ for distributions over the higher-level variable $\mathbf{z}^3$, corresponding to indirect evidence $q(\mathbf{z}_3|\boldsymbol{o}_1)$ and $q(\mathbf{z}_3|\boldsymbol{o}_2)$. Then our final PPC for $q(\mathbf{z}_3|\boldsymbol{o}_1, \boldsymbol{o}_2)$ would be simply $\mathbf{r}^{(3)} = A\mathbf{r}^{(3|1)} + B\mathbf{r}^{(3|2)}$ for some matrices A and B.

However, our illustrative goal with this example inference problem is to define how representations would be combined along the representations of $\mathbf{z}_1$ and $\mathbf{z}_2$, which requires a transformation, rather than simply assuming a population code directly for the higher-level variable. In this case we must describe a marginalization over the nuisance variables that distinguish $\mathbf{z}_1$ and $\mathbf{z}_2$ from $\mathbf{z}_3$.

Marginalizing is more difficult in PPCs, because it is nonlinear in neural activity. Nonetheless, past work has shown that for Gaussian posteriors, quadratic nonlinearities with divisive normalization provide one good transformation [Beck

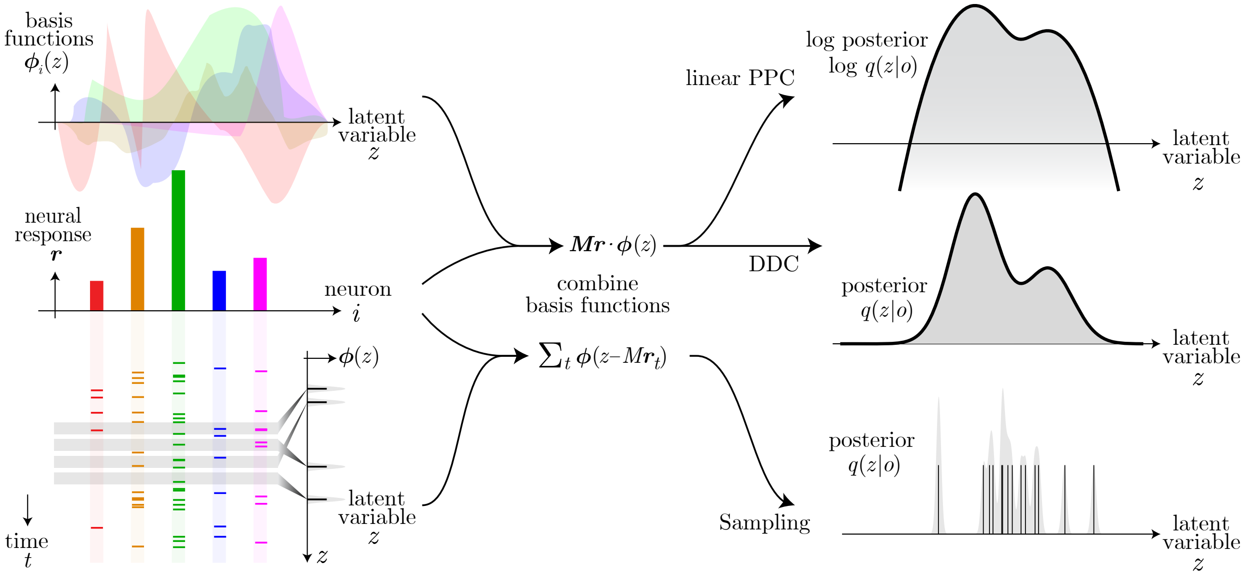

Figure 6: Illustration of the similarities and differences between linear PPCs, DDCs, and NSCs. In all three encoding schemes, the approximate posterior $q(z|o)$ can be written as a sum of basis functions $\phi_i(z)$. The essential difference between PPCs and DDCs on the one hand, and NSCs on the other, is that for PPCs & DDCs the neural responses $\mathbf{r}$ determine the *amplitude* of each basis function, while for NSCs the neural responses determine the *location* of each basis function. Another difference is that the number of basis functions for PPCs and DDCs is fixed, while for NSCs it is variable, increasing with time (one sample per time window). Note that while this illustration focuses on a scalar z, all three schemes also work for multidimensional $\mathbf{z}$. The basis functions for DDCs depend on an additional constraint like maximum entropy, similarly to the conversion of a sum of $\delta-$functions into a histogram for sampling. To maximize similarity between the schemes this illustration uses a distributed NSC scheme where samples are linearly read out from neural responses; most existing NSCs assume a 1-1 relationship between a single neuron's r_i and a corresponding z_i (also see Fig. 5).

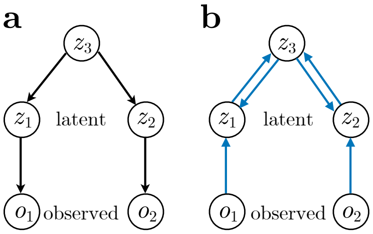

Figure 7: Illustration of probabilistic computations: cue integration: **(a)** Directed probabilistic graphical model. **(b)** Information flow during inference, i.e. the computation of $p(z_1, z_2, z_3|o_1, o_2)$.

et al., 2011], where quadratic operations and divisive normalization are both biologically plausible and well-described neural operations.

Putting these ideas together, the computation of $q(\mathbf{z}_3|\boldsymbol{o}_1, \boldsymbol{o}_2)$ for PPCs would start with populations $\mathbf{r}^{(1)}$ and $\mathbf{r}^{(2)}$, each encoding a posterior $q(\mathbf{z}_k|\boldsymbol{o}_k)$ over the lowel-level variables given the corresponding evidence. A third higher-level population $\mathbf{r}^{(3)}$ would then be driven by these lower ones as

$$r^{(3)}_\ell = \frac{\mathbf{r}^{(1)} A^\ell \mathbf{r}^{(1)} + \mathbf{r}^{(2)} B^\ell \mathbf{r}^{(2)}}{c^\ell + \boldsymbol{a}^\ell \mathbf{r}^{(1)} + \boldsymbol{b}^\ell \mathbf{r}^{(2)}} \tag{17}$$

where tensors A^ℓ_{ij} and B^ℓ_{ij} specify how each product of inputs $r^{(k)}_i$ and $r^{(k)}_j$ are weighted by neuron $r^{(3)}_\ell$, and vectors $\boldsymbol{a}^\ell_i$ and $\boldsymbol{b}^\ell_i$ specify how these inputs affect the divisive normalization. All of these weights are specified by the representation of the input population [Beck et al., 2011] and the coupling strength between the $\mathbf{z}$.

This describes the information flow from inputs to the representation of $q(\mathbf{z}_3|\boldsymbol{o}_1, \boldsymbol{o}_2)$. To condition the *lower*-level variable on *both* observations, a PPC representation would repeat this process to update the population $\mathbf{r}^{(1)}$, once again using quadratic operations and divisive normalization. This iterative updating defines a message-passing algorithm that converges to an equilibrium representation which reparameterizes the sufficient statistics of the joint distribution in terms of sufficient statistics for its marginals [Wainwright et al., 2003]. Since the neural activity in a PPC represents these statistics, the relevant dimensions of the neural activity also converges [Vasudeva Raju and Pitkow, 2016].

4.5.2 DDC

Unlike marginalization and chain inference, which can be easily implemented in the DDC framework, the cue combination needs elaborate computations [Sahani, 2021]. Various DDC-based methods have been proposed for implementing cue integration; here, we present a proportionality relationship for the DDC of the posterior distribution which is derived through the expansion technique described in 4.2. Let $r^{(1)}_i = \int p(z_1|o_1)\phi^{(1)}_i(z_1)dz_1$, $r^{(2)}_j = \int p(z_2|o_2)\phi^{(2)}_j(z_2)dz_2$, and $r^{(3)}_k = \int p(z_3|o_1, o_2)\phi^{(3)}_k(z_3)dz_3$ be the DDC values of the posterior distributions $p(z_1|o_1)$, $p(z_2|o_2)$, and $p(z_3|o_1, o_2)$, respectively. In DDC-based cue combination, ideally, we would like to compute the DDC values of the posterior $p(z_3|o_1, o_2)$, i.e. $r^{(3)}_k$, using the DDC values $r^{(1)}_i$ and $r^{(2)}_j$. It can be shown that $r^{(3)}_k \propto \int f_k(z_1, z_2)p(z_1|o_1)p(z_2|o_2)dz_1dz_2$, where $f_k(z_1, z_2) = \int \frac{p(z_3)p(z_1|z_3)p(z_2|z_3)}{p(z_1)p(z_2)}\phi^{(3)}_k(z_3)dz_3$. By approximating $f_k(z_1, z_2)$ through a bilinear expansion, $f_k(z_1, z_2) \approx \sum_{i,j} c_{i,j,k}\phi^{(1)}_i(z_1)\phi^{(2)}_j(z_2)$, we can find the proportionality relationship $r^{(3)}_k \propto \sum_{i,j} c_{i,j,k}r^{(1)}_i r^{(2)}_j$. Using a similar approach, we derive proportionality relationships for the marginal DDC values of z_1 and z_2. Let $\widehat{r}^{(1)}_k = \int p(z_1|o_1, o_2)\phi^{(1)}_k(z_1)dz_1$ denote the DDC of the marginal distribution $p(z_1|o_1, o_2)$. We can show that $\widehat{r}^{(1)}_k \propto \sum_{i,j} \widehat{c}_{i,j,k}r^{(1)}_i r^{(2)}_j$, where $\widehat{c}_{i,j,k}$ are the expansion coefficients of $g_k(z_1, z_2) = \phi^{(1)}_k(z_1) \int \frac{p(z_3)p(z_1|z_3)p(z_2|z_3)}{p(z_1)p(z_2)}dz_3$.

Note: In this implementation, the bilinear expansion is a challenging task. Moreover, the DDC values of $p(z_3|o_1, o_2)$ are merely proportional to the weighted sum of the product of DDC values $\mathbf{r}^{(1)}$ and $\mathbf{r}^{(2)}$, and the normalization factor is not easy to calculate. To deal with these limitations, other methods have been suggested for inference and learning in hierarchical models, such as DDC-Helmholtz machine with wake-sleep algorithm [Vértes and Sahani, 2018] (see section 4.2 for more information).

4.5.3 Neural sampling

Simple Langevin sampling from the posterior over latents $\mathbf{z}$ given observations $\boldsymbol{o}$ takes the form of stochastic dynamics of the form:

$$\dot{\mathbf{z}} = -\nabla \log p(\mathbf{z}|\boldsymbol{o}) + d\epsilon \tag{18}$$

where $d\epsilon$ is Brownian noise. Given the factorization of the posterior $p(\mathbf{z}|\boldsymbol{o}) = \frac{1}{Z}p(\boldsymbol{o}_1|\mathbf{z}_1)p(\boldsymbol{o}_2|\mathbf{z}_2)p(\mathbf{z}_1|z_3)p(\mathbf{z}_2|z_3)P(\mathbf{z}_3)$, with Z denoting the normalizing constant, the logarithm translates into dynamics that additively combine the contribution of each element. Assuming a one-to-one map between neurons and latent variables

$$\begin{aligned}
\dot{z}_1 &= -\log p(\boldsymbol{o}_1|z_1) - \log p(z_1|z_3) \to \dot{r}_1 = f_1(\boldsymbol{o}_1) + g_1(\mathbf{r}_3) && (19)\\
\dot{z}_2 &= -\log p(\boldsymbol{o}_2|z_2) - \log p(z_2|z_3) \to \dot{r}_2 = f_2(\boldsymbol{o}_2) + g_2(\mathbf{r}_3) && (20)\\
\dot{z}_3 &= -\log p(z_1|z_3) - \log p(z_2|z_3) - \log(\mathbf{r}_3) \to \dot{r}_3 = g_3(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) && (21)
\end{aligned}$$

These dynamics involve feedforward inputs $f_1(\cdot)$ to the neurons/subpopulations $\mathbf{r}_1$ and $\mathbf{r}_2$, respectively, with recurrent interactions $g_i(\cdot)$ structured by the information flow described in Fig. 7b. The exact form of these functions depends on the specifics of the graphical model, e.g. in the linear gaussian case feedforward input effects would be linear and the recurrent interactions would implement a stochastic linear dynamical system. More broadly, it is important to note that within the sampling framework inference is not naturally thought of as a transformation of representations but rather as a collection of recurrently interacting dynamical systems nodes which jointly represent one posterior distribution; marginals of that posteriors available by reading out information from single nodes in this system.

5 Interpretation of existing data

Distinguishing between different coding schemes using empirical data – both behavioral and neurophysiological – is complicated by the fact that the predictions generally depend not only on the coding scheme that relates a posterior probability to a neural response, but also on the generative model, $p(o, z)$, that is being used by a Bayesian brain. While it is possible to empirically evaluate such 'complete models', consisting of both an assumption about the $p(o, z)$ and the coding scheme, it is currently unclear whether there are empirical signatures that can distinguish between coding schemes irrespective of the assumed generative model. That is, prediction failures associated with any given coding scheme can be attributed to incorrect choice of generative model or, equivalently, incorrect assumptions about either the identity of the latent variables represented by a given population. In the next sections we will elucidate this fact by reviewing classic empirical observations and summarizing how they can be explained assuming different coding schemes, often involving different assumptions about the underlying generative models.

It is important to note that these theories only put constraints on, or make predictions for, a subset of the observable biophysical properties. For example, if the activity of only a subset of neurons represent posteriors, with other neurons performing auxiliary computations (e.g. as in [Pecevski et al., 2011, Aitchison and Lengyel, 2016, Echeveste et al., 2020]), then this will pose the extra empirical challenge of identifying those neurons. Similarly, if e.g. parameters of distributions represent low-dimensional projections of high dimensional neural activity, then neural activity in directions that are orthogonal to those projections will not be constrained. In general, this caveat is a special case of the general neural coding question that asks what aspect of neural activity is computationally relevant, often applied contrasting membrane potentials with spike times or firing rates [Dayan and Abbott, 2005].

The degeneracy that arises from the possibility of the different model components of probabilistic computations to trade off against each other suggests deeper theoretical work into 'equivalence classes' of different models that may all be compatible with the same biophysical system, yet involve different generative model – neural coding pairs, or pertaining to different aspects of neural activity (also see [Shivkumar et al., 2018, Lange et al., 2023]).

5.1 Tuning functions and their modulations

5.1.1 Tuning to a single stimulus dimension

When the average response of a neuron changes as a function of some variable, s, it is said to be 'tuned' to s. Typically, the considered variables are experimenter-defined, such as the orientation of a visual image on the retina, or frequency of an auditory stimulus. In the context of probabilistic inference, tuning arises when the neuronal response represents the posterior over latent variables $\mathbf{z}$ that depends on s, and that dependency changes the average response. In general, the tuning function is the consequence of both the coding scheme (how the response depends on the represented posterior), and how the internal variable $\mathbf{z}$ depends on the experimenter chosen variable s.

Each of these coding schemes predicts that neurons are tuned to the represented latents, $\mathbf{z}$. If s parameterizes a subspace of $\mathbf{z}$ then some of the neurons representing $p(\mathbf{z})$ will also be tuned to s. Moreover, if the latents are a deterministic function of s, i.e. $\mathbf{z} = f(s)$, then not only will the population of neurons be tuned to both $\mathbf{z}$ and s, but also the form of the probabilistic neural code (PPC/DDC/NSC) for $\mathbf{z}$ will be inherited by s. This suggests that complete knowledge of the relationship between represented latents $\mathbf{z}$ and laboratory variables s is not required for detecting the coding scheme.

5.1.2 Scaling of tuning curves with other stimulus parameters that influence uncertainty

Empirical observation: Two of the most-studied tuning curves, those to orientation in area V1, and those to motion direction in area MT, have been shown to be 'invariant' to the key stimulus aspects believed to influence the brain's uncertainty about the respective tuning variable: image contrast for orientation, and motion coherence for motion direction []. Invariant in this context means that the *shape* of the tuning curve is approximately invariant, and that changes in contrast and coherence have an approximately multiplicative effect on their magnitude across the entire stimulus range. From the perspective that sensory neurons 'represent' particular aspects of the input, this feature of the data appears paradoxical: while it is plausible that the neurons whose preferred stimulus is closest to the correct one would increase their firing with increasing certainty about the correct value, it is less clear why the same would be true for neurons representing stimulus values that have become less likely with increasing contrast or motion coherence [].

General probabilistic interpretation: In general, the shape and scaling of tuning curves with respect to some variable s will depend on the generative model defining the posterior over z, and the neural encoding scheme.

PPC interpretation: When visual contrast simply scales neural tuning over another feature such as orientation or motion direction, the PPC's logarithmic relationship between activity and posterior means that the posterior becomes narrower as the tuning amplitude rises. No study to date has investigated the tuning curves implied for a PPC with a generative model for e.g. natural images.

DDC interpretation: One of the key features of the DDC representation is the modulation of population sparsity by uncertainty. When the variance (uncertainty) of the posterior distribution increases, more DDC encoding functions overlap with the distribution and the sparsity of activity reduces. In other words, the diversity of neuronal activity increases with uncertainty (Fig.8) [Ujfalussy and Orbán, 2022].

It is also important to highlight the distinction between the DDC encoding functions and tuning functions. To compute the tuning function of a neuron, the experimenter sweeps over the parameter of interest $\boldsymbol{s}$ (e.g. the orientation of the grating) and measures the firing rate of the neuron. The value of the tuning function of a DDC neuron at $\boldsymbol{s}$ is equal to the expected value of the DDC encoding function $\phi(\mathbf{z})$ with respect to the posterior distribution $p(\mathbf{z}|\boldsymbol{o})$, where the sensory observation $\boldsymbol{o}$ is a function of $\boldsymbol{s}$.

The DDC framework suggests that decreasing the uncertainty of the posterior distribution should narrow the tuning functions of individual neurons. However, it has been observed that in V1, the tuning functions over the orientation are contrast-invariant, and decreasing the contrast does not broaden the tuning. Nevertheless, we should note that the generative model and the latent variables over which the DDC is defined play a critical role in this analysis; for example a multiplicative contrast term in the generative model results in a different posterior over the latent variables which correspond to the coefficients of the Gabor basis functions. This area warrants a more comprehensive investigation, and it is of great importance to study the impact of different generative models (e.g. for natural images) on the modulation of tuning functions in the DDC framework.

Sampling interpretation: Two principal generative models for natural images have been shown to produce tuning curves to orientation (as well as other dimensions like spatial frequency) that approximately scale with contrast: Gaussian scale-mixture models under the assumption that latents are represented by membrane potentials [Orbán et al., 2016], and linear Gaussian models with binary latents represented by spikes (Chattoraj et al. COSYNE 2016) [Shivkumar et al., 2018].

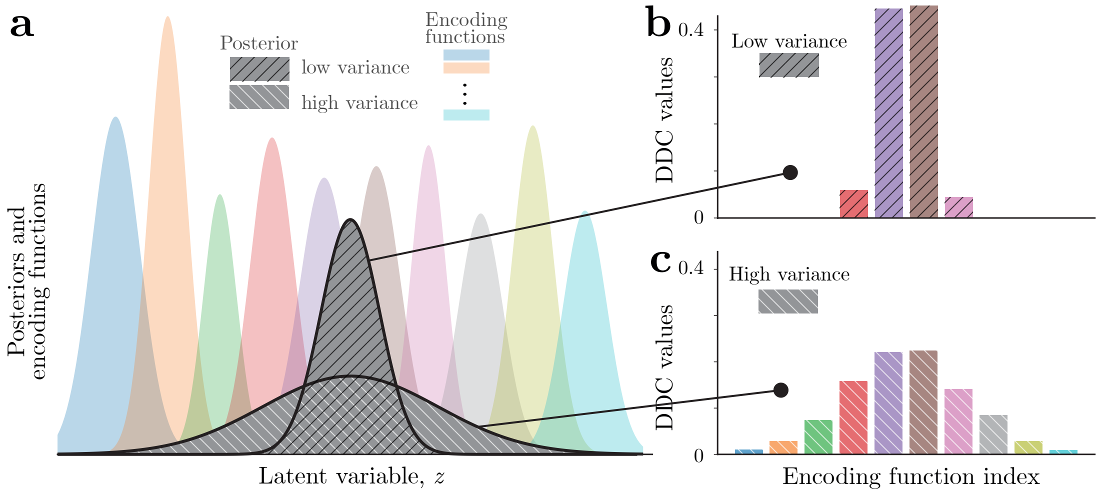

Figure 8: DDC sparsity inversely correlates with posterior variance. (**a**) DDC encoding functions and two posterior distributions with low and high variance. (**b**) DDC values of the posterior distribution with low variance. (**c**) DDC values of the posterior distribution with high variance.

5.2 Neural variability

Both membrane potentials and spiking responses of sensory neurons are known to be variable, even to repeated presentations to the same external stimulus [Tolhurst et al., 1983, de Ruyter van Steveninck et al., 1997]. Regardless of whether this variability is pure noise that is corrupting the underlying signal, or serves a computational function, it may be helpful for distinguishing between the different proposals described above.

5.2.1 Empirical observations

Neural variance Under natural conditions, neural responses are highly variable [Tolhurst et al., 1983]. This variability is different in different systems: in the retina, variability in spike counts is roughly as low as possible [Berry et al., 1997]; in the early auditory system, sound direction is computed using enormous and extremely reliable synapses [Joris and Trussell, 2018]; but in the cortex, variance is higher, and is observed to grow roughly proportional to the mean [Tolhurst et al., 1983, de Ruyter van Steveninck et al., 1997]. Indeed, the Fano factor, the ratio of the stimulus conditioned variance to the mean, is approximately one for most neurons in sensory cortical areas [Rieke et al., 1999]. This relationship is maintained across a wide range of stimulus contrast levels (a proxy for input information in visual tasks) so that higher contrast is associated with higher firing rates and more variability. There also appears to be an increase in variability from lower level to higher level areas [Kara et al., 2000] with at least some of that increase accounted for by variability shared among neurons [Goris et al., 2014]. In the absence of a sensory stimulus neural variability is higher and exhibits a marked decrease at the time of stimulus stimulus onset [Churchland et al., 2010] modulated by contrast in early visual areas. These findings apply both to membrane potentials and firing rate [Finn et al., 2007, Savin and Deneve, 2014, Orbán et al., 2016, Hennequin et al., 2018].

Neural covariance Neural responses often covary across neurons. The precise pattern of covariability may be important for their information content and computational function. Variation can be decomposed into different aspects: we expect some variation simply because the external input varies. This is often called 'signal correlation'. Variability across repeated presentations of identical sensory inputs is called 'noise correlations' [Gawne and Richmond, 1993, Averbeck and Lee, 2006]. Here, the scare quotes to serve as a reminder that noise may be a misnomer as these fluctuations could be a result of useful computational mechanisms, fluctuations in attention, or context effects that were not held fixed during the experiment. Indeed, much of the interest in neural covariation focuses on noise correlations precisely because they reflect the internal computations and structure of the neural code rather than merely the external drive to a circuit [Savin and Deneve, 2014, Lange and Haefner, 2017, Ruff et al., 2018].

Empirically, the noise covariance between two neurons typically grows with their mean responses, just like the variance for cortical neurons. The covariance reflects both the relationship between the neurons and scale of each neuron's relationship. The correlation coefficient is one way to approximately isolate relationship between the neurons from the modulation of each neuron separately (although see [De La Rocha et al., 2007, Pitkow and Meister, 2012]). The pattern of correlations is often related to the neural tunings: similarly tuned neurons often exhibit greater correlations [Zohary et al., 1994, Averbeck and Lee, 2003, Nardin et al., 2023], although there is substantial diversity around this pattern. Furthemore, some studies report changes in correlation with the stimulus [Ohiorhenuan et al., 2010, Ponce-Alvarez et al., 2013], motor outputs [Dadarlat and Stryker, 2017], brain state [Ecker et al., 2014], and attention [Cohen and Maunsell, 2009].

Noise covariance affects the information content that can be optimally extracted from a neuronal population [Zohary et al., 1994, Abbott and Dayan, 1999, Averbeck et al., 2006]. When signal covariance and noise covariance are similar, as observed in cortex, the resulting neural code is highly redundant with respect to the stimulus of interest[Moreno-Bote et al., 2014, Ecker et al., 2014]. While this redundancy appears limiting, it may serve a computational purpose (e.g. [Lange and Haefner, 2022, Nardin et al., 2023, Haimerl et al., 2023]).

5.2.2 Interpretation

There are two principal ways in which neural variability can arise among neural responses encoding a posterior belief: variability in the posterior itself, and a stochastic encoding of a fixed posterior [Lange and Haefner, 2022]. Since the latter is shared by all encoding schemes, we will discuss it first. Trial-by-trial variability in the posterior can arise even when the experimenter-controlled stimulus is kept constant as the result of external and internal sources of variability that can be mapped on to different aspects of computing a specific posterior: variability in the observation process (e.g. small eye-movements), variability in the likelihood computations, and variability in representation of prior expectations. Likelihoods and priors may be variable due to an approximate model of the outside world [Beck et al., 2012], approximate computations (e.g. implementations in stochastic neural circuits, or computations that only converge asymptotically).

Under some circumstances, e.g. in the context of learning a behavioral task, it has been possible to characterize the nature of the neural covariability due to the variability in the posterior, and to derive empirical predictions that match that reported in existing studies [Lange and Haefner, 2022].

Interpretation of neural variability by PPCs Unlike sampling based codes, neural variability does not play a computational role in PPCs. However, PPCs do make strong predictions for the relationship between tuning curves and the covariance structures (Eq. **??**) in the presence of nuisance parameters. While there are many ways to satisfy this relationship, one oft cited way is when Fano Factors are constant (but not necessarily one) for all values of the stimulus and nuisance parameters. As a result, the ubiquity of Fano factors near one across all stimulus conditions is often cited as evidence in favor of PPCs despite the fact that variable Fano factors can also be consistent with PPCs.

Interpretation of neural variability by DDCs Like for PPCs, in DDC, variability does not serve a computational function but is considered noise that contaminates the signal present in the firing rates. As noted in 4.2, various noise models can be incorporated to model neural variability in the DDC framework, including additive noise (with independent or correlated components), and Poisson noise Zemel et al. [1998].

Interpretation of neural variability by NSCs Neural variability is a necessary consequence of sampling-based inference: neural responses are variable since they are directly related to samples, which vary stochastically over time, with the amount of variability directly related to the uncertainty in the underlying beliefs.

For models based on continuous latent variables, this view predicts a dissociation of mean and variance reflecting underlying beliefs that can change in both mean and uncertainty. Importantly, how both depend on external stimulus parameters depends critically on the assumed generative model, $p(\boldsymbol{o}|\mathbf{z})$. For instance, [Orbán et al., 2016, Festa et al., 2021] found that the stimulus-dependence of spiking responses and membrane potentials of V1 neurons were compatible with the assumption that membrane potentials represent samples from mixture variables in a Gaussian scale mixture model.

For models based on binary latent variables, mean and variance are more tightly coupled since the variance of a binary variable is simply $p(1-p)$ where p is the probability of the variable being 1. For small p, and summing over many independent samples, the distribution over the count is approximately Poisson, suggesting that individual spikes (and absences of spikes) may be interpretable as samples from distributions over binary $\mathbf{z}$ [Buesing et al., 2011, Shivkumar et al., 2018]. While the Fano Factor of independent binary samples is $1-p$, i.e. sub-Poisson, samples that are generated using an MCMC algorithm often have positive autocorrelations. This will increase the variability of the number of spikes counted over an extended time window, making the resulting variability compatible with empirical observations from cortex *independent* of the specific nature of $\mathbf{z}$ and $p(\boldsymbol{o}|\mathbf{z})$.

Neural sampling further predicts that whenever the posterior over different latent variables, $p(z_1, z_2|\boldsymbol{o})$ is correlated, then this dependency in the posterior should directly be expressed in neural co-variability between the neurons representing z_1 and z_2, respectively. Starting from this insight, several studies have derived concrete predictions for noise correlations and choice correlations in the presence and absence of behavioral tasks [Haefner et al., 2016, Bondy et al., 2018, Bányai et al., 2019]. However, it is important to note that the trial-by-trial variability in the posterior (e.g. due to input noise and approximations as noted above) is closely related to the shape of the posterior itself. As a result, at least qualitatively, these predictions are shared by any coding scheme [**?**] and, without a more quantitative analysis (see e.g. [Ujfalussy and Orbán, 2022]), cannot be taken as direct evidence for the neural sampling hypothesis. Finally, the alignment of stimulus and noise correlations arises either as a consequence of approximate inference or due to a separate encoding process, e.g. in distributed sampling [Savin and Deneve, 2014].

5.3 Relationship between spontaneous and evoked neural activity

5.3.1 Empirical observations

Neurons in sensory cortex are active even in the absence of external inputs [Faisal et al., 2008]. This spontaneous activity does not appear random but instead appears to be structured similar to activity evoked by external sensory inputs [Tsodyks et al., 1999, Kenet et al., 2003, Fiser et al., 2004, Luczak et al., 2009]. Furthermore, the statistical structures of the spontaneous and the evoked activity appear to converge over the course of development [Berkes et al., 2011], providing a constraint on models of probabilistic computations in the brain (but see [Avitan and Stringer, 2022]).

Interpretation of the relationship between spontaneous and evoked activity in NSCs Under the assumption that complete darkness is interpreted by the brain's internal model as the absence of information about the internally represented variable, $\mathbf{z}$, the likelihood, $p(\boldsymbol{o}|\mathbf{z})$ is flat, and the posterior, $p(\mathbf{z}|\boldsymbol{o}) \propto p(\boldsymbol{o}|vz)p(\mathbf{z})$, should equal the prior. If neural activity represents the posterior, spontaneous activity should therefore reflect the brain's prior. Furthermore, inference in a well-calibrated generative model requires that the average posterior matches the prior, $p(\mathbf{z}) = \int p(\mathbf{z}|\boldsymbol{o})p(\boldsymbol{o})\mathrm{d}\boldsymbol{o}$. Under the assumption that neural activity represents samples from the posterior, the distribution over spontaneous activities is therefore predicted to match the distribution over activities evoked by natural stimuli, $\boldsymbol{o}$, provided that they are presented in proportion to their natural occurrence, $p(\boldsymbol{o})$ – as tested and confirmed by [Berkes et al., 2011].

Interpretation of the relationship between spontaneous and evoked activity in DDCs Since DDCs propose that neural responses represent extended moments of the posterior distribution, and therefore are linear functions of the posterior ('linear distributional codes' [Lange and Haefner, 2022]), the calibration argument described above predicts that average spontaneous activity equals average evoked activity. If moments are represented by average neural activity, i.e. firing rates, then this implies that the spontaneous firing rate should equal the average evoked firing rate. Note that this prediction is only a special case of the test in [Berkes et al., 2011] who found that the *distribution* over responses (spikes) equalled the *distribution* over evoked responses to natural stimuli. Assuming the DDC representation has, for instance, Poisson variability around the firing rate, this will not in general be compatible with the [Berkes et al., 2011] observation.

Interpretation of the relationship between spontaneous and evoked activity in PPCs For PPCs, as with DDCs and NSCs, spontaneous activity is assumed to represent the prior distribution over the relevant latent variables. However, due to PPCs' nonlinear relationship with the encoded distribution, average evoked activity is not expected to be related to spontaneous activity.

Energy efficient PPCs prefer to represent priors with low levels of neural activity to conserve spikes when there is no information to report. This static consideration, however, is overly simplistic, as it fails to take into account the dynamics needed to implement inference. PPCs can straightforwardly be adapted to implement predictive coding algorithms, further increasing efficiency when performing inference on hierarchical generative models. In this case, populations of neurons represent residual likelihoods rather than posterior distributions. Associated patterns of activity that represent these probabilistic error signals exhibit greater transient variability in the presence of noise generated by feed forward and feedback connections. Moreover, dynamics designed to implement probabilistic reasoning tend to turn that variability into patterns of activity that qualitatively look like patterns of activity associated with posterior distributions. Thus, while there is not necessarily a relationship between spontaneous and evoked activity in a PPC, there are many dynamical systems utilizing PPCs that exhibit strong transient patterns of activity that resemblance evoked activity [Grabska-Barwinska et al., 2013].

5.4 Oscillations

Oscillations are a ubiquituous feature of cortical activity [Buzsaki and Draguhn, 2004]; however it is currently unknown to what they can constrain the neural implementation of probabilistice inference. On one hand they are predicted by fast algorithms implementing neural sampling using non-normal dynamics in models of hippocampus [Savin et al., 2014] and sensory cortex[Aitchison and Lengyel, 2016, Echeveste et al., 2020]. On the other hand, it has been recently shown that by modeling the hippocampal formation as a Helmholtz machine, theta oscillations can be used to mediate the wake-sleep algorithm [George et al., 2024]. As a result, it can be suggested that the implementation of the DDC framework through the wake-sleep algorithm in a Helmholtz machine might be compatible with neural oscillations; further analysis is required to investigate this scheme. No studies currently exist on their compatibility with PPCs.

5.5 Neural–behavioral correlations

Many experiments have measured behaviors that accord with behaviors based on probabilistic inference [Fiser et al., 2010, Pouget et al., 2013]. Of course this does not explain how the brain accomplishes this. To understand the neural basis of such behaviors, neuroscientists have examined whether neural representations of sensory uncertainty are related to actions. This is often accomplished by directly examining the relationship between neural activity and behavior.

5.5.1 Empirical

One common approach to measuring the relationship between neural activity and behavior is measuring the correlation between choices and individual neural responses. This correlation is known as choice probabilities [Britten et al., 1996, Haefner et al., 2013] or choice correlations [Pitkow et al., 2015, Clery et al., 2017, Yang et al., 2021, Chicharro et al., 2021]. A second approach is to predict (decode) behavior from the activity of a neural population (e.g. [Mante et al., 2013] and many others). A more sophisticated variant of this approach is to decode from the neurons targeted latent properties hypothesized to be important for behavior, and then ascertain whether those decoded quantities predict behavior [Shahidi et al., 2019, Wu et al., 2020]. A recent paper used this approach to decode uncertainty about the stimulus from trial-by-trial fluctuations in neural activity, and found that this decoded uncertainty predicted shifts in decision criteria [Walker et al., 2020].

5.5.2 Interpretations

As explained in the variability section above, trial-to-trial variability in the observations will induce variability in the posterior over $\mathbf{z}$ and hence behavior, inducing correlations between neural representations of $p(\mathbf{z}|\boldsymbol{o})$ and behavior – regardless of the neural coding scheme. Interestingly, after learning, the structure of these correlations is approximately the same regardless of neural code [Lange and Haefner, 2022]. As a consequence, work showing agreement between empirical correlations [Haefner et al., 2016, Bondy et al., 2018], or decoded latents Walker et al. [2020], and predictions of probabilistic inference models based on particular codes, can only be taken as evidence in favor of the representation of posteriors rather than any one particular coding scheme over another.

5.6 Behavioral data

There are at least two ways in which behavioral data can constrain models of probabilistic computations in the brain. First, if behavior is close to optimal, then this places a constraint on the brain's internal model insofar as that the task model must be a special case of the brain's internal model. Second, any deviations in behavior from optimality place constraints either on the brain's internal model and/or the approximate inference algorithm. Plausible deviations of the internal model may result from being adapted to natural inputs as opposed to task-specific ones, or due to incomplete learning of the task. In those situations, Bayesian inference makes predictions about the direction of behavioral change in e.g. perceptual learning paradigms. Alternatively, even if the internal model is correct, the specific approximate inference algorithm employed by the brain will lead to deviations from optimal behavior. Such deviations will generally depend on the specific algorithm, and thereby observed behavior placing constraints on which algorithm is employed by the brain.

To test theories of probabilistic computations in the brain using neural data requires the specification of the link between computational quantities like samples or parameters, and neural responses (see 2.7). In analogy, testing the same theories using behavioral data requires a link between posteriors and actions. While much work exists on the nature of this link (e.g. [Kording, 2007]), we consider this beyond the scope of this manuscript, and will only briefly describe selected attempts in the hope to encourage further research in that direction and describe caveats.

5.6.1 Behavioral bias

As an example, Haefner et al. found that hierarchical inference by sampling in the context of a sequential evidence integration task led to an overweighting of evidence presented early in a trial[Haefner et al., 2016] . A follow-up study elaborating on this initial finding discovered that rather than being specific to sampling, this bias was also predicted by a variational inference model based on a parametric representation [Lange et al., 2021]. By interpolating between two related but different tasks, and by explaining data across both tasks with the same generative model structure, and the same inference algorithm, this study uses generalization across tasks as a way to address a general critique of probabilistic approaches: it is possible to explain any behavior as optimal inference on some generative model. However, this can be seen as a form of overfitting to a task, yielding a generative model that may not generalize to other tasks.

5.6.2 Behavioral variability

Just as for neural responses, variable behavior may arise as the result of variability in the posterior due to uncontrolled variability in the observations, or variability in the neural encoding or computation of the posterior [Drugowitsch et al., 2016, Lange et al., 2021, Shivkumar et al., 2022]. Furthermore, the variability in the posterior itself may be magnified by a mismatch between the internal model and the model generating the observations [Beck et al., 2012].

Much like neural variability, human and nonhuman primate perception has been shown to be variable, even for constant external inputs. Examples are images with ambiguous interpretation (e.g. the vase/face image), and dichoptic stimuli that induce perceptual switching between the image shown to the left and the right eye [Blake and Logothetis, 2002]. [Gershman et al., 2012] showed that sampling from a probabilistic model of bistable inputs implied a distribution over dominance times that qualitatively matched empirically observed distributions. [Moreno-Bote et al., 2011] showed that under the assumption that the posterior is represented as a linear PPC, an attractor network could generate samples from the posteriors that matched empirically observed distributions.

5.7 Does the entirety of the considered empirical data favor one of the neural codes?

Overall, most of the observations considered above can qualitatively be explained by all of the proposals. The only exception may the observation that spontaneous activity is very similar to the average evoked activity: while this is directly predicted by neural sampling, it seems significantly harder to explain by DDCs and PPCs.

6 Recommendations for future studies

Testing for relationships between the neural representations for different posteriors: While systems neuroscience has traditionally focused on stimulus-response relationships, focusing on the relationships between the responses holds the promise of empirical tests of the different coding schemes that do not require explicit knowledge of the $\mathbf{z}$ that is being represented. A key example of this approach exploits the calibration property of probabilistic models: that the average posterior should equal the prior. This was tested by [Berkes et al., 2011] who compared spontaneous activity in area V1 to average evoked activity and found that they are statistically indistinguishable in mature ferrets. This result is predicted for any system that represents posterior beliefs under the assumption that placing the ferrets in a completely dark room is interpreted as 'no input' by the visual system. Recent work [Lengyel et al., 2023] has shown how to exploit a linearity property obeyed by some neural codes (DDCs and NSCs), but not all (e.g. PPCs), to test whether neural responses to different stimuli are linearly related to each other in a way suggested by the underlying posteriors to those stimuli. Such an approach may allow for the development of a method to compare brains and probabilistic models that is akin to representational similarity analysis (RSA [Kriegeskorte et al., 2008]) that allows for the empirical test of generative models and neural codes requiring weaker assumptions than currently needed by explicating a 'complete' model.

Generalization across multiple experiments and datasets: Given many possible choices for latent variables $\mathbf{z}$, the generative model linking them to experimental variables $p(\mathbf{z}, \boldsymbol{o})$, $\boldsymbol{s}$, and the concrete mapping between the posterior and the neural responses $\mathbf{r}$, no single experimental dataset will be enough to constrain all of these degrees of freedom. However, we can make progress by focusing on the generalization properties of the probabilistic model. For data fitting to go beyond 'Bayesian just so' storytelling [Bowers and Davis, 2012], the same set of model choices should be able to explain *all* aspects of the data, not just a subset (e.g. tuning functions, their changes with uncertainty, or a given feature of response covariability, structure of spontaneous response covariability, behavioral response changes with uncertainty, etc). Testing for consistency of data-constrained probabilistic quantities across tasks (prior, latents, mapping to neural activity) thus seems to be a productive approach to validate such hypotheses, akin to behavioral-level attempts at constraining probabilistic descriptions of perception at the behavioral level [Maloney, 2002, Houlsby et al., 2013].

Develop a quantitative benchmark for the comparison of probabilistic models: Another promising direction is to design a benchmark consisting of all relevant aspects of neural activity in one cortical area (e.g. membrane potentials, spike times, and spike rates), for a specific set of stimuli and behavioral contexts. This would facilitate fair comparisons when comparing different models and encoding schemes, and may accelerate progress for the same reasons benchmarks have been helpful in machine learning. Recent work has taken a step in that direction by directly fitting a flexibly parameterized generative model to neural responses to natural images under the assumption of a neural sampling code [Shrinivasan et al., 2024].

Causal manipulations: The strongest way to test our understanding of the computations performed by a neural circuit is to causally manipulate that circuit. This allows us to directly attribute computational or behavioral consequences to the manipulated property. Modern neuroscience methods afford multiple ways of performing such causal interventions, including electrical, pharmacological, or optogenetic manipulations. As the controllable spatial and temporal resolution of these manipulations increases, our interventions can be more targeted to specific activity patterns of interest. However, even coarse interventions may provide illuminating tests for discriminating between probabilistic coding schemes.

For instance, cooling or optically inactivating an area implies that the marginal belief about the corresponding variable z_1 is either completely confident that the latent variable is zero, $p(z_1) = \delta(z_1)$ (for NSCs), or uninformative with $p(z_1) = \text{const}$ (for PPCs). As a result, NSCs predict that the neural variability in other areas representing z_2 should be decreased (NSCs), while PPCs predict that the mean activity should be reduced since $p(z_2)$ will be less certain.

Interactions across brain areas: We have emphasized the importance of computations because representations do not stand on their own. Their value lies in their use. Decades of past evidence has demonstrated that brain areas exhibit both some specialization and some hierarchical structure. Consequently, we expect that representations of different latent variables in different brain areas will influence each other in predictable ways. Some of this may be discernible through observing existing correlations [Semedo et al., 2019]. However, causal manipulations including inactivation [Lakshminarasimhan et al., 2018], noise injection, or patterned perturbation [Chettih and Harvey, 2019, Adesnik and Abdeladim, 2021] are valuable in distinguishing direct interactions from indirect ones or from common causes [Lakshminarasimhan et al., 2018, Das and Fiete, 2020]. For all codes, only a subspace of neural activity may dominate the encoding of the parameters of the encoded probability distribution, so only perturbations that affect those dimensions should affect computations in other brain areas. In PPCs and DDCs, these are dimensions that project onto the sufficient statistics of the posterior. In sampling, these are dimensions that contribute to fluctuations in latent variables. For example, if multiple neurons represent the same variable, such that their population mean represents the sample, then population dimensions that increase some neurons' firing while decreasing others' will have no effect. Experiments that measure and then manipulate (or track noise in) these dimensions will allow the testing of whether these dimensions affect downstream computation in the manner predicted by each theory.

Theoretical work on similarities and differences between the coding schemes: Theoretical work may find that these models are formally equivalent under some conditions. For example, recall that all of these models make predictions that depend critically on committing to assumptions of a generative model, including what latent variables $\mathbf{z}$ the distribution is over, what properties of neural activity $\mathbf{r}$ encode them, and how these are connected to observations $\boldsymbol{o}$. For example, [Lange et al., 2023] showed that sampling over basis function amplitudes can manifest as a PPC over orientations on a coarser timescale. Different choices of experimental variables $\boldsymbol{s}$ to probe possible latent variables $\mathbf{z}$ may also lead to indistinguishable conclusions about the brain's confidence when these two types of variables are highly informative about each other. Future studies may find that such relations hold more generally, underscoring the importance of defining and comparing the key assumptions from which predictions are derived. Deeper insights on the similarities and equivalences of different coding schemes will also yield a better understanding of the predictions on which these schemes actually disagree, and the kind of data and experiments that may be able to definitely distinguish between them.

Theoretical work on general framework which contains specific coding schemes as special cases: Following existing work, this paper treats different coding schemes as alternative hypothesis. However, it might be more productive to conceived them as special cases of a more general coding scheme. For instance, Lange et al. described a space in which variational and sampling-based inference represent two extreme points along a continuum of inference algorithms some of which may be a closer description of the brain's encoding scheme than either of the extremes.

6.1 Conclusion

There is growing empirical evidence that Bayesian inference, and Bayesian decision theory, is a useful framework for understanding human behavior. On the basis of this it is tempting to view neural activity through the same lens. However, the jury is still out whether this is a fruitful approach, and whether the Bayesian framework has predictive power for neural activity. For instance, outputs consistent with probabilistic computations emerge generically when a sufficiently flexible computational system is trained in a world where probabilistic inference is the best way to solve problems [Ramsey, 1926, Orhan and Ma, 2017]. However, whether the brain's implementation can be mapped onto Bayesian concepts like priors, likelihoods, posteriors, loss functions, variational parameters, moments, or samples, is less obvious. But if it can, and if these mappings generalize across sensory inputs and behavioral tasks – an empirical question – then this would greatly advance our understanding of the brain linking the computational level with the implementation level via the algorithmic level.

Acknowledgements

We would like to thank the participants of the CCN/GAC Kick-off workshop, in particular the presenters Alex Pouget, Rajesh Rao, Maneesh Sahani, Eero Simoncelli, Eszter Vertes, Balazs Ujfalussy, Emin Orhan, and Richard Lange. We also thank Richard Lange, Gergő Orban and Suhas Shrinivasan for comments on the manuscript, and Màtè Lengyel for helpful discussions. This work was financially supported by NIH U19NS118246 (RMH,XP), NSF/CAREER IIS-2143440 (RMH), Gatsby Charitable Foundation, Simons Foundation (SCGB 543039), and NSF Award No. 1922658 (CS).

References

Larry F Abbott and Peter Dayan. The effect of correlated variability on the accuracy of a population code. *Neural computation*, 11(1):91–101, 1999.

Hillel Adesnik and Lamiae Abdeladim. Probing neural codes with two-photon holographic optogenetics. *Nature neuroscience*, 24(10):1356–1366, 2021.

Laurence Aitchison and Máté Lengyel. The hamiltonian brain: Efficient probabilistic inference with excitatory-inhibitory neural circuit dynamics. *PLoS computational biology*, 12(12):e1005186, 2016.

Charles H Anderson. Basic elements of biological computational systems. *International Journal of Modern Physics C*, 5(02):313–315, 1994.

Charles H Anderson and David C Van Essen. Neurobiological computational systems. *Computational intelligence imitating life*, 213222, 1994.

Bruno B Averbeck and Daeyeol Lee. Neural noise and movement-related codes in the macaque supplementary motor area. *Journal of Neuroscience*, 23(20):7630–7641, 2003.

Bruno B Averbeck and Daeyeol Lee. Effects of noise correlations on information encoding and decoding. *Journal of neurophysiology*, 95(6):3633–3644, 2006.

Bruno B Averbeck, Peter E Latham, and Alexandre Pouget. Neural correlations, population coding and computation. *Nature reviews neuroscience*, 7(5):358–366, 2006.

Lilach Avitan and Carsen Stringer. Not so spontaneous: Multi-dimensional representations of behaviors and context in sensory areas. *Neuron*, 110(19):3064–3075, 2022.

Ben Baker, Richard Lange, Alessandro Achille, Rosa Cao, Nikolaus Kriegeskorte, Odelia Schwartz, and Xaq Pitkow. What makes representations "useful"? 2021a.

Ben Baker, Benjamin Lansdell, and Konrad Kording. A philosophical understanding of representation for neuroscience. *arXiv preprint arXiv:2102.06592*, 2021b.

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, Tim Hanks, Anne K Churchland, Jamie Roitman, Michael N Shadlen, Peter E Latham, and Alexandre Pouget. Probabilistic population codes for bayesian decision making. *Neuron*, 60(6): 1142–1152, 2008.

Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Marginalization in neural circuits with divisive normalization. *Journal of Neuroscience*, 31(43):15310–15319, 2011.

Jeffrey M Beck, Wei Ji Ma, Xaq Pitkow, Peter E Latham, and Alexandre Pouget. Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron*, 74(1):30–39, 2012.

David Beniaguev, Idan Segev, and Michael London. Single cortical neurons as deep artificial neural networks. *Neuron*, 109(17):2727–2739, 2021.

Pietro Berkes, Gergő Orbán, Máté Lengyel, and József Fiser. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331(6013):83–87, 2011.

Michael J Berry, David K Warland, and Markus Meister. The structure and precision of retinal spike trains. *Proceedings of the National Academy of Sciences*, 94(10):5411–5416, 1997.

Randolph Blake and Nikos K Logothetis. Visual competition. *Nature Reviews Neuroscience*, 3(1):13–21, 2002.

Adrian G Bondy, Ralf M Haefner, and Bruce G Cumming. Feedback determines the structure of correlated variability in primary visual cortex. *Nature neuroscience*, 21(4):598–606, 2018.

Jörg Bornschein, Marc Henniges, and Jörg Lücke. Are v1 simple cells optimized for visual occlusions? a comparative study. *PLoS computational biology*, 9(6):e1003062, 2013.

Jeffrey S Bowers and Colin J Davis. Bayesian just-so stories in psychology and neuroscience. *Psychological bulletin*, 138(3):389, 2012.

Kenneth H Britten, William T Newsome, Michael N Shadlen, Simona Celebrini, and J Anthony Movshon. A relationship between behavioral choice and the visual responses of neurons in macaque mt. *Visual neuroscience*, 13(1):87–100, 1996.

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 Comput Biol*, 7(11):e1002211, 2011.

Gyorgy Buzsaki and Andreas Draguhn. Neuronal oscillations in cortical networks. *science*, 304(5679):1926–1929, 2004.

Selmaan N Chettih and Christopher D Harvey. Single-neuron perturbations reveal feature-specific competition in v1. *Nature*, 567(7748):334–340, 2019.

Daniel Chicharro, Stefano Panzeri, and Ralf M Haefner. Stimulus-dependent relationships between behavioral choice and sensory neural responses. *Elife*, 10:e54858, 2021.

Mark M Churchland, M Yu Byron, John P Cunningham, Leo P Sugrue, Marlene R Cohen, Greg S Corrado, William T Newsome, Andrew M Clark, Paymon Hosseini, Benjamin B Scott, et al. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature neuroscience*, 13(3):369–378, 2010.

Stephane Clery, Bruce G Cumming, and Hendrikje Nienborg. Decision-related activity in macaque v2 for fine disparity discrimination is not compatible with optimal linear readout. *Journal of Neuroscience*, 37(3):715–725, 2017.

Marlene R Cohen and John HR Maunsell. Attention improves performance primarily by reducing interneuronal correlations. *Nature neuroscience*, 12(12):1594–1600, 2009.

Ferenc Csikor, Balazs Meszena, and Gergo Orban. Top-down perceptual inference shaping the activity of early visual cortex. *bioRxiv*, pages 2023–11, 2023.

Will Dabney, Zeb Kurth-Nelson, Naoshige Uchida, Clara Kwon Starkweather, Demis Hassabis, Rémi Munos, and Matthew Botvinick. A distributional code for value in dopamine-based reinforcement learning. *Nature*, 577(7792): 671–675, 2020.

Maria C Dadarlat and Michael P Stryker. Locomotion enhances neural encoding of visual stimuli in mouse v1. *Journal of Neuroscience*, 37(14):3764–3775, 2017.

Abhranil Das and Ila R Fiete. Systematic errors in connectivity inferred from activity in strongly recurrent networks. *Nature Neuroscience*, 23(10):1286–1296, 2020.

Peter Dayan and Laurence F Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT press, 2005.

Peter Dayan and Nathaniel D Daw. Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4):429–453, 2008.

Peter Dayan, Geoffrey E Hinton, Radford M Neal, and Richard S Zemel. The helmholtz machine. *Neural computation*, 7(5):889–904, 1995.

Jaime De La Rocha, Brent Doiron, Eric Shea-Brown, Krešimir Josić, and Alex Reyes. Correlation between neural spike trains increases with firing rate. *Nature*, 448(7155):802–806, 2007.

Rob R de Ruyter van Steveninck, Geoffrey D Lewen, Steven P Strong, Roland Koberle, and William Bialek. Reproducibility and variability in neural spike trains. *Science*, 275(5307):1805–1808, 1997.

Sophie Deneve. Bayesian spiking neurons i: inference. *Neural computation*, 20(1):91–117, 2008a.

Sophie Deneve. Bayesian spiking neurons ii: learning. *Neural computation*, 20(1):118–145, 2008b.

Jan Drugowitsch, Valentin Wyart, Anne-Dominique Devauchelle, and Etienne Koechlin. Computational precision of mental inference as critical source of human choice suboptimality. *Neuron*, 92(6):1398–1411, 2016.

Rodrigo Echeveste, Laurence Aitchison, Guillaume Hennequin, and Máté Lengyel. Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *Nature Neuroscience*, 23(9):1138–1149, 2020.

Alexander S Ecker, Philipp Berens, R James Cotton, Manivannan Subramaniyan, George H Denfield, Cathryn R Cadwell, Stelios M Smirnakis, Matthias Bethge, and Andreas S Tolias. State dependence of noise correlations in macaque primary visual cortex. *Neuron*, 82(1):235–248, 2014.

Valérie Ego-Stengel and Matthew A Wilson. Spatial selectivity and theta phase precession in ca1 interneurons. *Hippocampus*, 17(2):161–174, 2007.

Howard Eichenbaum. What versus where: Non-spatial aspects of memory representation by the hippocampus. *Behavioral Neuroscience of Learning and Memory*, pages 101–117, 2018.

A Aldo Faisal, Luc PJ Selen, and Daniel M Wolpert. Noise in the nervous system. *Nature reviews neuroscience*, 9(4): 292–303, 2008.

Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.

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):3635, 2021.

Ian M Finn, Nicholas J Priebe, and David Ferster. The emergence of contrast-invariant orientation tuning in simple cells of cat visual cortex. *Neuron*, 54(1):137–152, 2007.

József Fiser, Chiayu Chiu, and Michael Weliky. Small modulation of ongoing cortical dynamics by sensory input during natural vision. *Nature*, 431(7008):573–578, 2004.

József Fiser, Pietro Berkes, Gergő Orbán, and Máté Lengyel. Statistically optimal perception and learning: from behavior to neural representations. *Trends in cognitive sciences*, 14(3):119–130, 2010.

Timo Flesch, Jan Balaguer, Ronald Dekker, Hamed Nili, and Christopher Summerfield. Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences*, 115(44):E10313–E10322, 2018.

Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138, 2010.

Timothy J Gawne and Barry J Richmond. How independent are the messages carried by adjacent inferior temporal cortical neurons? *Journal of Neuroscience*, 13(7):2758–2771, 1993.

Tom M George, Kimberly L Stachenfeld, Caswell Barry, Claudia Clopath, and Tomoki Fukai. A generative model of the hippocampal formation trained with theta driven local learning rules. *Advances in Neural Information Processing Systems*, 36, 2024.

Samuel J Gershman. What does the free energy principle tell us about the brain? *arXiv preprint arXiv:1901.07945*, 2019.

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

James J Gibson. *The ecological approach to visual perception*. Houghton Mifflin, 1979.

Albert Gidon, Timothy Adam Zolnik, Pawel Fidzinski, Felix Bolduan, Athanasia Papoutsi, Panayiota Poirazi, Martin Holtkamp, Imre Vida, and Matthew Evan Larkum. Dendritic action potentials and computation in human layer 2/3 cortical neurons. *Science*, 367(6473):83–87, 2020.

Robbe LT Goris, J Anthony Movshon, and Eero P Simoncelli. Partitioning neuronal variability. *Nature neuroscience*, 17(6):858–865, 2014.

Agnieszka Grabska-Barwinska, Jeff Beck, Alexandre Pouget, and Peter 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.

Ralf M Haefner, Sebastian Gerwinn, Jakob H Macke, and Matthias Bethge. Inferring decoding strategies from choice probabilities in the presence of correlated variability. *Nature neuroscience*, 16(2):235–242, 2013.

Ralf M Haefner, Pietro Berkes, and József Fiser. Perceptual decision-making as probabilistic inference by neural sampling. *Neuron*, 90(3):649–660, 2016.

Caroline Haimerl, Douglas A Ruff, Marlene R Cohen, Cristina Savin, and Eero P Simoncelli. Targeted v1 comodulation supports task-adaptive sensory decisions. *Nature Communications*, 14(1):7879, 2023.

Guillaume Hennequin, Laurence Aitchison, and Máté Lengyel. Fast sampling-based inference in balanced neuronal networks. In *NIPS*, volume 27, pages 2240–2248. Citeseer, 2014.

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(4):846–860, 2018.

Erik P Hoel, Larissa Albantakis, and Giulio Tononi. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49):19790–19795, 2013.

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

Neil MT Houlsby, Ferenc Huszár, Mohammad M Ghassemi, Gergő Orbán, Daniel M Wolpert, and Máté Lengyel. Cognitive tomography reveals complex, task-independent mental representations. *Current Biology*, 23(21):2169–2175, 2013.

Patrik O Hoyer and Aapo Hyvärinen. Interpreting neural response variability as monte carlo sampling of the posterior. In *Advances in neural information processing systems*, pages 293–300, 2003.

David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

Przemyslaw Jarzebowski, Y Audrey Hay, Benjamin F Grewe, and Ole Paulsen. Different encoding of reward location in dorsal and intermediate hippocampus. *Current Biology*, 32(4):834–841, 2022.

Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

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

Shihao Ji, Ya Xue, and Lawrence Carin. Bayesian compressive sensing. *IEEE Transactions on signal processing*, 56(6): 2346–2356, 2008.

Ilenna Simone Jones and Konrad Paul Kording. Might a single neuron solve interesting machine learning problems through successive computations on its dendritic tree? *Neural Computation*, 33(6):1554–1571, 2021.

Philip X Joris and Laurence O Trussell. The calyx of held: a hypothesis on the need for reliable timing in an intensity-difference encoder. *Neuron*, 100(3):534–549, 2018.

Prakash Kara, Pamela Reinagel, and R Clay Reid. Low response variability in simultaneously recorded retinal, thalamic, and cortical neurons. *Neuron*, 27(3):635–646, 2000.

Johannes HB Kemperman. The general moment problem, a geometric approach. *The Annals of Mathematical Statistics*, 39(1):93–122, 1968.

Tal Kenet, Dmitri Bibitchkov, Misha Tsodyks, Amiram Grinvald, and Amos Arieli. Spontaneously emerging cortical representations of visual attributes. *Nature*, 425(6961):954–956, 2003.

Daniel Kersten and Alan Yuille. Bayesian models of object perception. *Current opinion in neurobiology*, 13(2):150–158, 2003.

David C Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, 2004.

David C Knill and Whitman Richards. *Perception as Bayesian inference*. Cambridge University Press, 1996.

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

Konrad Kording. Decision theory: what" should" the nervous system do? *Science*, 318(5850):606–610, 2007.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, page 4, 2008.

Anna Kutschireiter and Jean-Pascal Pfister. Particle-filtering approaches for nonlinear bayesian decoding of neuronal spike trains. *arXiv preprint arXiv:1804.09739*, 2018.

Kaushik J Lakshminarasimhan, Alexandre Pouget, Gregory C DeAngelis, Dora E Angelaki, and Xaq Pitkow. Inferring decoding strategies for multiple correlated neural populations. *PLoS computational biology*, 14(9):e1006371, 2018.

Richard D Lange and Ralf M Haefner. Characterizing and interpreting the influence of internal variables on sensory activity. *Current opinion in neurobiology*, 46:84–89, 2017.

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):e1009557, 2022.

Richard D Lange, Ankani Chattoraj, Jeffrey M Beck, Jacob L Yates, and Ralf M Haefner. A confirmation bias in perceptual decision-making due to hierarchical approximate inference. *PLoS Computational Biology*, 17(11): e1009517, 2021.

Richard D Lange, Ari S Benjamin, Ralf M Haefner, and Xaq Pitkow. Interpolating between sampling and variational inference with infinite stochastic mixtures. In *Uncertainty in Artificial Intelligence*, pages 1063–1073. PMLR, 2022.

Richard D Lange, Sabyasachi Shivkumar, Ankani Chattoraj, and Ralf M Haefner. Bayesian encoding and decoding as distinct perspectives on neural coding. *Nature Neuroscience*, pages 1–10, 2023.

Pierre Simon Laplace. Théorie analytique des probabilités, 2 vols. *Paris: Courcier Imprimeur*, 1812.

Pierre Simon Laplace. *Théorie analytique des probabilités*. Courcier, 1820.

Tai Sing Lee and David Mumford. Hierarchical bayesian inference in the visual cortex. *JOSA A*, 20(7):1434–1448, 2003.

Robert Legenstein and Wolfgang Maass. Ensembles of spiking neurons with noise support optimal probabilistic inference in a dynamically changing environment. *PLoS computational biology*, 10(10):e1003859, 2014.

Gabor Lengyel, Sabyasachi Shivkumar, and Ralf M Haefner. A general method for testing bayesian models using neural data. In *UniReps: the First Workshop on Unifying Representations in Neural Models*, 2023.

Duncan Luce, David Krantz, Patrick Suppes, and Amos Tversky. Foundations of measurement, vol. iii: Representation, axiomatization, and invariance. 1990.

Artur Luczak, Peter Barthó, and Kenneth D Harris. Spontaneous events outline the realm of possible sensory responses in neocortical populations. *Neuron*, 62(3):413–425, 2009.

Wei Ji Ma. Organizing probabilistic models of perception. *Trends in cognitive sciences*, 16(10):511–518, 2012.

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.

Laurence T Maloney. *Statistical decision theory and biological vision*. na, 2002.

Laurence T Maloney and Pascal Mamassian. Bayesian decision theory as a model of human visual perception: Testing bayesian transfer. *Visual neuroscience*, 26(1):147–155, 2009.

Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78–84, 2013.

Gianluigi Mongillo, Omri Barak, and Misha Tsodyks. Synaptic theory of working memory. *Science*, 319(5869): 1543–1546, 2008.

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

Rubén Moreno-Bote, Jeffrey Beck, Ingmar Kanitscheider, Xaq Pitkow, Peter Latham, and Alexandre Pouget. Information-limiting correlations. *Nature neuroscience*, 17(10):1410–1417, 2014.

Michele Nardin, Jozsef Csicsvari, Gašper Tkačik, and Cristina Savin. The structure of hippocampal ca1 interactions optimizes spatial coding across experience. *Journal of Neuroscience*, 43(48):8140–8156, 2023.

Ifije E Ohiorhenuan, Ferenc Mechler, Keith P Purpura, Anita M Schmid, Qin Hu, and Jonathan D Victor. Sparse coding and high-order correlations in fine-scale cortical networks. *Nature*, 466(7306):617–621, 2010.

John O'Keefe and Jonathan Dostrovsky. The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. *Brain research*, 1971.

Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 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(1):1–14, 2017.

Andrew J Parker and William T Newsome. Sense and the single neuron: probing the physiology of perception. *Annual review of neuroscience*, 21(1):227–277, 1998.

Dejan Pecevski, Lars Buesing, and Wolfgang Maass. Probabilistic inference in general graphical models through sampling in stochastic networks of spiking neurons. *PLoS computational biology*, 7(12):e1002294, 2011.

Benjamin Peters, James J DiCarlo, Todd Gureckis, Ralf Haefner, Leyla Isik, Joshua Tenenbaum, Talia Konkle, Thomas Naselaris, Kimberly Stachenfeld, Zenna Tavares, et al. How does the primate brain combine generative and discriminative computations in vision? *arXiv preprint arXiv:2401.06005*, 2024.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Xaq Pitkow. Compressive neural representation of sparse, high-dimensional probabilities. In *Proceedings of the 25th International Conference on Neural Information Processing Systems-Volume 1*, pages 1349–1357, 2012.

Xaq Pitkow and Markus Meister. Decorrelation and efficient coding by retinal ganglion cells. *Nature neuroscience*, 15 (4):628–635, 2012.

Xaq Pitkow, Sheng Liu, Dora E Angelaki, Gregory C DeAngelis, and Alexandre Pouget. How can single sensory neurons predict behavior? *Neuron*, 87(2):411–423, 2015.

Stephan Pohl, Edgar Y Walker, David L Barack, Jennifer Lee, Rachel N Denison, Ned Block, Florent Meyniel, and Wei Ji Ma. Desiderata of evidence for representation in neuroscience. *arXiv preprint arXiv:2403.14046*, 2024.

Panayiota Poirazi, Terrence Brannon, and Bartlett W Mel. Pyramidal neuron as two-layer neural network. *Neuron*, 37 (6):989–999, 2003.

Adrián Ponce-Alvarez, Alexander Thiele, Thomas D Albright, Gene R Stoner, and Gustavo Deco. Stimulus-dependent variability and noise correlations in cortical mt neurons. *Proceedings of the National Academy of Sciences*, 110(32): 13162–13167, 2013.

Alexandre Pouget, Jeffrey M Beck, Wei Ji Ma, and Peter E Latham. Probabilistic brains: knowns and unknowns. *Nature neuroscience*, 16(9):1170–1178, 2013.

Dobromir Rahnev, Ned Block, Janneke Jehee, and Rachel Denison. Is perception probabilistic? 2020.

Frank P Ramsey. Truth and probability. In *Readings in formal epistemology*, pages 21–45. Springer, 1926.

Rajesh PN Rao. Bayesian computation in recurrent neural circuits. *Neural computation*, 16(1):1–38, 2004.

Fred Rieke, David Warland, Rob de Ruyter Van Steveninck, and William Bialek. *Spikes: exploring the neural code*. MIT press, 1999.

Mattia Rigotti, Omri Barak, Melissa R Warden, Xiao-Jing Wang, Nathaniel D Daw, Earl K Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, 2013.

Douglas A Ruff, Amy M Ni, and Marlene R Cohen. Cognition as a window into neuronal population space. *Annual review of neuroscience*, 41:77–97, 2018.

Camille Rullán Buxó and Cristina Savin. A sampling-based circuit for optimal decision making. *Advances in Neural Information Processing Systems*, 34:14163–14175, 2021.

Maneesh Sahani. Theoretical neuroscience lecture notes. 2021.

Maneesh Sahani and Peter Dayan. Doubly distributional population codes: simultaneous representation of uncertainty and multiplicity. *Neural Computation*, 15(10):2255–2279, 2003.

Noor Sajid, Philip J Ball, Thomas Parr, and Karl J Friston. Active inference: demystified and compared. *Neural computation*, 33(3):674–712, 2021.

Mehrdad Salmasi and Maneesh Sahani. Learning neural codes for perceptual uncertainty. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 2463–2468. IEEE, 2022.

Cristina Savin and Sophie Deneve. Spatio-temporal representations of uncertainty in spiking neural networks. In *NIPS*, volume 27, pages 2024–2032, 2014.

Cristina Savin, Peter Dayan, and Máté Lengyel. Two is better than one: distinct roles for familiarity and recollection in retrieving palimpsest memories. *Advances in Neural Information Processing Systems*, 24:1305–1313, 2011.

Cristina Savin, Peter Dayan, and Máté Lengyel. Optimal recall from bounded metaplastic synapses: predicting functional adaptations in hippocampal area ca3. *PLoS Comput Biol*, 10(2):e1003489, 2014.

Konrad Schmüdgen et al. *The moment problem*, volume 9. Springer, 2017.

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

João D Semedo, Amin Zandvakili, Christian K Machens, M Yu Byron, and Adam Kohn. Cortical areas interact through a communication subspace. *Neuron*, 102(1):249–259, 2019.

Michael N Shadlen, Roozbeh Kiani, Timothy D Hanks, and Anne K Churchland. An intentional framework. *Better than conscious*, pages 71–101, 2008.

Neda Shahidi, Paul Schrater, Tony Wright, Xaq Pitkow, and Valentin Dragoi. Population coding of strategic variables during foraging in freely-moving macaques. *BioRxiv*, page 811992, 2019.

Charles Scott Sherrington. Observations on the scratch-reflex in the spinal dog. *The Journal of physiology*, 34(1-2): 1–50, 1906.

Sabyasachi Shivkumar, Richard D Lange, Ankani Chattoraj, and Ralf M Haefner. A probabilistic population code based on neural samples. *arXiv preprint arXiv:1811.09739*, 2018.

Sabyasachi Shivkumar, Madeline S Cappelloni, Ross K Maddox, and Ralf M Haefner. Inferring sources of suboptimality in perceptual decision making using a causal inference task. *bioRxiv*, pages 2022–04, 2022.

Suhas Shrinivasan, Konstantin-Klemens Lurz, Kelli Restivo, George Denfield, Andreas Tolias, Edgar Walker, and Fabian Sinz. Taking the neural sampling code very seriously: A data-driven approach for evaluating generative models of the visual system. *Advances in Neural Information Processing Systems*, 36, 2024.

Fabian H Sinz, Xaq Pitkow, Jacob Reimer, Matthias Bethge, and Andreas S Tolias. Engineering a less artificial intelligence. *Neuron*, 103(6):967–979, 2019.

Jørgen Sugar and May-Britt Moser. Episodic memory: Neuronal codes for what, where, and when. *Hippocampus*, 29 (12):1190–1205, 2019.

David J Tolhurst, J Anthony Movshon, and Andrew F Dean. The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision research*, 23(8):775–785, 1983.

Misha Tsodyks, Tal Kenet, Amiram Grinvald, and Amos Arieli. Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science*, 286(5446):1943–1946, 1999.

Balazs B Ujfalussy and Gergő Orbán. Sampling motion trajectories during hippocampal theta sequences. *Elife*, 11: e74058, 2022.

Rajkumar Vasudeva Raju and Zachary Pitkow. Inference by reparameterization in neural population codes. *Advances in Neural Information Processing Systems*, 29:2029–2037, 2016.

Eszter Vértes and Maneesh Sahani. Flexible and accurate inference and learning for deep generative models. In *NeurIPS*, 2018.

Eszter Vértes and Maneesh Sahani. A neurally plausible model learns successor representations in partially observable environments. *Advances in Neural Information Processing Systems*, 32:13714–13724, 2019.

Heinrich Von Helmholtz. Treatise on physiological optics vol. iii. 1867.

Martin J Wainwright and Michael Irwin Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

Martin J Wainwright, Odelia Schwartz, and Eero P Simoncelli. 10 natural image statistics and divisive normalization. *Probabilistic models of the brain*, page 203, 2002.

Martin J Wainwright, Tommi S Jaakkola, and Alan S Willsky. Tree-based reparameterization framework for analysis of sum-product and related algorithms. *IEEE Transactions on information theory*, 49(5):1120–1146, 2003.

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(1):122–129, 2020.

Li Kevin Wenliang and Maneesh Sahani. A neurally plausible model for online recognition and postdiction in a dynamical environment. *Advances in Neural Information Processing Systems*, page 672089, 2020.

Zhengwei Wu, Minhae Kwon, Saurabh Daptardar, Paul Schrater, and Xaq Pitkow. Rational thoughts in neural codes. *Proceedings of the National Academy of Sciences*, 117(47):29311–29320, 2020.

Qianli Yang, Edgar Walker, R James Cotton, Andreas S Tolias, and Xaq Pitkow. Revealing nonlinear neural decoding by analyzing choices. *Nature communications*, 12(1):1–13, 2021.

Richard Zemel and Peter Dayan. Distributional population codes and multiple motion models. *Advances in neural information processing systems*, 11, 1998.

Richard S Zemel, Peter Dayan, and Alexandre Pouget. Probabilistic interpretation of population codes. *Neural computation*, 10(2):403–430, 1998.

Ehud Zohary, Michael N Shadlen, and William T Newsome. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, 370(6485):140–143, 1994.