bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sampling-based Bayesian inference in recurrent circuits of stochastic spiking neurons

Wen-Hao Zhang[1−4,†], Si Wu[5], Krešimir Josić[6,7*] Brent Doiron[1−4*]

[1]Departments of Neurobiology and Statistics, University of Chicago, Chicago, IL, USA.

[2]Grossman Center for Quantiative Biology and Human Behavior, University of Chicago, Chicago, IL, USA.

[3]Department of Mathematics, University of Pittsburgh, Pittsburgh, PA, USA.

[4]Center for the Neural Basis of Cognition, Pittsburgh, PA, USA.

[5]School of Electronics Engineering and Computer Science, IDG/McGovern Institute for Brain Research, Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China.

[6]Department of Mathematics, University of Houston, TX, USA.

[7]Department of Biology and Biochemistry, University of Houston, TX, USA.

† Present address: Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center, Dallas, TX, USA.

*Equal contribution.

Correspondence should be addressed to: bdoiron@uchicago.edu or kresimir.josic@gmail.com

Abstract

Two facts about cortex are widely accepted: neuronal responses show large spiking variability with near Poisson statistics and cortical circuits feature abundant recurrent connections between neurons. How these spiking and circuit properties combine to support sensory representation and information processing is not well understood. We build a theoretical framework showing that these two ubiquitous features of cortex combine to produce optimal sampling-based Bayesian inference. Recurrent connections store an internal model of the external world, and Poissonian variability of spike responses drives flexible sampling from the posterior stimulus distributions obtained by combining feedforward and recurrent neuronal inputs. We illustrate how this framework for sampling-based inference can be used by cortex to represent latent multivariate stimuli organized either hierarchically or in parallel. A neural signature of such network sampling are internally generated differential correlations whose amplitude is determined by the prior stored in the circuit, which provides an experimentally testable prediction for our framework.

Keywords: Sampling-based Bayesian inference, Poisson spiking neurons, Recurrent network dynamics, Differential correlations.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

In an uncertain and changing world, it is imperative for the brain to reliably represent and interpret external stimuli. The cortex is essential for the representation of the sensory world, and it is believed that populations of neurons collectively code for richly structured sensory scenes [1]. However, two central characteristics of cortical circuits remain to be properly integrated into population coding frameworks. First, neuronal activity in sensory cortices is often noisy, showing significant variability of spiking responses evoked by the same stimulus [2, 3]. In many traditional coding frameworks such spiking variability degrades the representation of stimuli by cortical activity [4]. Why cortical responses display large spiking variability while isolated cortical neurons can respond reliably remains a mystery. Second, the primary source of synaptic inputs to cortical neurons does not come from upstream centers which convey sensory signals, but rather from recurrent pathways between cortical neurons [5–7]. While such recurrent connections are often organized about a stimulus feature axis [8, 9], it is not obvious whether or how their presence improves overall representation. We propose a biologically motivated inference coding scheme where these two ubiquitous cortical circuit features, variability in spike generation and recurrent connections, together support a probabilistic representation of stimuli in rich sensory scenes.

Numerous studies have framed sensory processing in the cortex in terms of Bayesian inference (e.g., [10–16]). Specifically, the 'Bayesian brain' hypothesis posits that sensory cortex infers and synthesizes a posterior distribution of the latent stimuli which describes the probability of possible stimuli that could have given rise to the sensory inputs. Performing Bayesian inference requires cortex to store an internal model that represents how sensory inputs and external stimuli are generated. Once a sensory input is received, cortical dynamics inverts this internal model in a process termed 'analysis-by-synthesis' [12], and represents the posterior distributively across neurons and/or across time [15, 16]. In this study, we propose that recurrent connections in cortical circuits store the prior of latent stimuli to produce the posterior distribution when combined with evidence from sensory inputs. Moreover, we posit that Poisson spiking variability provides a source of fluctuations needed for generating random samples from the inferred posterior.

To test these hypotheses we consider a recurrent circuit model where neurons receive stochastic feedforward inputs which carry information about the external world, and respond with Poisson-distributed spiking activity. We find that such Poissonian spiking provides the variability that allows the network to generate samples from posterior stimulus distributions with differing uncertainties. We use this sampling framework to illustrate circuit-based Bayesian inference given two distinct generative models of stimuli in the external world: one organized hierarchically with a stimulus variable that depends on a latent context variable, and a second where a pair of latent stimuli are organized in parallel. In both cases a recurrent circuit is able to generate samples from the joint posterior, and infer the values of the latent variables. We show through both analytic derivation

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and simulations that recurrent connections represent the correlation structure of these models, and the weight of these connections can be tuned to optimally capture the prior distribution of stimuli in the external world. The stronger the correlation between the latent variables, the stronger the recurrent connections need to be for the network to generate samples from the correct posterior distribution.

Finally, a neural signature of this circuit-based sampling mechanism is internally generated population noise correlations aligned with the stimulus response direction, often referred to as "differential correlations" [4, 17]. In our framework, the amplitude of internally generated differential correlations is determined by the recurrent connection strength, which also determines the prior stored by the circuit. Since optimal inference requires a specific magnitude of recurrent connectivity, differential correlations resulting from such recurrent connectivity are a potential signature of optimal coding. This is in contrast to the deleterious impact of externally generated differential correlations. We thus predict that the correlation structure of the external world shapes recurrent wiring in neural circuits, and is reflected in the pattern of differential noise correlations. We use this logic to provide testable predictions from our framework for sampling-based Bayesian inference by recurrent, stochastic cortical circuits.

Results

Recurrent circuitry and spiking variability do not improve conventional neural codes

We start with the classic example of a sensory stimulus, s, encoded in neuronal population activity, $\mathbf{r}$, from which a stimulus estimate $\hat{s}$ can be decoded (Fig. 1A, top) [18]. It is reasonable to expect that neuronal circuitry is adapted to accurately represent ethologically relevant stimuli. However, as we will show next, in simple coding schemes two ubiquitous features of cortical circuits – internal spiking variability and recurrent connectivity – are at best irrelevant for, and in many cases degrade, the accuracy of these representations.

In population coding frameworks stimuli are encoded by a neuronal population with individual neurons tuned to a preferred stimulus value. The preferred values of all neurons cover the whole range of stimuli [18–20] (Fig. 1B, bottom); if s ranges over a periodic domain (such as the orientation of a bar in a visual scene, or the direction of an arm reach) then it is commonly assumed that the neurons' preferred stimuli are distributed on a ring (Fig. 1B, top). To generate neuronal responses from such a population we simulate a network of neurons whose spiking activity, $\mathbf{r}_t$, at time t is Poissonian with instantaneous firing rate $\boldsymbol{\lambda}_t$ (Eq. 11). For simplicity we assume linear (or linearized) neuronal transfer and synaptic interactions (Eqs. 10-11), so that the firing rate is a linear function of the feedforward and recurrent inputs. We couple excitatory (E) neurons with similar stimulus preferences more strongly [8, 9] to one another, compared to neuron pairs with dissimilar tuning. In

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

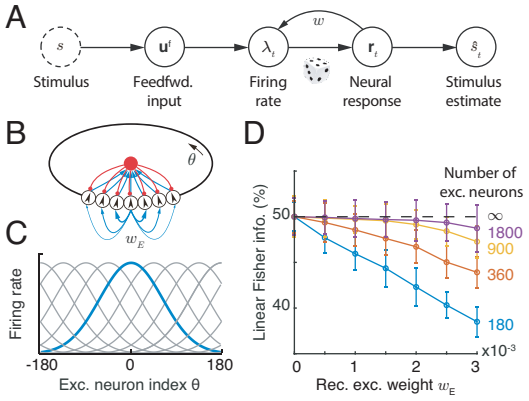

Figure 1: A network with structured recurrent connections limits the linear Fisher Information (LFI) about external stimuli. (A) A schematic diagram showing how a stimulus, s, is encoded in neuronal response, $\mathbf{r}_t$. A stimulus estimate, $\hat{s}_t$, can be obtained from $\mathbf{r}_t$. (B) A recurrent ring model (top) where the connections between excitatory neurons are dependent on their distance along the ring. Blue arrows: excitatory synapses with line width denoting connection strength; red arrows: inhibitory synapses. (C) The population activity of excitatory neurons in the ring model, $\mathbf{r}_t$, dependent on a stimulus, s. The blue curve shows the population activity in response to $s = 0$, and gray curves the activities in response to stimuli with values at the peak locations of the curves. (D) For finite size networks (colored lines; ratio of excitatory to inhibitory neurons kept constant) LFI decreases as w_E increases. In the limit of infinite network size LFI does not depend on w_E (dashed line). Since neural responses are variable, LFI in the neuronal response converges to only half of the LFI in the feedforward input.

this way the recurrent E connectivity has the same circular symmetry as the stimulus (Fig. 1B). In contrast, connections between inhibitory (I) neurons are unstructured, and inhibitory activity acts to stabilize network activity [21]. A stimulus, e.g. $s = 0$, results in elevated activity of E neurons with the corresponding preference (Fig. S1A). As expected, an increase in the strength of recurrent excitatory connections increases both the firing rates and the trial-to-trial pairwise covariability (i.e. noise correlations) in the responses [2] (Fig. S2A). This canonical network model has been widely used to explain cortical network dynamics and neural coding [21–23].

We use linear Fisher Information (LFI) to quantify the impact of recurrent connectivity and internal spiking variability on the accuracy of the stimulus estimate, $\hat{s}_t$, from the activity vector $\mathbf{r}_t$ (see details in Eq. S39 in Supplemental Information). The inverse of LFI provides a lower bound on the expected square of the difference between the true value, s, and the estimate, $\hat{s}_t$, made by a linear decoder [1, 4, 17–19, 24]. In the limit of an infinite number of neurons available to the decoder LFI is unaffected by recurrent connectivity strength, w_E (Fig. 1D, dashed line). This is because the mean response of the network is linear in its inputs, and an (invertible) linear transformation can neither increase nor decrease LFI (see Eq. S38 in Supplemental Information). For networks with a finite number of neurons, the variability from spike generation is shared between neurons via recurrent interactions. Consequently an increase in coupling strength, w_E, reduces LFI in finite networks (Fig. 1D, colored lines).

In sum, recurrent connectivity and spiking variability do not improve, and often degrade, stim-

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ulus representation in the network (as measured by LFI). Since synaptic coupling is biologically expensive, a network that most accurately and cheaply represents a stimulus is then one with no recurrent connections (i.e., $w_E = 0$) and minimal spiking variability. Nevertheless, connectivity in mammalian cortex is highly recurrent [5–7, 9], and neural responses are highly variable [2, 3]. What is then the purpose of these extensive recurrent connections between cortical neurons, and why are their responses so noisy?

While classical population code theory often explains how to generate point estimates of a stimulus (Fig. 1A), numerous studies suggest that the brain performs Bayesian inference to synthesize and estimate the probability distribution of latent stimuli from sensory inputs (e.g., [10–15, 25, 26]). To compute this posterior a neural circuit needs to combine a stored representation of the prior distribution of the stimulus with the likelihood conveyed by feedforward inputs. We propose that recurrent connectivity can be used to represent the prior and spiking variability can generate samples from this posterior distribution. Before we present our full model we first show how sampling-based inference can be implemented in a population of spiking neurons.

Internally generated Poisson spiking variability drives sampling-based Bayesian inference

Many studies suggest that neuronal response variability is a signature of sampling in neural circuits (e.g., [16, 27–32]). In these studies the instantaneous population responses, $\mathbf{r}_t$, represent a sample of a latent stimulus, and the empirical distribution of stimulus samples collected over time is an approximation of the posterior distribution. Furthermore, response variability is typically modeled using a continuous (e.g., Gaussian) distribution [27, 29–33]. However, spike trains from cortical neurons are often Poissonian, and spike counts are discrete [3, 34]. It is unclear if discrete Poissonian variability can generate samples from stimuli with continuous probability distributions (e.g., orientation, moving direction) with the flexibility needed to represent different stimulus uncertainties.

We address this question using a theory based on a simple model network composed of excitatory (E) Poissonian neurons (Eqs. 10-11), and subsequently support our findings by simulating a network containing both E and inhibitory (I) neurons (e.g. Fig. 1B). We start by showing that Poissonian spiking in a population of tuned neurons can drive sampling from a well–defined distribution. We assume that the instantaneous firing rates of a population of E neurons, $\boldsymbol{\lambda}_t$, have a bell-shaped (Gaussian) profile (Fig. 2B), so that for the j^{th} neuron $\boldsymbol{\lambda}_{tj} = R\exp[\mathbf{h}_j(\bar{s}_t)] = R\exp[-(\bar{s}_t-\theta_j)^2/2a^2]$ (See Eq. 12 in Methods). Here θ_j is the preferred stimulus of neuron j, a is the width of the tuning curve, and $\bar{s}_t$ is the location of the peak of the firing rate profile, $\boldsymbol{\lambda}_t$, in stimulus space (x-axis in Fig. 2B). Note that the value of $\bar{s}_t$ is arbitrary here, but we will later relate it to the input to the population. The (smooth) Gaussian tuning curves simplify the analysis, but are not essential for

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

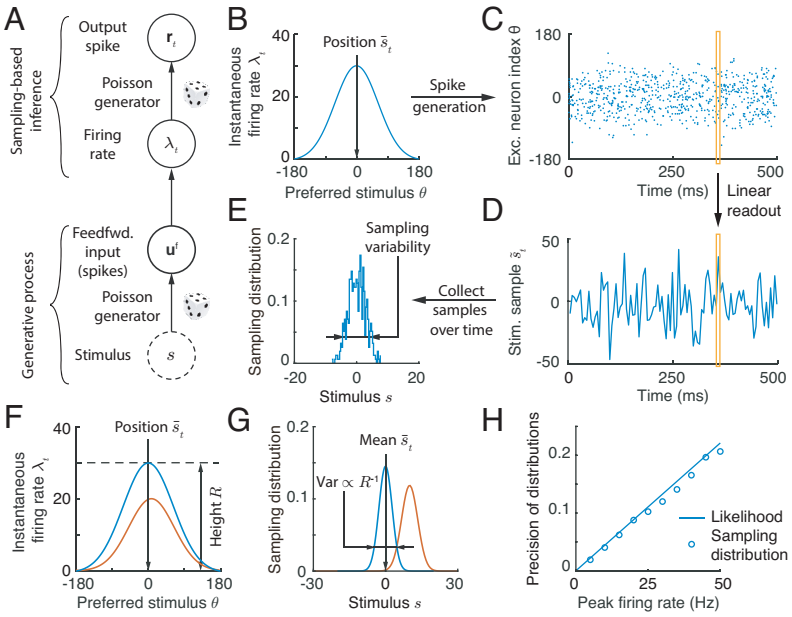

Figure 2: Spike generation with Poissonian variability can support sampling-based Bayesian inference. (A) We use a feedforward network model (no recurrent connections) to demonstrate how spiking variability drives sampling. Neurons receive feedforward inputs, $\mathbf{u}^{\mathrm{f}}$, modeled as independent Poisson spike trains, resulting in a Poissonian population response, $\mathbf{r}_t$, with means determined by the instantaneous firing rate vector, $\boldsymbol{\lambda}_t$. (B-E) Demonstration of sampling via stochastic spike generation. A population of neurons with Gaussian tuning and firing rates $\boldsymbol{\lambda}_t$ (B) generates a realization a population response, $\mathbf{r}_t$ (C). A sample from the posterior distribution of the stimulus (D, orange box) can be linearly read out from the population response (C, orange box). (E) The sampling distribution is obtained by collecting stimulus samples over time. (F-G) The profile of population firing rates (F) determines the sampling distribution (G). The position of the population firing rate, $\bar{s}_t$, determines the mean of the sampling distribution, and the variance of the sampling distribution is inversely proportional to the peak firing rate, R. We show two population activity profiles, one in blue and the other in orange, to illustrate these points. (H) In an E-I network, the precision of the sampling distribution (the inverse of sampling variability) read out from E neurons increases with the height of firing rate, and is consistent with the likelihood directly read out from the feedforward input.

the argument. Finally, the preferred stimuli of the E neurons, $\{\theta_j\}_{j=1}^{N_E}$, are uniformly distributed over the stimulus range (Fig 1B). In each time interval the population activity is given by a vector of independent Poisson random variables, $\mathbf{r}_t$, with means determined by the instantaneous firing rate vector $\boldsymbol{\lambda}_t$ (Fig. 2B-C). At each time, t, this spiking activity produces a stimulus sample, $\tilde{s}_t$, from the probability distribution determined by the instantaneous firing rates, $\boldsymbol{\lambda}_t$ (Fig. 2D, see Methods),

$$\tilde{s}_t \sim p(\tilde{s}|\boldsymbol{\lambda}_t) \propto \exp[\mathbf{h}(\tilde{s})^\top \boldsymbol{\lambda}_t] \propto \mathcal{N}(\tilde{s}|\bar{s}_t, \Lambda^{-1}). \tag{1}$$

With the Gaussian firing rate profile we use here, the stimulus sample, $\tilde{s}_t$, can be read out as $\tilde{s}_t = \sum_j \mathbf{r}_{tj}\theta_j / \sum_j \mathbf{r}_{tj}$ (Eq. 14 and Fig. 2D), which can be thought of as the location of the response, $\mathbf{r}_t$, in stimulus space (y-axis in Fig. 2C). The collection of stimulus samples across time ($\{\tilde{s}_t\}$; Fig. 2E), determines the sampling distribution $q(s) = T^{-1}\sum_t \delta(s - \tilde{s}_t)$ which approximates the

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

distribution $p(s|\boldsymbol{\lambda}_t)$, i.e., $p(s|\boldsymbol{\lambda}_t) \approx q(s)$ [16, 35]. Here $\delta(\cdot)$ is the Dirac delta function and T is the number of samples.

To use this mechanism to produce samples from the posterior distribution of a stimulus, we must define a generative model for the feedforward inputs evoked by a stimulus. We take the feedforward input to the neural population, $\mathbf{u}^{\mathsf{f}}$, to be a vector of independent Poisson spike counts with Gaussian tuning over the stimulus, s. Following assumptions widely used in previous studies of probabilistic population codes (PPC) [36, 37], we assume that the mean input spike count to the j^{th} excitatory neuron in the population is $\langle \mathbf{u}^{\mathsf{f}}_j(s) \rangle \propto \exp[\mathbf{h}_j(s)] = \exp[-(s-\theta_j)^2/2a^2]$. A single realization of the input, $\mathbf{u}^{\mathsf{f}}$, in a time interval encodes the whole likelihood function over the stimulus, $p(\mathbf{u}^{\mathsf{f}}|s)$ [36]. This likelihood is proportional to a Gaussian due to the Gaussian profile of feedforward input (Eq. 19),

$$\begin{aligned} p(\mathbf{u}^{\mathsf{f}}|s) &= \textstyle\prod_{j=1}^{N_E} \text{Poisson}\big[\langle \mathbf{u}^{\mathsf{f}}_j(s) \rangle\big], \\ &\propto \exp\big[\mathbf{h}(s)^\top \mathbf{u}^{\mathsf{f}}\big], \\ &\propto \mathcal{N}(s|\mu_{\mathsf{f}}, \Lambda_{\mathsf{f}}^{-1}). \end{aligned} \tag{2}$$

Here the likelihood mean, μ_{f}, is determined by the location of $\mathbf{u}^{\mathsf{f}}$ in stimulus space, and the precision, Λ_{f}, is proportional to the spike count (or height) of $\mathbf{u}^{\mathsf{f}}$ (Eq. 20). Since a realization of the feedforward input encodes the whole likelihood function, we present a fixed $\mathbf{u}^{\mathsf{f}}$ to the network over time (dropping the time index t), and describe how samples from the posterior $p(s|\mathbf{u}^{\mathsf{f}})$ are generated by the network.

A simple example of inference via sampling is provided by a population of E neurons without recurrent connections and instantaneous firing rates equal to the feedforward input, $\boldsymbol{\lambda}_t = \mathbf{u}^{\mathsf{f}}$ (Eq. 10), and hence constant in time (Fig. 2A). In this feedforward network Poisson spike generation produces samples from the normalized likelihood, i.e., $\tilde{s}_t \sim p(\tilde{s}|\boldsymbol{\lambda}_t) \propto p(\mathbf{u}^{\mathsf{f}}|\tilde{s})$, and consequently the network represents a uniform stimulus prior (i.e., $p(s)$ is a constant).

To test our theory, we simulated the response of a network of tuned excitatory (E) and untuned inhibitory (I) neurons (Fig. 2A,C) to a fixed but randomly generated feedforward input (Eq. 18). While the E neurons shared no recurrent connections, the E and I neurons were connected to maintain stable network activity. To confirm that the overall firing rate dictated the sampling variability (Eq. 1), we increased the feedforward input rate, which reduced the width of the likelihood (Eq. 2). As a result, the sampling precision (inverse of the sampling variance) increased and matched the precision of the likelihood (Fig. 2G, H), even as the normalized response variability (measured the by Fano factor) of single neurons remained unchanged.

While the above analysis introduces the key components of a sampling-based theory of inference, stimulus sampling using a feedforward network is unnecessary: A single observation of the response $\mathbf{r}$ in a deterministic feedforward network ($\mathbf{r} = \mathbf{u}^{\mathsf{f}}$ after removing spike generation in Eq. 11) would

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

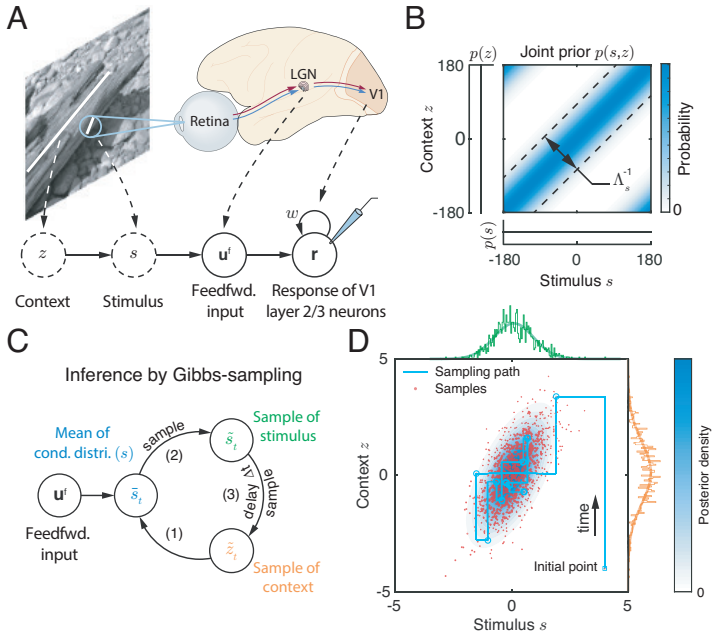

Figure 3: A hierarchical generative model and posterior inference via Gibbs sampling. (A) An example of sensory feedforward input generation: The context, z, is the orientation of the tree trunk, and the stimulus, s, is the orientation of the bark texture located in the classical receptive field of a V1 hypercolumn. The recurrent circuit generates samples from the joint posterior over stimulus and context. Solid circles: observations and responses in the brain; dashed circles: latent variables in the external world. (Natural image and brain schematic adapted from [38] and [39] respectively). (B) The joint prior over the context, z, and stimulus, s, is concentrated on the diagonal. The correlation between context and stimulus is determined by parameter Λ_s. (C) The posterior over context and stimulus can be approximated via Gibbs sampling (Eqs. 4a-4c) by iteratively generating samples of s and z from their respective conditional distributions. (D) The resulting approximations of the joint and marginal posterior over the latent stimulus, s, and context, z. Light blue contour: the posterior distribution (Eq. 24); Red dots: Samples obtained using Gibbs sampling. The green and orange projections are the marginal posterior distributions of the stimulus s and context z respectively.

also represent the whole likelihood [36], avoiding the costly process of collecting samples $\tilde{s}_t$ across time. We next consider more interesting cases, and show that spiking variability in recurrent networks can drive sampling from more complex posterior distributions.

Recurrent cortical circuit samples a hierarchical generative model

Recurrent networks can store a variety of generative model structures; to demonstrate the generality of our sampling framework we provide two example generative models which serve as building blocks for more complex models. We first consider a two-stage hierarchical model of feedforward inputs received by the cortical circuit (Fig. 3A). The first stage of our model consists of a stimulus, s, and a context, z, both of which are one dimensional for simplicity. The structure of the world is described by the joint distribution, $p(s, z)$. Using the visual system as motivation, s, could be the orientation of the visual texture within a classical receptive field (local information) of a

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

hypercolumn of V1 neurons, while the orientation within a non-classical receptive field of these cells could describe the corresponding context, z (Fig. 3A). The likelihood of the stimulus based on a given context, $p(s|z) = \mathcal{N}(s|z, \Lambda_s^{-1})$, is Gaussian with precision Λ_s. For simplicity, we assume that the context prior, $p(z)$, is uniform, which implies that the marginal prior of s, is also uniform (Fig. 3B). This assumption is not essential for our main conclusions but does simplify the analysis. Importantly, the joint prior of stimulus and context, $p(s, z)$, can have non-trivial structure with the density concentrated around the diagonal $s = z$ (Fig. 3B). The precision Λ_s measures how strongly the context, z, and the stimulus, s, are related, and thus determines how strongly their joint distribution is concentrated around the diagonal.

The second stage of the generative model describes how the feedforward input depends on the stimulus, s; this is identical to our prior treatment (See Eq. 2). Combining these two stages provides a complete description of the generative model for the feedforward input received by neurons in the population,

$$\begin{aligned} p(\mathbf{u}^{\mathsf{f}}|s)p(s|z)p(z) &\propto \textstyle\prod_{j=1}^{N_E} \text{Poisson}\big(\mathbf{u}_j^{\mathsf{f}}|s\big)p(s|z), \\ &\propto \mathcal{N}(s|\mu_{\mathsf{f}}, \Lambda_{\mathsf{f}}^{-1})\mathcal{N}(s|z, \Lambda_s^{-1}). \end{aligned} \tag{3}$$

Given this hierarchical model we can show that the joint posterior over stimulus and context features, $p(s, z|\mathbf{u}^{\mathsf{f}})$ is a bivariate normal distribution (see Eq. 24), and we next use it to evaluate the accuracy of the sampling distribution.

Gibbs sampling of the joint stimulus and context posterior

One approach to approximate the joint distribution over stimulus and context is Gibbs sampling [31, 35, 40, 41] which starts with an initial guess for the value of the two latent variables, and proceeds by alternately generating samples of one variable from the distribution conditioned on the value of the second variable. More precisely, to approximate the joint posterior of s and z (Eq. 3), Gibbs sampling proceeds by generating a sequence of samples, $(\tilde{s}_t, \tilde{z}_t)$ indexed by time t, through recursive iteration of the following steps (Fig. 3C and Eq. 25),

$$\text{Compute}: \quad p(\tilde{s}|\tilde{z}_t, \mathbf{u}^{\mathsf{f}}) \propto p(\mathbf{u}^{\mathsf{f}}|\tilde{s})p(\tilde{s}|\tilde{z}_t) \equiv \mathcal{N}(\tilde{s}|\bar{s}_t, \Lambda^{-1}), \tag{4a}$$

$$\text{Sample}: \quad \tilde{s}_t \sim p(\tilde{s}|\tilde{z}_t, \mathbf{u}^{\mathsf{f}}), \tag{4b}$$

$$\text{Sample}: \quad \tilde{z}_{t+\Delta t} \sim p(\tilde{z}|\tilde{s}_t) = \mathcal{N}(\tilde{z}|\tilde{s}_t, \Lambda_s^{-1}). \tag{4c}$$

Here Δt is the time increment between successive samples. The samples (red dots in Fig. 3D) are generated by alternately fixing the values of the two variables, so that sampling trajectories alternate between horizontal and vertical jumps (cyan lines in Fig. 3D). The empirical distribution of samples, i.e., $q(s, z|\mathbf{u}^{\mathsf{f}}) = T^{-1} \sum_t \delta\left[(s, z)^\top - (\tilde{s}_t, \tilde{z}_t)^\top\right]$ with $\top$ denoting vector transpose, approximates the

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

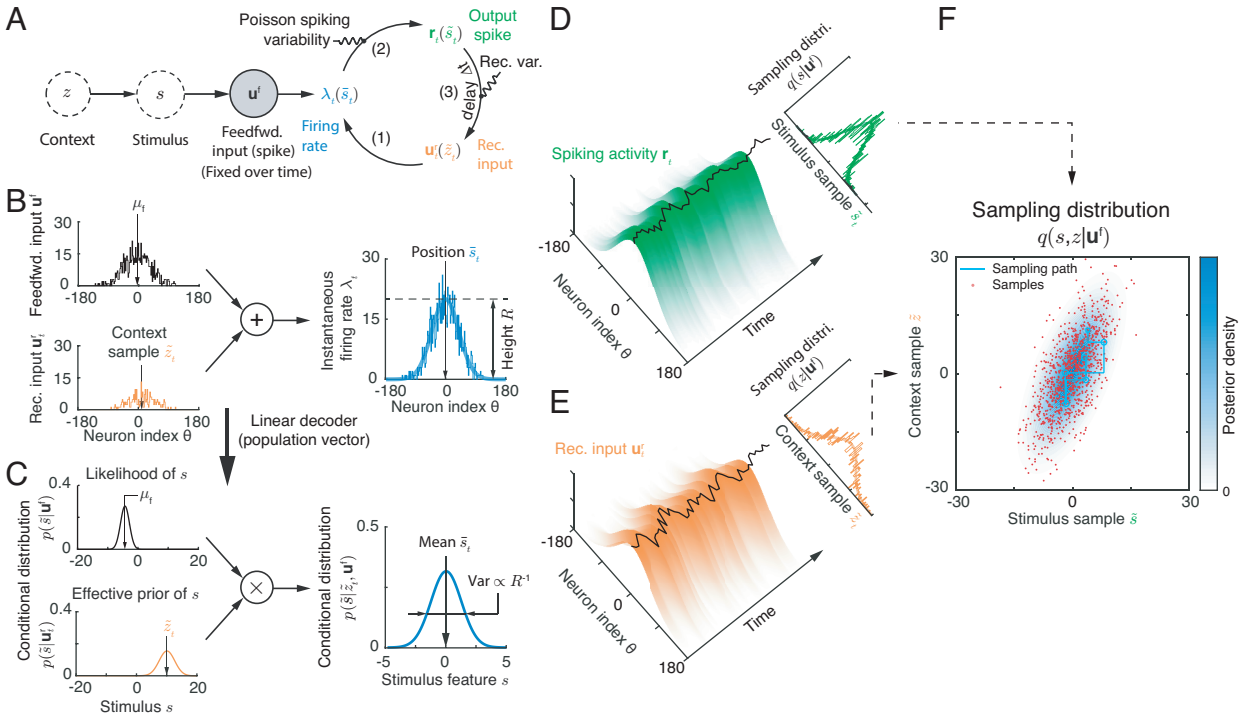

Figure 4: A recurrent circuit generates samples from the posterior defined by a hierarchical generative model. (A) Schematic of recurrent circuit dynamics, in which stimulus, s, and context, z, features are encoded respectively in the population response, $\mathbf{r}_t$, and recurrent inputs, $\mathbf{u}^{\mathsf{r}}_t$. (B-C) When the feedforward inputs and recurrent inputs share the same tuning profile, summing the two inputs to define the instantaneous firing rate (B) is equivalent to multiplying the conditional distributions encoded by the two inputs to obtain the conditional distribution of the stimulus, $p(s|\tilde{z}_t, \mathbf{u}^{\mathsf{f}})$. (C) The conditional distributions of the stimulus can be explicitly read out from corresponding population responses by a linear decoder (B). (D-F) Reading out the joint sampling distribution from the recurrent circuit. The projection of the spiking activity (Eq. 14) and recurrent inputs (Eq. 29) onto the stimulus subspace (black curves), can be read out linearly from the population activity and interpreted as a sample of stimulus and context respectively (Eqs. 4b-4c). Top right insets: the empirical marginal distributions of samples and marginal posteriors (smooth lines). (F) The joint value (red dots) of instantaneous samples of stimulus (black curve on the surface in D), and context (black curve on surface in E) represent samples from the joint posterior of the stimulus and context. The true joint posterior is represented by the blue contour.

joint posterior $p(s, z|\mathbf{u}^{\mathsf{f}})$ (blue contour map in Fig. 3D, Eq. 24) [35]. To approximate $p(s|\mathbf{u}^{\mathsf{f}})$, the marginal posterior distribution of s, we can use only samples $\tilde{s}_t$ to obtain the approximating distribution $q(s|\mathbf{u}^{\mathsf{f}})$ (compare the two green lines at the margin in Fig. 3D). The same is true for the marginal posterior over z.

Implementing Gibbs sampling of stimulus and context in a recurrently coupled cortical circuit

An implementation of Gibbs sampling in a recurrent E circuit can be intuitively understood by comparing the recurrent network dynamics (Fig. 4A) with the dynamics described by the Gibbs sampling algorithm (Fig. 3C). In the recurrent network a stimulus sample, $\tilde{s}_t$, is represented by the activity of E cells, $\mathbf{r}_t$, while a context sample, $\tilde{z}_t$, is represented by recurrent inputs, $\mathbf{u}^{\mathsf{r}}_t$. To generate

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

correct samples we require that the conditional distribution that is represented by the instantaneous firing rate, $\boldsymbol{\lambda}_t$ (Eq. 1), matches the conditional distribution used in the Gibbs sampling algorithm (Eq. 4b), so that $p(\tilde{s}|\tilde{z}_t, \mathbf{u}^{\mathsf{f}}) = p(\tilde{s}|\boldsymbol{\lambda}_t) \propto \exp[\mathbf{h}(\tilde{s})^\top \boldsymbol{\lambda}_t]$. Equating the two distributions (see Eqs. 4a and 10) yields the relation,

$$\begin{aligned} &\ln p(\tilde{s}|\tilde{z}_t, \mathbf{u}^{\mathsf{f}}) = \ln p(\mathbf{u}^{\mathsf{f}}|\tilde{s}) + \ln p(\tilde{s}|\tilde{z}_t), \\ \Leftrightarrow \quad &\mathbf{h}(\tilde{s})^\top \boldsymbol{\lambda}_t = \mathbf{h}(\tilde{s})^\top \mathbf{u}^{\mathsf{f}} + \mathbf{h}(\tilde{s})^\top \mathbf{u}^{\mathsf{r}}_t. \end{aligned} \tag{5}$$

This equation holds when two constraints are satisfied: First, the firing rate vector, $\boldsymbol{\lambda}_t$, needs to have a Gaussian profile peaked at $\bar{s}_t$, i.e., the mean of $p(\tilde{s}|\tilde{z}_t, \mathbf{u}^{\mathsf{f}})$ (Eq. 4a). Second, the peak firing rate, R, needs to be proportional to the precision of $p(\tilde{s}|\tilde{z}_t, \mathbf{u}^{\mathsf{f}})$, i.e., $R \propto \Lambda$ (see Fig. 2F-G). In a neural circuit one way for $\boldsymbol{\lambda}_t$ to satisfy these constraints is for feedforward inputs, $\mathbf{u}^{\mathsf{f}}$, and recurrent inputs, $\mathbf{u}^{\mathsf{r}}_t$, to both have Gaussian profiles with the same width, a, as that of $\boldsymbol{\lambda}_t$ (by sharing the same $\mathbf{h}(\tilde{s})$, Eqs. 5 and 12). This is because the sum of two Gaussian-profile inputs with the same width, a, gives a firing rate, $\boldsymbol{\lambda}_t$, with the same tuning, as long as the difference of the locations of two inputs is much smaller than the width, a. Our generative model (Eq. 3) produces feedforward input, $\mathbf{u}^{\mathsf{f}}$, with a Gaussian profile and encodes the likelihood function $p(\mathbf{u}^{\mathsf{f}}|\tilde{s})$. The recurrent input, $\mathbf{u}^{\mathsf{r}}_t$, then need to represent the conditional distribution $p(\tilde{s}|\tilde{z}_t)$. Hence, to satisfy Eq. (5) the recurrent input $\mathbf{u}^{\mathsf{r}}_t$ should have the same Gaussian profile as $\mathbf{u}^{\mathsf{f}}$ (Eq. 29), with its location and magnitude determined by the mean and precision of $p(\tilde{s}|\tilde{z}_t)$, respectively.

If recurrent interactions are absent (setting $\mathbf{u}^{\mathsf{r}}_t = 0$), then network activity, $\mathbf{r}_t$, generates samples from the normalized likelihood, $p(\mathbf{u}^{\mathsf{f}}|\tilde{s})$, as we showed previously when describing feedforward networks (Fig. 2). When neurons only receive recurrent inputs (setting $\mathbf{u}^{\mathsf{f}} = 0$), the network generates samples from the conditional distribution $p(\tilde{s}|\tilde{z}_t)$. Driven by a sum of recurrent and feedforward inputs the network generates samples from a distribution given by the product of the conditional distributions encoded by both inputs respectively (Fig. 4B-C).

The recurrent weights must be adjusted so that the recurrent input has the appropriate magnitude and width to encode the likelihood $p(s|z)$. To simplify the exposition we first assume that E neurons are only self-connected, so that the width of recurrent input trivially matches that of the feedforward input (otherwise recurrence will broaden the profile of the firing rate activity $\boldsymbol{\lambda}_t$ over the network). To constrain the magnitude of the recurrent weights we require that the sum of the recurrent inputs satisfies $\sum_j \mathbf{u}^{\mathsf{r}}_{tj} \propto \Lambda_s$. Since $\mathbf{u}^{\mathsf{r}}_j = w_E \mathbf{r}_j$ and the width of $\mathbf{u}^{\mathsf{r}}_j$ and $\mathbf{r}_j$ are equal, the magnitude of the recurrent weights that result in samples from the correct posterior must satisfy:

$$w^*_E = \frac{\langle \mathbf{u}^{\mathsf{r}}_j \rangle}{\langle \mathbf{r}_j \rangle} = \frac{\langle \sum_j \mathbf{u}^{\mathsf{r}}_j \rangle}{\langle \sum_j \mathbf{r}_j \rangle} = \frac{\Lambda_s}{\Lambda_{\mathsf{f}} + \Lambda_s}, \tag{6}$$

where Λ_s and Λ_{f} are the precision of likelihood $p(s|z)$ and $p(\mathbf{u}^{\mathsf{f}}|s)$ respectively (Eq. 3). The optimal

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

recurrent weight, w_E^*, thus encodes the correlation between the stimulus s and the context z. An increase in correlation between s and z, resulting in a narrower diagonal band in $p(s, z)$ (Fig. 3B), requires an increase in the recurrent weight w_E^* for optimal sampling. When context and stimulus are uncorrelated so that $\Lambda_s = 0$, the hierarchical generative model (Fig. 3A) is equivalent to the generative model without context (Fig. 2A) and recurrent interactions are not needed for sampling (i.e., w_E^*=0). Our framework (Eq. 6) thus predicts that optimal Bayesian inference is achieved with recurrent synaptic weights which depend on the correlative structure of the external world. We numerically test this prediction in the next section.

A stochastic E-I spiking network jointly samples stimulus and context

To confirm the predictions of this analysis, we simulated a full recurrent network consisting of both E and I neurons with Poisson spike train statistics (see details in Eqs. 47-50). The E neurons were synaptically connected to each other (Eq. 49, see Fig. 1A), in contrast to the simple network of self-connected E neurons we described above. While recurrent E to E coupling broadens the tuning of excitatory recurrent input, lateral inhibition can sharpen Gaussian firing rate profiles so that it matches that of the feedforward inputs (as required by Eq. 5).

The activity of the recurrent network in response to a fixed but randomly generated feedforward input (Eq. 3) can be decoded to produce samples from the bivarite posterior distribution of the stimulus and context. As above, samples from the conditional stimulus distribution are represented by the activity of E neurons (Eq. 14), while samples from the conditional context distribution are represented by recurrent inputs received by E neurons (Eq. 29; black curves overlaid on the top of population responses in Fig. 4D and E respectively). To update recurrent inputs we only used neuronal activity at the previous time step. Thus, the activities of E neurons and their recurrent inputs were updated in alternation, consistent with Gibbs sampling. The trajectory obtained by plotting the stimulus sample read out from the network activity on one axis, and plotting the context sample read out from recurrent E inputs on another axis then exhibits the characteristics of Gibbs sampling (Fig. 4F, cyan line). The resulting sampling distribution provides a good approximation to the joint posterior of stimulus and context (compare red dots and blue contour in Fig. 4F). Inhibitory neurons again did not respond selectively to either the stimulus or the context.

For the network to generate samples from the joint posterior, the recurrent connectivity should depend on the correlation between the stimulus and the context (Eq. 6). To verify this prediction, we fixed the generative model (Eq. 3) and changed only the recurrent weights in the network. For simplicity, we only varied the peak E weight, w_E (Eq. 49), and maintained network stability by fixing the ratio between E and I synaptic weights. While increasing w_E did not change the sampling mean, it did increase the variance of the context sampling distribution, and increased the correlation between stimulus and context samples (Fig. 5A).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

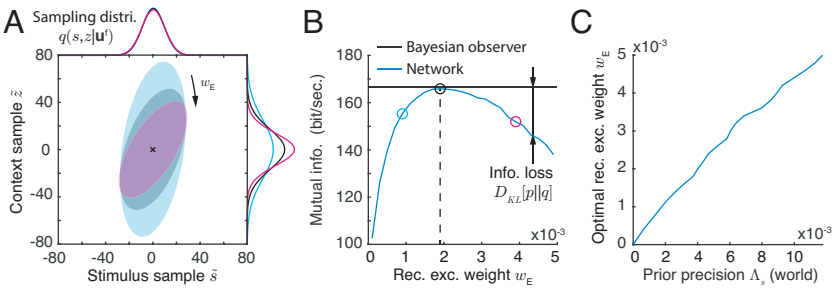

Figure 5: The joint sampling distribution of stimulus and context changes with the recurrent weight in the network. (A) The sampling distribution for different recurrent excitatory weights, w_E. The ratio of excitatory and inhibitory weights was fixed. Ellipses capture three standard deviations from the mean of the joint sampling distribution. Different colors correspond to the three values of w_E, denoted by different symbols in panel B. (B) The mutual information between the latent variables, s and z, and the feedforward inputs for an ideal Bayesian observer (black horizontal line) and for the sampling distribution generated by the network model (blue curve). The difference between the two lines is the KL divergence between the posterior, $p(s, z|\mathbf{u}^f)$, and the sampling distribution, $q(s, z|\mathbf{u}^f)$. KL divergence is minimized when the weight in the recurrent network is set to a value, w_E^*, at which the sampling distribution, q, best matches the true posteriori, p (black circle). (C) This optimal weight, w_E^*, increases with prior precision, Λ_s.

We use Kullback-Leibler (KL) divergence to measure the distance between the sampling distribution, $q(s, z|\mathbf{u}^f)$, and the true posterior, $p(s, z|\mathbf{u}^f)$ (Eq. 24). The KL divergence quantifies the loss of mutual information, measured in bits, between the latent variables (s and z) and the feedforward inputs, $\mathbf{u}^f$, when the true posterior, p, is approximated by the distribution, q (Eq. 42) [35]. The mutual information loss in the network is minimized at a unique value of the recurrent weight, w_E^*, at which the sampling distribution, q, best matches the posterior, p (Fig. 5B, black circle). To confirm that this optimal recurrent weight, w_E^*, increases with the correlation in the prior (precision Λ_s, Eq. 6), we numerically obtained the recurrent weight that minimizes the mutual information loss for each value of Λ_s in the generative model. These results confirmed the predictions of our theory (Eq. 6, Fig. 5C): When $\Lambda_s = 0$, i.e. when context and stimulus are uncorrelated, a network with no interactions performs best ($w_E^* = 0$), while for small Λ_s (relative to Λ_f) the optimal weight w_E^* is positive and increases with Λ_s. In total, we have described a potential mechanism for a recurrent network of spiking neurons to perform sampling-based Bayesian inference.

Generating samples from multi-dimensional posteriors with coupled neural circuits

To demonstrate the generality of the proposed neural code we next consider a world described by a broad, rather than deep (hierarchical) generative model. Information about each of two latent stimuli, $\mathbf{s} = (s_1, s_2)$, is relayed by corresponding feedforward inputs received by a neural circuit (Fig. 6A). We assume the prior is a bivariate Gaussian distribution (Fig. 6B), i.e., $p(\mathbf{s}) \propto \exp[-\Lambda_s(s_1-s_2)^2/2] \equiv \mathcal{N}(s_1-s_2, \Lambda_s^{-1})$, so that Λ_s ($\Lambda_s \geq 0$) characterizes the correlation between s_1 and s_2. Furthermore, each stimulus, s_m, independently generates feedforward spiking inputs, $\mathbf{u}_m^f$, each of which is received by a separate network and produces responses $\mathbf{r}_m$ for $m = 1, 2$ (Fig. 6A).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

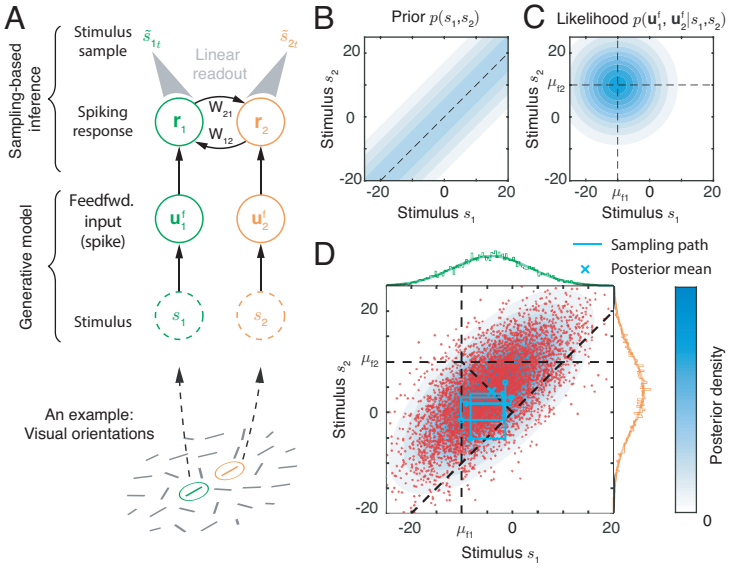

Figure 6: Distributed sampling from a multivariate posterior distributions using coupled networks. (A) Network m ($m = 1, 2$) receives a feedforward input evoked by a stimulus, s_m. The coupling between the two networks represents the stimulus prior. A linear readout from each network, m, can be interpreted as a sample from the posterior of the stimulus, s_m. (B-C) Examples of a prior (B) and likelihood (C). The prior distribution is concentrated around the diagonal line (dashed line), indicating the two stimuli are more likely to be colinear. In panel (C), $\mu_{f1} = -10$ and $\mu_{f2} = 10$ are the means of the likelihoods of s_1 and s_2 respectively. (D) The joint posterior of stimuli and the corresponding approximate sampling distribution generated by the coupled networks. A sample from the joint posterior can be read out individually from the activity of the corresponding network (shown in A). Light blue contour: the posterior distribution (Eq. 34); Red dots: stimulus samples generated by the network.

Thus, the full generative model of the input has the form,

$$\begin{aligned} p(\mathbf{u}^{\mathrm{f}}|\mathbf{s})p(\mathbf{s}) &= \Big[\prod_{m=1}^{2} p(\mathbf{u}_m^{\mathrm{f}}|s_m)\Big]p(s_1, s_2), \\ &\propto \Big[\prod_{m=1}^{2} \mathcal{N}(s_m|\mu_{\mathrm{f}m}, \Lambda_{\mathrm{f}m}^{-1})\Big]\mathcal{N}(s_1 - s_2, \Lambda_s^{-1}). \end{aligned} \tag{7}$$

The likelihood $p(\mathbf{u}_m^{\mathrm{f}}|s_m)$ is the same as that given previously (Eq. 2), where the feedforward inputs, $\mathbf{u}_m^{\mathrm{f}}$, are again described by conditionally independent Poisson spike counts with Gaussian tuning over stimulus s_m. As a concrete example, the two stimuli, s_m, could represent orientations of local edges falling in the central receptive fields of a V1 hypercolumn (Fig. 6A, bottom), with each V1 hypercolumn modeled by a network producing the response $\mathbf{r}_m$ (Fig. 6A, top). Then Λ_s characterizes *a priori* tendency of the stimuli to share similar orientations, and determines how likely two local edges are to be part of a global line, as in the case of contour integration [42, 43]. However, the generative model defined by Eq. (7) is quite general and has been also used to explain multisensory cue integration [10] and sensorimotor learning [13].

The posterior is a bivariate Gaussian distribution (Fig. 6D, Eq. 34) whose mean is shifted from the likelihood mean (Fig. 6C) towards to the diagonal line, because of the correlations between

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the stimuli in the prior (Fig. 6B). We can again use Gibbs sampling to approximate the posterior $p(\mathbf{s}|\mathbf{u}^{\mathsf{f}})$ using the following steps,

$$\text{Compute}: \quad p(\tilde{s}_1|\mathbf{u}_1^{\mathsf{f}}, \tilde{s}_{2,t-\Delta t}) \propto p(\mathbf{u}_1^{\mathsf{f}}|\tilde{s}_1)p(\tilde{s}_{2,t-\Delta t}|\tilde{s}_1), \tag{8a}$$

$$\text{Sample}: \quad \tilde{s}_{1t} \sim p(\tilde{s}_1|\mathbf{u}_1^{\mathsf{f}}, \tilde{s}_{2,t-\Delta t}), \tag{8b}$$

where $\tilde{s}_{1t}$ and $\tilde{s}_{2t}$ are instantaneous samples at time t of stimuli s_1 and s_2 respectively. We only give the steps needed to produce samples from the conditional distribution of s_1, as samples from the conditional distribution of s_2 can be obtained using the same steps after exchanging indices.

These sampling steps can be implemented distributively in a coupled neural circuit using a mechanism similar to that we described in the case of a hierarchical generative model. The activity of each network, $\mathbf{r}_m$, individually represents samples from the (marginal) posterior of s_m (Fig. 6A, top). The joint posterior is then approximated as the collection of samples represented by the activity pairs $(\mathbf{r}_1, \mathbf{r}_2)$. Taking network $m = 1$ as an example, spike response $\mathbf{r}_{1t}$ produces a stimulus sample $\tilde{s}_{1t}$ as long as the instantaneous firing rate $\boldsymbol{\lambda}_{1t}$ represents the conditional distribution $p(\tilde{s}_1|\mathbf{u}_1^{\mathsf{f}}, \tilde{s}_{2,t-\Delta t})$ (Eq. 8a). Since the feedforward input, $\mathbf{u}_1^{\mathsf{f}}$, represents the likelihood $p(\mathbf{u}_1^{\mathsf{f}}|\tilde{s}_1)$, to obtain the appropriate firing rates, $\boldsymbol{\lambda}_{1t}$, the recurrent input from network 2 to network 1, $\mathbf{u}_{12,t}^{\mathsf{r}}$, must encode the correct conditional distribution, $p(\tilde{s}_{2,t-\Delta t}|\tilde{s}_1)$. As in the case of the mechanism we proposed to implement sampling as described by Eq. (5), $\mathbf{u}_{12,t}^{\mathsf{r}}$ needs to have the same Gaussian profile as the firing rate $\boldsymbol{\lambda}_{1t}$, the position of $\mathbf{u}_{12,t}^{\mathsf{r}}$ on the stimulus space should match the mean of $p(\tilde{s}_{2,t-\Delta t}|\tilde{s}_1)$, i.e., $\tilde{s}_{2,t-\Delta t} = \sum_j \mathbf{u}_{12,tj}^{\mathsf{r}}\theta_j / \sum_j \mathbf{u}_{12,tj}^{\mathsf{r}}$, and the magnitude of $\mathbf{u}_{12,t}^{\mathsf{r}}$ must be proportional to the prior correlation, $\Lambda_s \propto \sum_j \mathbf{u}_{12,tj}^{\mathsf{r}}$ (Eq. 39). Hence, each network can sum the feedforward input and the recurrent input from its counterpart to obtain an update to the instantaneous conditional distribution given by Eq. (8a), and generate independent Poisson spikes to produce a sample from the instantaneous conditional distribution (Eq. 8b). Notably, the sample of each stimulus can be locally read out from corresponding network (Eq. 41, Fig. 6A), even if the activities of two networks are correlated.

Since the recurrent input strength represents the stimulus correlation in the prior determined by precision Λ_s, the coupling between the two networks needs to be tuned to generate the appropriate recurrent input. Indeed, in a network with only E neurons, and connections only between neurons with the same preferred stimulus value but in different networks, the optimal homogeneous connection strength is $w_{mn}^* = \langle \mathbf{u}_{mn,j}^{\mathsf{r}} \rangle / \langle \mathbf{r}_{n,j} \rangle = \Lambda_s/(\Lambda_{\mathsf{f}n} + \Lambda_s)$ (Eq. 40). This mirrors the result obtained with the hierarchical model presented earlier in Eq. (6).

Coupled E-I spiking networks sample bivariate dimensional posteriors

To test the feasibility of the proposed mechanisms for generating samples from a bivariate posterior we simulated a pair of bidirectionally coupled circuits consisting of E and I neurons (Fig. 7A).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

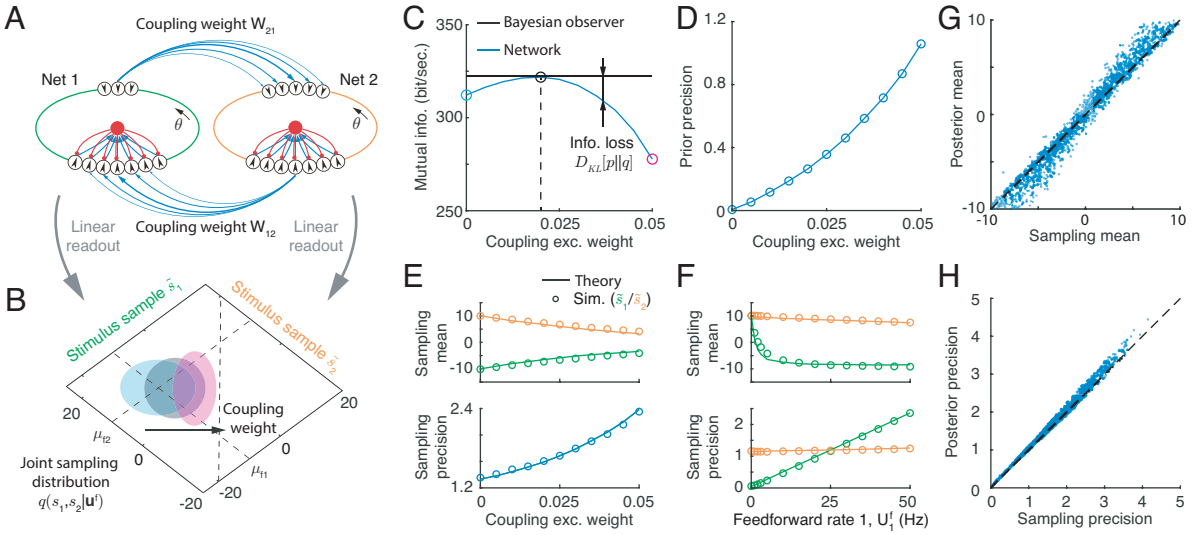

Figure 7: The statistics of the multivariate sampling distribution of stimuli generated by coupled E-I circuits. (A) Each of the two circuits individually generate a sample of a corresponding stimulus which can be read out linearly from that circuit's activity. Combining the readouts from the two networks yields the joint sampling distribution. The ring color indicates the stimulus sample the circuit generates: green and orange represent the stimulus s_1 and s_2, respectively. Blue arrows: E synapses with width denoting connection strength; red arrows: I synapses. (B) The sampling distribution shifts from the likelihood mean to the diagonal line as the coupling between the networks increases. Ellipses capture one standard deviation from the mean of the sampling distribution. Different colors correspond to the three different coupling weights between the circuits shown in panel (C). (C) The mutual information between latent variables and the feedforward inputs for the ideal Bayesian observer (black) and the sampling distributions generated by the network with different coupling weights between the two circuits. (D) The optimal coupling weight that minimizes information loss also increases with prior precision (which is inversely proportional to the width of the band in Fig. 6B). (E) The mean and precision of the sampling distribution over the two stimuli change with the coupling weight between the circuits when the feedforward input is fixed. (F) The mean and precision of the sampling distribution over the two stimuli change with the firing rate of feedforward input to network 1, with other network parameters fixed. (G-H) Comparison of the mean (G) and precision (H) of the sampling distributions with the posteriors under different combinations of feedforward inputs and coupling weights. Different dots are obtained from the sampling distributions obtained under different combinations of input direction and strength, and coupling weight between networks.

This neural circuit model can be extended to generate samples from higher dimensional posterior distribution (see Discussion). Each circuit receives feedforward input generated by one of the two stimuli. On every time step the sample of each stimulus, $\tilde{s}_{mt}$, can be individually and linearly read out from the response of corresponding network, $\mathbf{r}_{mt}$ (Eq. 41). Jointly, the two stimulus samples, one each from both networks, $\tilde{\mathbf{s}}_t = (\tilde{s}_{1t}, \tilde{s}_{2t})^\top$, provide a sample from the joint posterior of the two latent stimuli (Fig. 7B). We assumed that the synaptic connections between the networks, w_{mn} $(m, n = 1, 2;\ m \neq n)$, are excitatory, but target both E and I neurons, while inhibitory connections are local to each network. We also adjusted network parameters so that the profiles of the inputs across networks (e.g., the inputs from network 2 to 1) have the same tuning profile as the feedforward inputs (see Methods). Since we assumed uniform marginal priors (see Eq. 32),

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

recurrent connections between E neurons within the a circuit were absent, while E and I neurons within a circuit were recurrently connected to ensure network stability. For simplicity, we chose parameters so that the two circuits were symmetric, but the strength of the feedforward inputs to each could differ.

We asked whether the activity of the two coupled circuits can generate samples from bivariate posteriors, and how the sampling distribution depends on the coupling, w_{mn}, between the two circuits. An increase in synaptic coupling between the two networks caused the sampling distribution to shift from the likelihood mean towards the diagonal (Fig. 7B), resulting in stimulus samples, $\tilde{s}_{1t}$ and $\tilde{s}_{2t}$ that were more similar. This is consistent with an increase in stimulus correlation in the multivariate prior, Λ_s (Eq. 7). To confirm our prediction that the optimal coupling strength between the two networks, w^*_{mn}, increases with the stimulus correlation in the prior, Λ_s, we numerically obtained the coupling weight that minimizes the loss of mutual information between latent stimuli and feedforward inputs (Fig. 7C). The optimal synaptic weight between the circuits increased with stimulus correlation in the prior. At the optimal weight, w^*_{mn}, the sampling distribution was close to the true posterior, showing that a properly tuned circuit can generate samples from the correct distribution (Fig. 7D).

We next asked how the sampling distribution in the network depends on network and feedforward input parameters. As the coupling between the two circuits increased, the sample means of both stimuli converge (Fig. 7E, top) and the sampling precision of both stimuli increased as well (Fig. 7E, bottom), in agreement with a more correlated stimulus prior. We also tested whether a network with fixed parameters can generate samples from a family of posteriors with different uncertainties. To do so, we changed the uncertainty of the likelihood of s_1 by changing the firing rate in the feedforward input $\mathbf{u}^{\mathrm{f}}_1$ received by network 1. We observed that with a narrower likelihood of s_1, the sample means of both stimuli shifted towards the mean of likelihood of s_1 (-10°), and sampling precision increased, consistent with a change in the posterior distribution (Fig. 7F). Lastly, to demonstrate the robustness of this network implementation of sampling-based inference we compare the sampling distributions to the true posteriors under different combinations of input and network parameters (Fig. 7G-H), in each case setting the recurrent coupling to the optimal value, w^*_{mn}, obtained numerically. Across different parameter values we observe excellent agreement in both the mean (Fig. 7G) and precision (Fig. 7H) of the two densities. In sum, our recurrent network of spiking neuron models can be extended to support sampling-based Bayesian inference with multi-dimensional stimuli.

A signature of stimulus sampling: internally generated differential noise correlations

A central prediction of our circuit framework for sampling-based Bayesian inference is that an increase in the correlation between stimuli in the sensory world should result in stronger synapses

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

between neurons whose activities represent these stimuli (see Eq. 6). This is a difficult prediction to test since measuring synaptic connectivity along a functional axis is already challenging [44], let alone measuring a change in synaptic strength owing to a change in stimulus statistics. Here we outline a testable prediction of our theory by identifying a measurable, population-level signature of changes in functionally related recurrent synaptic strengths.

In response to a fixed feedforward input the responses of a recurrent circuit implementing stimulus sampling will fluctuate. The alignment of the recurrent circuitry and neuronal stimulus tuning causes a portion of these activity fluctuations to align with the subspace in which stimuli are coded. As an example, consider the sampling implemented by a single recurrent network (Fig. 4A), and suppose the population response fluctuates around its mean position (0° in the example of Fig. 8A), ignoring fluctuations along other directions in neuronal response space. The activity of neuron pairs with stimulus preference both above or below the mean position are positively correlated (the black and blue neurons in Fig. 8A), while the activity of neuron pairs with preferences straddling the mean are negatively correlated (the black and red neurons in Fig. 8A). Such stimulus sampling generates a covariance component which is proportional to the outer product of the derivative of neuronal tuning (Fig. 8B), i.e., $\mathbf{f}'_s\mathbf{f}'^{\top}_s$, where $\mathbf{f}'_s$ denotes the derivative of tuning $\mathbf{f}(s) = \langle \boldsymbol{\lambda}_t \rangle$ (mean firing rate) over stimulus s. Such noise correlations have been referred to as *differential correlations* [4, 17], and are generally viewed as deleterious to stimulus coding. Stochastic sampling in coupled networks (Fig. 6A) produces similar differential noise correlations (see Supplemental Information).

In our network implementation of sampling, the amplitude of internally generated differential correlations is not arbitrary, but is determined by the recurrent connection strength, w_E^*. Here, the differential covariance matrix of population responses has the form (see Eq. 44)

$$\begin{aligned} &\Sigma_{DC} = V(\bar{s}|\mathbf{u}^{\mathsf{f}})\mathbf{f}'_s\mathbf{f}'^{\top}_s, \\ &\text{where} \quad V(\bar{s}|\mathbf{u}^{\mathsf{f}}) = \frac{\Lambda_s}{\Lambda_{\mathsf{f}}(\Lambda_{\mathsf{f}} + \Lambda_s)} = a^2 n_{\mathsf{f}}^{-1} w_E^*, \end{aligned} \tag{9}$$

where $V(\bar{s}|\mathbf{u}^{\mathsf{f}})$ is the variance of $\bar{s}_t$ in equilibrium over time, and $\bar{s}_t$ is the mean of the instantaneous conditional distribution (Eq. 4a) represented by the position of instantaneous firing rate $\boldsymbol{\lambda}_t$ (Fig. 2B). Importantly, the amplitude of differential correlations increases with the recurrent weight, w_E^*, which is set by the prior precision Λ_s (Eq. 6; Fig. 8C). Thus, in our framework internally generated differential correlations are a by-product of inference by sampling from posterior distributions of stimuli in a structured world.

Distinguishing external and internal differential correlations

The previous analysis of internally generated differential correlations in a circuit implementing sampling-based inference is based on the assumption of a fixed feedforward input (Eq. 9). However,

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

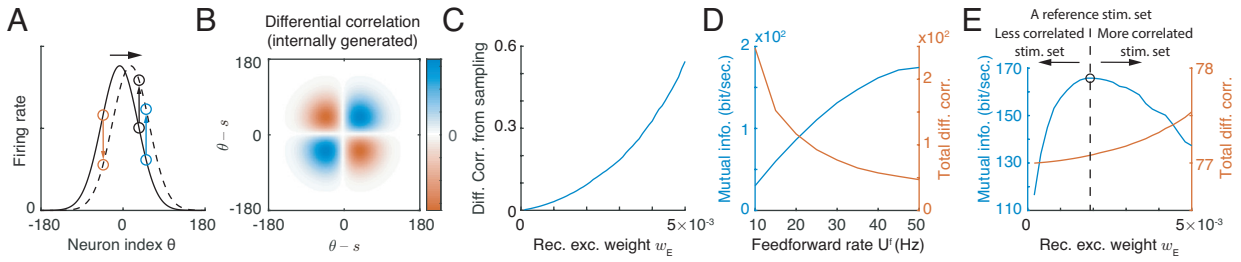

Figure 8: Stimulus sampling by a network is reflected in the internally generated differential correlations, whose impact differs from differential correlations inherited from feedforward inputs. (A) Stimulus sampling via spike generation causes the population firing rate to fluctuate along the stimulus subspace (x-axis). (B) The pattern of internally generated differential correlation in a network implementing sampling composed of neurons with Gaussian tuning. (C) Internally generated differential correlations in such a network increase with recurrent weight, w_E. (D) The rate in feedforward input decreases the externally generated correlations, and increases the mutual information between the feedforward inputs and latent stimulus. (E) Recurrent network weights increase internally generated differential correlations. Mutual information between stimulus and feedforward inputs changes non-monotonically with recurrent weight. The direction of arrows indicates the predicted direction of change of the recurrent weights after an animal is retrained using a new stimulus set with different correlations compared to the reference stimulus set.

in typical neurophysiology experiments an external stimulus, s, is fixed, while the feedforward input, $\mathbf{u}^{\mathrm{f}}$, fluctuates due to variability in sensory acquisition and transmission noise (Eqs. 3 and 7). Hence, differential correlations of neuronal population responses are a combination of correlations inherited from feedforward input [45], and correlations generated by recurrent network interactions that align with the population stimulus tuning [24]. When the feedforward input is described by a hierarchical generative model (Eq. 2), the total magnitude of differential correlations in the evoked response is $a^2 n_{\mathrm{f}}^{-1} w_E \mathbf{f}_s' \mathbf{f}_s'^{\top} + a^2 n_{\mathrm{f}}^{-1} \mathbf{f}_s' \mathbf{f}_s'^{\top}$ (see Eq. 46), where the second term reflects differential correlations inherited from the feedforward input (compare with Eq. 9). Although the two sources of differential correlations are intertwined in the neuronal response, they impact the information content differently thus offering a potential way to distinguish between them in neural data.

Externally generated differential correlations decrease with feedforward input rate which could be modulated by visual stimulus strength such as contrast (Fig. 8D, red curve). As a consequence, the mutual information (the information between feedforward inputs $\mathbf{u}^{\mathrm{f}}$ and the latent variables, i.e., s and z, sampled by recurrent network in Fig. 4A, Eq. 42) increases with feedforward input intensity (Fig. 8A, blue curve). We therefore have a monotonic, decreasing relationship between externally generated differential correlations and mutual information. This is expected since such inherited correlations always impair information processing, as observed previously [4, 17]. In contrast, an increase in recurrent weights, w_E, increases internally generated differential correlations, but results in a non-monotonic change in mutual information (Fig. 8B). Hence there is a non-monotonic relation between internally generated differential correlations and the mutual information between stimulus and feedforward inputs. In sum, the impact of external and internal differential correlations on stimulus coding can be distinguished by their respective monotonic and non-monotonic relation

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with the mutual information between stimulus and response.

Discussion

We have presented a framework in which neuronal response variability and recurrent synaptic connections, two ubiquitous features of cortex, are jointly used to implement sampling-based Bayesian inference in neuronal circuit models. Combining mathematical analysis and network simulations we established that stereotypical Poisson variability of discrete spike counts can drive flexible sampling from a family of continuous distributions. The sampling statistics are determined by the structure of recurrent coupling, which stores information about the stimulus prior, and feedforward inputs which convey the stimulus likelihood. Sampling-based inference is implemented in two steps: the instantaneous firing rate, determined by the sum of feedforward and recurrent inputs, represents the instantaneous conditional distribution of latent stimulus, while Poissonian variability in spike generation is used to generate a random stimulus sample from this conditional distribution. A simple circuit model is able to generate samples from multi-dimensional posteriors of latent variables organized hierarchically or in parallel, which underlies the computational basis of a wide range of perceptual and cognitive processes [46].

Comparison with other neural coding frameworks

The neural code we described shares some features with codes described in previous studies, including parametric representations in probabilistic population codes (PPCs) [15, 36, 37], and sampling-based codes (SBCs) [16, 27–32]. In our framework the conditional distributions of latent variables is represented by instantaneous firing rates which linearly encode the logarithms of these conditional distributions, and have a mathematical form that is similar to that used in past studies describing PPCs (e.g., Eq. 5). Further, the posterior is represented by stimulus samples generated through a random process, a feature of all SBCs. Despite these similarities, there are fundamental differences between the neural code we described and previously proposed PPCs and SBCs.

PPCs are generally implemented in networks with no internally generated variability, with stochasticity inherited from the stimulus. In contrast, our proposed network is doubly stochastic: The Poisson variability in the feedforward input allows a single realization of the feedforward input to represent the whole stimulus likelihood [36], while internally generated Poisson variability drives stimulus sampling. Further, in PPCs the posterior is represented parametrically by a one-shot neuronal response, while in our proposed network the joint posterior is approximated by a sequence of samples, each obtained as a linear readout from the instantaneous neuronal responses. Although it takes time to collect sufficient samples to approximate the posterior, a computational benefit compared with PPCs is that inference of a multivariate posterior can be implemented by linearly coupled networks (Fig. 6), while in PPCs nonlinear coupling between networks is required [47].

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Conventional SBCs are used to generate samples directly in a neural space whose dimension is given by the number of neurons in the population [16, 27, 28, 30–33], where a neuronal response, $\mathbf{r}_t$, is interpreted directly as a sample from the (marginal) posterior of neuronal responses, $p(\mathbf{r})$. Hence the posterior mean is the temporally averaged population response, and the covariance of population responses is the posterior covariance. In contrast, our proposed network generates samples in a low dimensional stimulus subspace embedded in high dimensional neural activity space. The linear projection of network activity, $\mathbf{r}_t$, onto the stimulus subspace represents a sample from the stimulus posterior, similar to a previous study [29]. A computational benefit of sampling in a low dimensional stimulus subspace is convergence speed, as the volume of the stimulus subspace is significantly smaller than that of the neural activity space. Indeed, in our examples sequences of samples generated by a single recurrent network (Fig. 4) and coupled networks (Fig. 6) can both converge to an equilibrium distribution in less than 20ms, which is fast enough to complete inference on a behaviorally relevant time scale (Fig. S6). Furthermore, the multiplication of probability distributions of latent stimulus, which is central to Bayesian inference (e.g., cue combination, decision making, see review in [15]), can be implemented by summing the inputs to a neuronal population (Eq. 5). This follows from the fact that the instantaneous population input (or firing rate) linearly encodes the logarithm of a probability distribution (Eqs. 1 and 5). In contrast, producing samples in neural activity space using conventional SBCs requires nonlinear operations in neural circuits in order to multiply probability distributions (or histograms) of the samples [15].

A recent study demonstrated that an E-I recurrent network of rate-based neurons can be numerically optimized for sampling-based Bayesian inference [32]. In contrast, we used a theoretical approach to derive a network model of simplified spiking neurons which implements sampling-based inference. This allowed us to explicitly describe the putative neural mechanisms needed for such sampling. Although the two studies use different generative models and neural representations, the network models in both studies share some common characteristics: ring structure, Poisson-like response variability, and tuning-dependent noise correlation (Fig. S1D). This implies that the seemingly different generative models and neural representations in the two studies reflect more general principles, as suggested in [48]. It will be interesting to extend our theoretical approach to dynamical spiking neurons to determine how the timescales of neuronal dynamics and neuronal oscillations impact inference in rich, dynamic sensory scenes (see below).

Testing the prediction that recurrent synaptic strength is determined by correlations between latent stimuli

Differential noise correlations generated by recurrent network interactions are a signature of network sampling in our framework (Fig. 5C and 8C). This is in contrast to earlier studies where differential correlations were inherited from feedforward inputs [17, 49]. While internally generated differential

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

correlations could also emerge from a recurrent circuit which is not implementing inference [22, 24, 49–52] or implementing inference via other algorithms [53], in our framework the relation between the magnitude of internally generated differential correlations, the posterior uncertainty, and the strength of the recurrent synaptic weight (Eq. 9) provides a clear test which can be used to verify our proposed circuit mechanism of sampling-based inference. One possible experimental approach would modulate the functional recurrent strength by using a perceptual learning task. Specifically, after using a reference stimulus set with a prescribed correlation between latent stimuli to fully train an animal, we expect that recurrent synaptic weights will strengthen or weaken to improve inference (Fig. 8E, dashed line). This will result in a fixed value of differential noise correlations in the population response due to the recurrent circuitry. Re-training with a stimulus set that has more (less) correlated latent stimuli compared to the reference set will cause the recurrent weights to increase (decrease) (Fig. 8E, red line). When the reference stimulus set is again used to drive task behavior, then performance (as a proxy of mutual information) will decrease, regardless of whether differential correlations have increased or decreased compared to those resulting from the reference stimulus set (Fig. 8E, arrows). In brief, the non-monotonic relationship between differential noise correlations and the mutual information between stimulus and responses which support Bayesian inference offers a clear (and falsifiable) experimental prediction.

Extensions of circuit-based Bayesian inference

Implementing sampling-based inference in our proposed network requires that feedforward and recurrent inputs have the same tuning profile over the stimulus (Eq. 5). This assumption is supported by experiments in layers 4 and 2/3 in mouse V1 [8]. Moreover, the recurrent connections in our network model are translation-invariant in the stimulus subspace, an assumption widely used in continuous attractor networks (CAN) [22, 51, 54, 55]. Translation-invariant connections simplify the mathematical analysis, but are not required for a circuit to implement sampling. Adding randomness in recurrent connectivity only increases the variance of the sampling distribution. In the past, CANs have been shown to achieve maximal likelihood estimation (point estimate) via template matching [15, 55, 56]. Here we have shown that a network with CAN-like structure and internally Poisson spiking variability is able to perform sampling-based Bayesian inference. In our network correlations in the stimulus prior are represented by the strength of recurrent synaptic activity, which implies that the (subjective) prior precision in the network increases with the feedforward input strength. To maintain a fixed prior in the network recurrent weights need to decrease with increased feedforward input strength which encodes the likelihood precision, $\Lambda_{\mathbf{f}}$ (Eq. 6). Therefore, the (subjective) prior stored in the network with fixed recurrent weights may differ from the objective stimulus prior in the world (Λ_s in Eqs. 3 and 7) with feedforward inputs of different strengths. This could be solved by short-term synaptic depression which decreases the synaptic

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

efficacy at increased neuronal firing rates [57]. On the other hand, since the proposed recurrent circuit is general, this result may explain the origin of inductive bias [58] or confirmation bias [59] in cortical processing. Another possibility is that the recurrent circuit represents a more complex generative model which better captures the statistical structure of natural stimuli [30, 32, 60]. We only considered sampling driven by spiking variability with a Fano factor of 1, while cortical responses often have Fano factors that differ from 1 [61, 62]. In the latter case, our theory can still work by changing the feedforward connection weight to compensate for the change in Fano factor, as suggested in a recent study [63].

To keep our exposition transparent we only presented models with minimal complexity. Our proposed network mechanism of sampling-based inference can be generalized to more complex generative models, since the assumption of Gaussianity (Eqs. 21 and 22) and the analytical expression in Eq. (24) are not essential, and several relaxed frameworks may be explored. First, similar networks can generate samples from other multi–dimensional distributions where the conditional distribution of each latent variable belongs to the linear exponential family [35, 36]. This could be done by changing the tuning functions of neurons to another appropriate profile, as the logarithm of tuning determines the type of sampling distribution (Eq. 1). When sampling from non-Gaussian distributions, the stimulus samples can be linearly read out with the weight determined by the tuning profile (i.e., $\mathbf{h}(s)$ in Eq. 1, [36]). Second, the tuning of recurrent inputs does not need to be the same as that of feedforward inputs. Instead the logarithm of recurrent input tuning can have a form of the conjugate prior with the likelihood conveyed by feedforward inputs. Third, the network model could also be used to infer the latent variables with a non-uniform marginal prior, if, for example, the preferred stimuli of neurons in the population are not distributed uniformly in the stimulus subspace [64]. Lastly, we considered only non-structured inhibition for simplicity. Structured inhibitory connections could modulate the position of excitatatory responses in the stimulus subspace, i.e., the mean of the conditional distribution. Such interplay between E and I neurons with structured inhibition has the potential to implement Hamiltonian sampling, where the I neurons represent the sample of auxiliary variables [33, 35].

In conclusion, we have shown that a recurrent circuit of neurons with Poisson spiking statistics can implement sampling from a family of multivariate posterior distributions, with internal spiking variability driving the generation of stimulus samples, and the recurrent connections representing the stimulus prior. The proposed neural code may help us understand the structure of neuronal activity, provide a building blocks for more complicated population computations.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

A linear network of excitatory neurons

We study how a generic recurrent network model consisting solely of N_E excitatory (E) neurons with Poisson spiking statistics (no inhibitory neurons) can implement sampling-based Bayesian inference to approximate the stimulus posterior. We describe neuronal activity using a time–discretized Hawkes process (a type of multivariate, inhomogeneous Poisson process [65]). The instantaneous firing rates of the neurons in the network at time t, $\boldsymbol{\lambda}_t$, obey the following recurrent equations:

$$\boldsymbol{\lambda}_t \Delta t = \mathbf{u}^{\mathsf{f}} + \mathbf{u}^{\mathsf{r}}_t = \mathbf{u}^{\mathsf{f}} + (w_E \mathbf{r}_{t-\Delta t} + \sigma_r \boldsymbol{\xi}_t)\,, \tag{10}$$

$$\mathbf{r}_t \sim \prod_{j=1}^{N_E} \text{Poisson}\,(\boldsymbol{\lambda}_{tj} \Delta t)\,, \tag{11}$$

where $\mathbf{u}^{\mathsf{f}}$ is the feedforward Poisson spiking input (described below; Eq. 18), $\mathbf{u}^{\mathsf{r}}_t$ is the continuous valued recurrent input at time t, and $\boldsymbol{\xi}_t$ is a N_E dimensional independent Gaussian white noise. Hence, over each time interval $[t - \Delta t, t]$ the activity of the neurons in the network is modeled by a vector of independently generated Poisson spike counts, $\mathbf{r}_t$, with means determined by the rates $\boldsymbol{\lambda}_t$. The parameters w_E and σ_r determine the excitatory recurrent weight and recurrent variability, respectively.

Poisson spike generation samples stimulus

Independent Poisson spike generation in the network whose activity is described by Eq. (11) can drive sampling across time or across trials from a conditional stimulus distribution determined by the instantaneous firing rate $\boldsymbol{\lambda}_t$. Below we compute the distribution of stimulus samples given $\boldsymbol{\lambda}_t$. We assume that the instantaneous firing rate, $\boldsymbol{\lambda}_t$, has a smooth bell-shaped profile and can be parameterized as,

$$\boldsymbol{\lambda}_{tj} = R \exp[-(\bar{s}_t - \theta_j)^2 / 2a^2] = R \exp[\mathbf{h}_j(\bar{s}_t)], \tag{12}$$

where $\bar{s}_t$ characterizes the position of the population firing rate on the stimulus subspace (Fig. 1B, x-axis), while R and a denote the height and width of the population firing rate, respectively. Further, θ_j is the preferred stimulus value of neuron j, and the preferred stimuli of all neurons, $\{\theta_j\}_{j=1}^{N_E}$, are uniformly distributed over the range of stimulus s (Fig. 1B).

To simplify the analysis, we first assume that the instantaneous firing rate is fixed over time. When generating Poisson spikes $\mathbf{r}_t$ from $\boldsymbol{\lambda}_t$, the probability of observing a stimulus sample $\tilde{s}_t$

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(embedded in $\mathbf{r}_t$) can be derived as (see details in Supplemental Information),

$$\begin{aligned} p(\mathbf{r}_t|\boldsymbol{\lambda}_t) &= \textstyle\prod_{j=1}^{N_E} \text{Poisson}\left(\mathbf{r}_{tj}|\boldsymbol{\lambda}_{tj}\Delta t\right), \\ &\propto \exp[\mathbf{h}(\bar{s}_t)^\top \mathbf{r}] \cdot \left[n_{\boldsymbol{\lambda}}^{n_\mathbf{r}} \exp(-n_{\boldsymbol{\lambda}})\right], \\ &\propto \mathcal{N}\left(\tilde{s}_t|\bar{s}_t, a^2 n_\mathbf{r}^{-1}\right) \text{Poisson}(n_\mathbf{r}|n_{\boldsymbol{\lambda}}), \end{aligned} \tag{13}$$

where $n_\mathbf{r} = \sum_j \mathbf{r}_{tj}$ is the number of emitted spikes across the whole neural population, and $n_{\boldsymbol{\lambda}} = \sum_j \langle \boldsymbol{\lambda}_j \rangle \Delta t$ is the sum of population firing rate. Here $\mathcal{N}(s|\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2, and $\mathbf{h}(\bar{s}_t)$ is a vector with the j^{th} element as $\mathbf{h}_j(\bar{s}_t)$ shown in Eq. (12). The logarithm of the firing rate profile, $\mathbf{h}(\bar{s}_t)$, determines how the stimulus sample $\tilde{s}_t$ and its mean, $\bar{s}_t$, can be read out respectively from $\mathbf{r}_t$ and $\boldsymbol{\lambda}_t$,

$$\tilde{s}_t = \textstyle\sum_j \mathbf{r}_{tj}\theta_j / \sum_j \mathbf{r}_{tj}, \quad \bar{s}_t = \sum_j \boldsymbol{\lambda}_{tj}\theta_j / \sum_j \boldsymbol{\lambda}_{tj}, \tag{14}$$

where $\tilde{s}_t$ and $\bar{s}_t$ characterizes the position of $\mathbf{r}_t$ and $\boldsymbol{\lambda}_t$ on the stimulus subspace.

The sampling variability of $\tilde{s}_t$ in a single time step depends on the number of emitted spikes, $n_\mathbf{r}$. When the fixed rates, $\boldsymbol{\lambda}_t$, repeatedly generate spikes over time, the sampling distribution of $\tilde{s}_t$ can be calculated by marginalizing the likelihood (Eq. 13, last line) over different values of $n_\mathbf{r}$ since $n_\mathbf{r}$ varies across time (detailed calculation by using Laplacian approximation can be seen in Supplemental Information),

$$\begin{aligned} p(\tilde{s}_t|\boldsymbol{\lambda}_t) &= \textstyle\sum_{n_\mathbf{r}} \mathcal{N}\left(\tilde{s}_t|\bar{s}_t, a^2 n_\mathbf{r}^{-1}\right) \text{Poisson}(n_\mathbf{r}|n_{\boldsymbol{\lambda}}), \\ &\approx \mathcal{N}\left(\tilde{s}_t|\bar{s}_t, a^2 n_{\boldsymbol{\lambda}}^{-1}\right). \end{aligned} \tag{15}$$

Each stimulus sample, $\tilde{s}_t$, is thus drawn from a conditional distribution determined by the instantaneous firing rate, $p(\tilde{s}|\boldsymbol{\lambda}_t)$, and can be written as

$$\tilde{s}_t \sim p(\tilde{s}|\boldsymbol{\lambda}_t) = \mathcal{N}\left(\tilde{s}|\bar{s}_t, a^2 n_{\boldsymbol{\lambda}}^{-1}\right) \propto \exp[\mathbf{h}(\tilde{s})^\top \boldsymbol{\lambda}_t]. \tag{16}$$

The last proportionality in the above equation is satisfied by a Gaussian profile in the firing rate (more general derivation can be found in Supplemental Information). Introducing $\Lambda = a^{-2} n_{\boldsymbol{\lambda}}$ gives Eq. (1) shown in the main text.

Eq. (16) suggests that the type of sampling distribution (or the conditional distribution) that is obtained from spike generation variability is determined by the profile of the instantaneous firing rate, i.e., $\mathbf{h}(\bar{s}_t)$ (Eq. 12). Although the sampling distribution belongs to the linear exponential family of distributions which is similar with the probabilistic population code (PPC) [36], there are different ways in representing these distributions. In PPCs the likelihood over $\bar{s}_t$ is parametrically represented by a single realization of independent neuronal response $\mathbf{r}$ (Eq. 13), while in our work

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the distribution is approximated by a sequence of samples, $\tilde{s}_t$, effectively generated by conditionally independent Poisson spike discharges.

The above analysis can be extended to the case where the instantaneous firing rate, $\boldsymbol{\lambda}_t$, in a time step deviates from a smooth Gaussian profile (Eq. 12), which is the case in the actual network simulations. In general, $\boldsymbol{\lambda}_t$ can be expressed as,

$$\boldsymbol{\lambda}_{tj} = R_t \exp[\mathbf{h}_j(\bar{s}_t)] + \delta_\perp \boldsymbol{\lambda}_{tj}, \tag{17}$$

where $\delta_\perp \boldsymbol{\lambda}_t$ denotes the deviation from a smooth Gaussian profile. Note that the sampling distribution only depends on the position, $\bar{s}_t$, and the sum of instantaneous firing rate, $n_{\boldsymbol{\lambda}}$ (Eq. 16), which corresponds to two perpendicular directions in the N_E dimensional space of $\boldsymbol{\lambda}_t$. For any instantaneous firing rate vector, $\boldsymbol{\lambda}_t$, we can always find $\bar{s}_t$ and R_t that make the deviation $\delta_\perp \boldsymbol{\lambda}_t$ perpendicular to the two directions, i.e., $\sum_j \delta_\perp \boldsymbol{\lambda}_{tj} \theta_j = 0$, and $\sum_j \delta_\perp \boldsymbol{\lambda}_{tj} = 0$. This observation imples that deviations from Gaussian firing rate profiles do not affect our theory.

Feedforward spiking input conveys the likelihood of stimulus

We model the feedforward inputs to the E neurons in the network, $\mathbf{u}^{\mathsf{f}}$, as independent Poisson spikes, with Gaussian tuning over stimulus s,

$$\begin{aligned} p(\mathbf{u}^{\mathsf{f}}|s) &= \textstyle\prod_{j=1}^{N_E} \text{Poisson}\big[\mathbf{u}^{\mathsf{f}}_j \big| \langle \mathbf{u}^{\mathsf{f}}_j(s) \rangle\big], \\ \langle \mathbf{u}^{\mathsf{f}}_j(s) \rangle &= U^{\mathsf{f}} \exp[\mathbf{h}_j(s)] = U^{\mathsf{f}} \exp[-(\theta_j - s)^2/2a^2]. \end{aligned} \tag{18}$$

Here $\mathbf{u}^{\mathsf{f}}_j$ denotes the feedforward input received by the j^{th} E neuron, and $\langle \mathbf{u}^{\mathsf{f}}_j(s) \rangle$ is the tuning of the feedforward input. This mathematical description of feedforward input is the same as the one used in the definition of typical PPCs [15, 36, 37]. Since the preferred stimulus values, $\{\theta_j\}_{j=1}^{N_E}$, of all feedforward inputs are uniformly distributed in stimulus space then the likelihood of s given a single observation of the input, $\mathbf{u}^{\mathsf{f}}$, satisfies [36, 37],

$$\begin{aligned} p(\mathbf{u}^{\mathsf{f}}|s) &\propto \exp\big[\mathbf{h}(s)^\top \mathbf{u}^{\mathsf{f}}\big], \\ &\propto \mathcal{N}\big(s|\mu_{\mathsf{f}}, \Lambda_{\mathsf{f}}^{-1}\big). \end{aligned} \tag{19}$$

The logarithm of tuning, $\mathbf{h}(s)$, determines the type of likelihood [15]. Specifically, the Gaussian tuning leads to a Gaussian likelihood (Eq. 19), whose mean, μ_{f}, and precision, Λ_{f}, are both linear functions of the inputs,

$$\mu_{\mathsf{f}} = n_{\mathsf{f}}^{-1} \textstyle\sum_j \mathbf{u}^{\mathsf{f}}_j \theta_j, \quad \Lambda_{\mathsf{f}} = a^{-2} n_{\mathsf{f}} = a^{-2} \textstyle\sum_j \mathbf{u}^{\mathsf{f}}_j. \tag{20}$$

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The mean, μ_{f}, represents the position of $\mathbf{u}^{\mathsf{f}}$ in stimulus subspace, and the precision, Λ_{f}, is proportional to the sum of total feedforward spike counts, n_{f}.

A recurrent network samples hierarchical latent variables

A hierarchical generative model

We consider a hierarchical generative model for which inference can be implemented in a recurrent circuit of Poisson neurons. We extend the simple generative model of feedforward input (Eq. 19) by considering the stimulus s to depend on a one dimensional context variable, z. For simplicity, we assume that z follows a uniform distribution (Fig. 3B, marginal plots)

$$p(z) = \mathcal{U}(-180^\circ, 180^\circ), \tag{21}$$

where $\mathcal{U}(a, b)$ denotes a uniform distribution over $[a, b]$. The assumption of a uniform prior, $p(z)$, simplifies our model significantly, as it implies the spatial homogeneity of the network model as given by Eqs. (18-19). However, this assumption is not essential for our main results. Due to the differences between the stimulus (local) and context (global) aspects of the sensory scene, the stimulus, s, is not identical to the context z, but we assume that the two are correlated, so that

$$p(s|z, \Lambda_s) = \mathcal{N}\left(s|z, \Lambda_s^{-1}\right). \tag{22}$$

In sum, the whole generative model is determined by,

$$\begin{aligned} p(\mathbf{u}^{\mathsf{f}}, s, z) &= p(\mathbf{u}^{\mathsf{f}}|s)p(s|z)p(z), \\ &\propto \mathcal{N}\big(s|\mu_{\mathsf{f}}, \Lambda_{\mathsf{f}}^{-1}\big)\mathcal{N}\big(s|z, \Lambda_s^{-1}\big), \end{aligned} \tag{23}$$

where $p(\mathbf{u}^{\mathsf{f}}|s)$ is the same as in Eq. (19).

Approximate Bayesian inference via Gibbs sampling

The joint posterior of stimulus and context can be analytically derived given the generative model (Eq. 23),

$$\begin{aligned} p(s, z|\mathbf{u}^{\mathsf{f}}) &= \mathcal{N}\big[(s, z)^\top|\boldsymbol{\mu}_p, \mathbf{K}_p^{-1}\big], \\ \boldsymbol{\mu}_p &= (\mu_{\mathsf{f}}, \mu_{\mathsf{f}})^\top, \quad \mathbf{K}_p = \begin{pmatrix} \Lambda_{\mathsf{f}} + \Lambda_s & -\Lambda_s \\ -\Lambda_s & \Lambda_s \end{pmatrix}. \end{aligned} \tag{24}$$

We will use this expression to verify that the samples produced by our algorithm converge to ththe output of the algorithm.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We use the stochastic response of our recurrent network (Eqs. 10-11), as a basis for Gibbs sampling [31, 35, 41] (a type of Monte Carlo method) to approximate the joint posterior of stimulus, s, and context, z. To describe the iterative Gibbs algorithm, we assume that a context sample, $\tilde{z}_t$, is provided at time t, which is then combined with the feedforward input to update the conditional distribution of stimulus s (step 1 in Fig. 3C),

$$\begin{aligned} p(\tilde{s}|\tilde{z}_t, \mathbf{u}^{\mathsf{f}}) \propto p(\mathbf{u}^{\mathsf{f}}|\tilde{s}) p(\tilde{s}|\tilde{z}_t) \propto \mathcal{N}\left(s|\bar{s}_t, \Lambda^{-1}\right), \\ \bar{s}_t = \frac{\Lambda_{\mathsf{f}}\mu_{\mathsf{f}} + \Lambda_s \tilde{z}_t}{\Lambda_{\mathsf{f}} + \Lambda_s}, \quad \Lambda = \Lambda_{\mathsf{f}} + \Lambda_s. \end{aligned} \tag{25}$$

The next step in the algorithm is to draw a sample, $\tilde{s}_t$, from the conditional distribution $p(\tilde{s}|\tilde{z}_t, \mathbf{u}^{\mathsf{f}})$ (step 2 in Fig. 3C),

$$\tilde{s}_t \sim p(\tilde{s}|\tilde{z}_t, \mathbf{u}^{\mathsf{f}}) = \mathcal{N}\left(\tilde{s}|\bar{s}_t, \Lambda^{-1}\right).$$

Next, the conditional distribution of context, z, is updated given this new sample, $\tilde{s}_t$, and a new sample, $\tilde{z}_{t+\Delta t}$, is drawn (step 3 in Fig. 3C),

$$\tilde{z}_{t+\Delta t} \sim p(\tilde{z}|\tilde{s}_t) = \mathcal{N}(\tilde{z}|\tilde{s}_t, \Lambda_s^{-1}). \tag{26}$$

These three steps in the Gibbs sampling algorithm (Eqs. 25-26) are performed iteratively until sufficiently many samples, $\tilde{s}_t$ and $\tilde{z}_t$, are generated to approximate the true posterior distribution with sufficient accuracy (Fig. 3D; compare the red dots with the blue contour map).

Implementing the Gibbs sampling in a recurrent circuit model

Gibbs sampling of the stimulus (Eq. 4b) can be implemented via independent Poisson spike generation, as long as the conditional distribution encoded in $\boldsymbol{\lambda}_t$ (Eq. 16) is the same as the conditional distribution in the Gibbs sampling algorithm (Eq. 4a), i.e., $\ln p(\tilde{s}|\boldsymbol{\lambda}_t) = \mathbf{h}(\tilde{s})^\top \boldsymbol{\lambda}_t = \ln p(\tilde{s}|\tilde{z}_t, \mathbf{u}^{\mathsf{f}})$. This condition can be realized in the recurrent circuit by relating the expressions describing the neural dynamics (Eq. 10) and those describing the Gibbs sampling distribution (Eq. 4a) to yield,

$$\begin{aligned} \ln p(\tilde{s}|\tilde{z}_t, \mathbf{u}^{\mathsf{f}}) &= \mathbf{h}(\tilde{s})^\top \boldsymbol{\lambda}_t, \\ &= \mathbf{h}(\tilde{s})^\top \mathbf{u}^{\mathsf{f}} + \mathbf{h}(\tilde{s})^\top \mathbf{u}^{\mathsf{r}}_t, \\ &= \ln p(\mathbf{u}^{\mathsf{f}}|\tilde{s}) + \ln p(\tilde{s}|\tilde{z}_t). \end{aligned} \tag{27}$$

The generative model for the feedforward input $\mathbf{u}^{\mathsf{f}}$ (Eq. 19) suggests that $\ln p(\mathbf{u}^{\mathsf{f}}|\tilde{s}) = \mathbf{h}(\tilde{s})^\top \mathbf{u}^{\mathsf{f}}$. Hence to satisfy Eq. (27) we require

$$\ln p(\tilde{s}|\tilde{z}_t) = \mathbf{h}(\tilde{s})^\top \mathbf{u}^{\mathsf{r}}_t, \tag{28}$$

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

which implies that the recurrent input, $\mathbf{u}^{\mathsf{r}}_t$, should approximately have a Gaussian profile,

$$\begin{aligned} \mathbf{u}^{\mathsf{r}}_{tj}(\tilde{z}_t) &= U^{\mathsf{r}} \exp[-(\theta_j - \tilde{z}_t)^2/2a^2] + \delta_{\perp}\mathbf{u}^{\mathsf{r}}_{tj}, \\ \tilde{z}_t = \textstyle\sum_j \mathbf{u}^{\mathsf{r}}_{tj}\theta_j / \sum_j \mathbf{u}^{\mathsf{r}}_{tj}, &\quad \Lambda_s = a^{-2}\textstyle\sum_j \mathbf{u}^{\mathsf{r}}_{tj}, \end{aligned} \tag{29}$$

whose position on the stimulus subspace is $\tilde{z}_t$, and the sum of input (height) is determined by Λ_s, the precision of conditional distribution $p(s|\tilde{z}_t)$. In a similar fashion to Eq. (17), $\delta_{\perp}\mathbf{u}^{\mathsf{r}}_t$ denotes the deviation from a smooth Gaussian and is perpendicular to the direction of $\tilde{z}_t$ and Λ_s.

The optimal recurrent weight can be derived by combining Eq. (29) and Eq. (17). We notice the recurrent input, $\mathbf{u}^{\mathsf{r}}$, and neuronal responses, $\mathbf{r}_t$, have the same tuning width, a, in a network with only E neurons. This can only be achieved if E neurons are only self-connected (Eq. 10), as lateral connection broaden their tuning. The optimal recurrent weight generating recurrent input with appropriate strength is then,

$$w^*_E = \frac{\langle \mathbf{u}^{\mathsf{r}}_j \rangle}{\langle \mathbf{r}_j \rangle} = \frac{\sum_j \langle \mathbf{u}^{\mathsf{r}}_j \rangle}{\sum_j \langle \mathbf{r}_j \rangle} = \frac{\sum_j \langle \mathbf{u}^{\mathsf{r}}_j \rangle}{\sum_j \left(\langle \mathbf{u}^{\mathsf{f}}_j \rangle + \langle \mathbf{u}^{\mathsf{r}}_j \rangle \right)} = \frac{\Lambda_s}{\Lambda_{\mathsf{f}} + \Lambda_s}, \tag{30}$$

which yields Eq. (6) in the main text. Note that the self-connection is a result of the simplifying assumption that the network consists solely of E neurons (Eq. 10), which can be relaxed in a full network consisting both E and I neurons as we show below.

The sampling of the context variable (Eq. 4c) can be implemented through variability in the recurrent input. To do this, we include diffusive term in the recurrent interactions, $\mathbf{u}^{\mathsf{r}}_t$, and we equate the variance of the fluctuations with the mean to mimic a Poisson distribution:

$$\mathbf{u}^{\mathsf{r}}_t = \bar{\mathbf{u}}^{\mathsf{r}}_t + \sqrt{[\bar{\mathbf{u}}^{\mathsf{r}}_t]_+}\boldsymbol{\xi}_t, \quad \bar{\mathbf{u}}^{\mathsf{r}}_t = w^*_E \mathbf{r}_{t-\Delta t}, \tag{31}$$

where $[\cdot]_+$ denotes negative rectification. Here $\boldsymbol{\xi}_t$ is a N_E dimensional Gaussian white noise with $\langle \boldsymbol{\xi}_t(i)\boldsymbol{\xi}_{t'}(j)\rangle = \delta_{ij}\delta(t-t')$, δ_{ij} and $\delta(t-t')$ are Kronecker and Dirac delta functions respectively, $\bar{\mathbf{u}}^{\mathsf{r}}_t$ represents the conditional distribution $p(\tilde{z}|\tilde{s}_{t-\Delta t})$, and $\mathbf{u}^{\mathsf{r}}_t$ represent a context sample $\tilde{z}_t$ (Eq. 29). The multiplicative variability on recurrent interaction may come from synaptic noise [66, 67].

Coupled circuits sample a multi-dimensional posterior

We consider a generative model which has multiple latent stimuli, $\mathbf{s} = (s_1, s_2, \cdots, s_m)$, which are organized in parallel (Fig. 6A). Without loss of generality, we consider the simplest case where $m = 2$, and the same mechanism can be straightforwardly extended to any $m > 2$. We assume the

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

joint prior of $\mathbf{s}$ is a multivariate normal distribution,

$$p(\mathbf{s}) = \mathcal{N}(\mathbf{s}|\boldsymbol{\mu}_s, \boldsymbol{\Lambda}_s^{-1}) \propto \exp[-\Lambda_s(s_1 - s_2)^2/2], \quad \text{with} \quad \boldsymbol{\Lambda}_s = \Lambda_s \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \tag{32}$$

and each stimulus s_m is uniformly distributed in $(-180°, 180°]$ with periodic boundary imposed. The definition of Gaussian distribution in a circular space works well as long as the variance of the distribution is much smaller than the range of stimulus space. Here $\boldsymbol{\Lambda}_s$ is the precision matrix, while the scalar variable Λ_s ($\Lambda_s \geq 0$) characterizes the correlation between s_1 and s_2. Note that the covariance matrix $\boldsymbol{\Lambda}_s^{-1}$ is not defined, and the prior (Eq. 32) is improper. The mean, $\boldsymbol{\mu}_s$, is a free parameter, because it doesn't appear in the detailed expression of the prior (Eq. 32), which is a consequence from the zero determinant of the precision matrix, i.e., $|\boldsymbol{\Lambda}_s| = 0$. A further consequence is that the prior is not centered at $\boldsymbol{\mu}_s$, but instead has a band structure along the diagonal, and the marginal prior of each stimulus feature $p(s_m)$ ($m = 1, 2$) is uniform (Fig. 6B). The uniform marginal prior simplifies our theoretical derivation as it implies the spatial homogeneity of the network model but doesn't impact the proposed neural coding mechanism.

Each stimulus s_m ($m = 1, 2$) individually generates feedforward spiking input $\mathbf{u}_m^{\mathsf{f}}$, whose likelihood $p(\mathbf{u}_m^{\mathsf{f}}|s_m)$ is exactly the same as Eq. (2). Combined together, the generative model is

$$\begin{aligned} p(\mathbf{u}^{\mathsf{f}}|\mathbf{s})p(\mathbf{s}) &= \Big[\prod_{m=1}^{2} p(\mathbf{u}_m^{\mathsf{f}}|s_m)\Big] p(s_1, s_2), \\ &\propto \Big[\prod_{m=1}^{2} \mathcal{N}(s_m|\mu_{\mathsf{f}m}, \Lambda_{\mathsf{f}m}^{-1})\Big] \mathcal{N}(\mathbf{s}|\boldsymbol{\mu}_s, \boldsymbol{\Lambda}_s^{-1}), \\ &\propto \mathcal{N}(\mathbf{s}|\boldsymbol{\mu}_{\mathsf{f}}, \boldsymbol{\Lambda}_{\mathsf{f}}^{-1}) \mathcal{N}(\mathbf{s}|\boldsymbol{\mu}_s, \boldsymbol{\Lambda}_s^{-1}), \end{aligned} \tag{33}$$

where $\boldsymbol{\mu}_{\mathsf{f}} = (\mu_{\mathsf{f}1}, \mu_{\mathsf{f}2})^\top$, and the likelihood precision matrix $\boldsymbol{\Lambda}_{\mathsf{f}} = \mathrm{diag}(\Lambda_{\mathsf{f}1}, \Lambda_{\mathsf{f}2})$ is a diagonal matrix.

Gibbs sampling of the multi-dimensional posterior in a coupled neural circuit

Given the generative model (Eq. 33), the joint posterior of s_1 and s_2 is a bivariate normal distribution, i.e., $p(\mathbf{s}|\mathbf{u}^{\mathsf{f}}) = \mathcal{N}\big(\mathbf{s}|\boldsymbol{\mu}_p, \mathbf{K}_p^{-1}\big)$, whose precision matrix $\mathbf{K}_p$ and the mean $\boldsymbol{\mu}_p$ are,

$$\mathbf{K}_p = \boldsymbol{\Lambda}_{\mathsf{f}} + \boldsymbol{\Lambda}_s, \quad \boldsymbol{\mu}_p = \mathbf{K}_p^{-1} \boldsymbol{\Lambda}_{\mathsf{f}} \boldsymbol{\mu}_{\mathsf{f}}. \tag{34}$$

The precision matrix of the posterior is the sum of the precision of the likelihood and the prior, implying increased reliability of the distribution after combining with the prior. Meanwhile, the posterior mean is the weighted average of the means of the two likelihoods, with the weight proportional to the precision of each likelihood. We use this expression for the posterior to evaluate the performance of the proposed sampling-based algorithm.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Using Gibbs sampling to approximate the posterior (Eq. 34) involves the following steps:

$$\text{Compute}: \quad p(\tilde{s}_1|\mathbf{u}_1^{\mathsf{f}}, \tilde{s}_{2,t-\Delta t}) \propto p(\mathbf{u}_1^{\mathsf{f}}|\tilde{s}_1)p(\tilde{s}_{2,t-\Delta t}|\tilde{s}_1), \tag{35a}$$

$$\text{Sample}: \quad \tilde{s}_{1t} \sim p(\tilde{s}_1|\mathbf{u}_1^{\mathsf{f}}, \tilde{s}_{2,t-\Delta t}). \tag{35b}$$

We note that we only describe the sampling from the posterior distribution of s_1; as samples from the posterior of s_2 can be obtained similarly after exchanging indices. This sampling can be implemented in a neural circuit model consisting of several coupled networks, in which each network generates samples from the posterior distribution of the corresponding stimulus. Therefore the number of networks in the coupled circuit equals the dimension of the latent stimuli. The dynamics of the coupled neural circuit is defined by:

$$\boldsymbol{\lambda}_{1t} = \mathbf{u}_1^{\mathsf{f}} + \mathbf{u}_{12,t}^{\mathsf{r}} = \mathbf{u}_1^{\mathsf{f}} + w_{12}\mathbf{r}_{2,t-\Delta t}, \tag{36}$$

$$\mathbf{r}_{1t} \sim \prod_{j=1}^{N_E} \text{Poisson}(\boldsymbol{\lambda}_{1t,j}), \tag{37}$$

We again note the dynamics of network 2 can be similarly obtained by changing indices. To implement Gibbs sampling (Eqs. 35a-35b) in the coupled circuit (Eqs. 36-37), spike generation in network 1 (Eq. 37) can be used to produce stimulus samples, $\tilde{s}_{1t}$, when the conditional distribution determined by $\boldsymbol{\lambda}_{1t}$ matches the conditional distribution required in the definition of Gibbs sampling (Eq. 35a), i.e., $\ln p(\tilde{s}_1|\mathbf{u}_1^{\mathsf{f}}, \tilde{s}_{2,t-\Delta t}) = \ln p(\tilde{s}_{1t}|\boldsymbol{\lambda}_{1t}) = \mathbf{h}(\tilde{s}_1)^\top \boldsymbol{\lambda}_{1t}$. Taking the logarithm of Eq. (35a) yields,

$$\ln p(\tilde{s}_1|\mathbf{u}_1^{\mathsf{f}}, \tilde{s}_{2,t-\Delta t}) = \ln p(\mathbf{u}_1^{\mathsf{f}}|\tilde{s}_1) + \ln p(\tilde{s}_{2,t-\Delta t}|\tilde{s}_1). \tag{38}$$

Comparing this expression with Eq. (36), we see that the feedforward input, $\mathbf{u}_1^{\mathsf{f}}$, matches the conditional distribution $p(\mathbf{u}_1^{\mathsf{f}}|\tilde{s}_1)$ (Eq. 33). We therefore require the recurrent input from network 2 to network 1 to encode the conditional distribution $p(\tilde{s}_{2,t-\Delta t}|\tilde{s}_1)$, i.e., $\ln p(\tilde{s}_{2,t-\Delta t}|\tilde{s}_1) = \mathbf{h}(\tilde{s}_1)^\top \mathbf{u}_{12,t}^{\mathsf{r}}$. This implies that $\mathbf{u}_{12,t}^{\mathsf{r}}$ should approximately have a Gaussian profile,

$$\begin{aligned} \mathbf{u}_{12,tj}^{\mathsf{r}} &= U_{12}^{\mathsf{r}} \exp[-(\theta_j - \tilde{s}_{t-\Delta t})^2/2a^2] + \delta_\perp \mathbf{u}_{12,tj}^{\mathsf{r}}, \\ \tilde{s}_{2,t-\Delta t} &= \textstyle\sum_j \mathbf{u}_{12,tj}^{\mathsf{r}} \theta_j / \sum_j \mathbf{u}_{12,tj}^{\mathsf{r}}, \quad \Lambda_s = a^{-2} \sum_j \mathbf{u}_{12,tj}^{\mathsf{r}}, \end{aligned} \tag{39}$$

where $\delta_\perp \mathbf{u}_{12,tj}^{\mathsf{r}}$ quantifies the deviation from a perfect Gaussian profile, and does not affect the decoded value $\tilde{s}_{2,t-\Delta t}$ and Λ_s.

The recurrent input, $\mathbf{u}_{12}^{\mathsf{r}}$, (Eq. 39) has the same width a as the neuronal response, $\mathbf{r}_1$. In circuit containing only E neurons, if the two networks have the same number of neurons, then across networks only neurons having the same preferred stimulus should be connected. The optimal

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

recurrent weight between two networks is then

$$w_{mn} = \frac{\langle \mathbf{u}^{\mathsf{r}}_{mn,j} \rangle}{\langle \mathbf{r}_{nj} \rangle} = \frac{\sum_j \langle \mathbf{u}^{\mathsf{r}}_{mn,j} \rangle}{\sum_j \langle \mathbf{r}_{nj} \rangle} = \frac{\Lambda_s}{\Lambda_s + \Lambda^{\mathsf{f}}_n}, \quad (m \neq n) \tag{40}$$

Since each network individually generate a stimulus sample, the sample of stimulus m can be locally read out from network m's responses even if the activities of two networks are correlated (Fig. 6A), which greatly simplifies readout. Furthermore, due to the population firing rate of each network has Gaussian profile, the stimulus sample $\tilde{s}_{mt}$ can be linearly read out from $\mathbf{r}_{mt}$ as

$$\tilde{s}_{mt} = \textstyle\sum_j \theta_j \mathbf{r}_{mt,j} / \sum_j \mathbf{r}_{mt,j}. \tag{41}$$

We note that the circuit implementation of Gibbs sampling from a multi-dimensional posterior (Eq. 8a) does not require the recurrent connections between E neurons within a network. This is due to the assumption that the marginal priors of each stimulus feature, $p(s_m)$, are uniform. For a non-uniform marginal prior $p(s_m)$, recurrent connections between E neurons within a network would be required for generating samples from a distribution that matches the true posterior.

Inference from an information-theoretic point of view

The goal of the sampling algorithm is to approximate the posterior distribution of a latent variables, Θ, given a feedforward input, $\mathbf{u}^{\mathsf{f}}$. Specifically, the latent variables $\Theta = \{s, z\}$ in the hierarchical generative model (Eq. 23), or $\Theta = \mathbf{s} = \{s_1, s_2\}$ in the generative model with breadth (Eq. 33). When the sampling algorithm uses an internal model which does not match the structure of the generative model, the sampling distribution $q(\Theta|\mathbf{u}^{\mathsf{f}})$ will differ from the true posterior, $p(\Theta|\mathbf{u}^{\mathsf{f}})$ (Eq. 24). In this case the mutual information between the sampling distribution of the latent variables, Θ, and $\mathbf{u}^{\mathsf{f}}$ will be smaller than in the case when samples come from the true posterior, $p(\Theta|\mathbf{u}^{\mathsf{f}})$,

$$\begin{aligned} I(\Theta, \mathbf{u}^{\mathsf{f}}) &= -\mathbb{E}_{p(\Theta)}[\log p(\Theta)] + \mathbb{E}_{p(\Theta,\mathbf{u}^{\mathsf{f}})}[\log p(\Theta|\mathbf{u}^{\mathsf{f}})] \\ &\geq -\mathbb{E}_{p(\Theta)}[\log p(\Theta)] + \mathbb{E}_{p(\Theta,\mathbf{u}^{\mathsf{f}})}[\log q(\Theta|\mathbf{u}^{\mathsf{f}})] \equiv I_q(\Theta; \mathbf{u}^{\mathsf{f}}), \end{aligned} \tag{42}$$

It is straightforward to show that the difference between $I(\Theta, \mathbf{u}^{\mathsf{f}})$ and $I_q(\Theta, \mathbf{u}^{\mathsf{f}})$ is the Kullback-Leibler (KL) divergence between p and q, i.e., $D_{KL}[p||q] = I(\Theta, \mathbf{u}^{\mathsf{f}}) - I_q(\Theta, \mathbf{u}^{\mathsf{f}}) = \mathbb{E}_p(\ln p - \ln q) \geq 0$. Equality in Eq. (42) holds only if the distribution q matches the true posterior p.

The mutual information $I_q(\Theta; \mathbf{u}^{\mathsf{f}})$ can be computed analytically when the approximating distribution $q(\Theta|\mathbf{u}^{\mathsf{f}}) = \mathcal{N}\left(\Theta|\boldsymbol{\mu}_q, \mathbf{K}_q^{-1}\right)$ is a bivariate normal (substituting Eqs. 23 and 24 into Eq. 42),

$$I_q(\Theta; \mathbf{u}^{\mathsf{f}}) = \log L + \frac{1}{2}\left[1 + \log \frac{|\mathbf{K}_q|}{2\pi\Lambda_s} - \mathrm{tr}(\mathbf{K}_q \mathbf{K}_p^{-1}) - (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^\top \mathbf{K}_q (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)\right]. \tag{43}$$

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Here $L = 360°$ is the length of the stimulus feature subspace, while $\boldsymbol{\mu}_p$ and $\mathbf{K}_p$ are the mean and the precision matrix of the posterior distribution (Eqs. 24 or 34). When q matches the posterior distribution, p, we have, $I(\Theta; \mathbf{u}^{\mathsf{f}}) = \log L - \frac{1}{2}[1 + \log(2\pi\Lambda_s) - \log|\mathbf{K}_p|]$.

The neuronal response distribution conditioned on external stimulus

We compute the distribution of neuronal responses $\mathbf{r}$ over time/trial in response to an external stimulus s, i.e., $p(\mathbf{r}|s)$, in order to find a neural signature of network sampling and compare it with experimental data. For a fixed external stimulus s, the neuronal response $\mathbf{r}$ fluctuates due to both sensory transmission noise described by $p(\mathbf{u}^{\mathsf{f}}|s)$ (Eq. 18), as well as the internally generated variability described by $p(\mathbf{r}|\mathbf{u}^{\mathsf{f}})$ (Fig. 4A). Therefore, the distribution of $\mathbf{r}$ in response to an external stimulus s has the form

$$p(\mathbf{r}|s) = \int p(\mathbf{r}|\mathbf{u}^{\mathsf{f}}) p(\mathbf{u}^{\mathsf{f}}|s) d\mathbf{u}^{\mathsf{f}}.$$

For simplicity, we only compute the covariability of $p(\mathbf{r}|\mathbf{u}^{\mathsf{f}})$ along the stimulus subspace (Fig. 1B, x-axis), because the covariability along other directions is not related with stimulus sampling. By approximating the Poissonian spiking variability $p(\mathbf{r}|\boldsymbol{\lambda})$ with a multivariate normal distribution (Eq. 11), and considering the limit of weak fluctuations in $\boldsymbol{\lambda}$ along the stimulus subspace over time, $p(\mathbf{r}|\mathbf{u}^{\mathsf{f}})$ can be computed approximately as (see math details in Supplemental Information),

$$\begin{aligned} p(\mathbf{r}|\mathbf{u}^{\mathsf{f}}) &= \int p(\mathbf{r}|\boldsymbol{\lambda}) p(\boldsymbol{\lambda}|\mathbf{u}^{\mathsf{f}}) d\boldsymbol{\lambda}, \\ &\approx \mathcal{N}\big[\mathbf{r}\big|\mathbf{f}(s), \mathrm{diag}(\mathbf{f}(s)) + V(\bar{s}|\mu_{\mathsf{f}}) \mathbf{f}'_s \mathbf{f}'^{\top}_s\big], \quad \text{where } s = \mu_{\mathsf{f}}. \end{aligned} \tag{44}$$

$\mathbf{f}(s) = \langle \boldsymbol{\lambda}_t \rangle$ denotes the temporally averaged population response. The covariance structure of the neuronal response includes two terms: $\mathrm{diag}(\mathbf{f}(s))$, a diagonal matrix whose entries equal that of the vector $\mathbf{f}(s)$ denoting the (independent) Poisson spiking variability (Eq. 23), and $V(\bar{s}|\mu_{\mathsf{f}}) \mathbf{f}'_s \mathbf{f}'^{\top}_s$, a term that captures the covariability due to firing rate fluctuations along the stimulus subspace (Fig. 8A), where $\mathbf{f}'_s = d\mathbf{f}(s)/ds$ is the derivative of $\mathbf{f}(s)$ over the stimulus feature s. The covariance $\mathbf{f}'_s \mathbf{f}'^{\top}_s$ is often termed differential (noise) correlations [4, 17]. With the Gaussian profile of $\mathbf{f}(s)$ (Eqs. 18 and 29), $\mathbf{f}'_s \mathbf{f}'^{\top}_s$ exhibits anti-symmetric structure (Fig. 8B) [17, 22, 50, 68, 69].

In Eq. (44), $V(\bar{s}|\mu_{\mathsf{f}})$ is the variance of $\bar{s}_t$ (the mean of conditional distribution in Eq. 4a) over time and characterizes the amplitude of internally generated differential correlations. In network implementation, $\bar{s}_t$ and μ_{f} are represented as the position of $\boldsymbol{\lambda}_t$ and $\mathbf{u}^{\mathsf{f}}$ on the stimulus subspace respectively (Eqs. 14 and 20). The dynamics of Gibbs sampling (Eq. S20 in Supplemental Information) and the network structure (Eq. 6) imply that

$$V(\bar{s}|\mu_{\mathsf{f}}) = \frac{\Lambda_s}{\Lambda_{\mathsf{f}}(\Lambda_{\mathsf{f}} + \Lambda_s)} = a^2 n_{\mathsf{f}}^{-1} w_E^*. \tag{45}$$

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Note that $V(\bar{s}|\mu_{\mathsf{f}})$ is constrained by network connections, in that it is internally generated and shared within the network (for $w_E^* > 0$).

An expression for $p(\mathbf{r}|s)$ can be derived similarly, and includes an additional term contributing to differential correlations compared with $p(\mathbf{r}|\mathbf{u}^{\mathsf{f}})$ (Eq. 44) due to fluctuations in the feedforward inputs,

$$\begin{aligned} p(\mathbf{r}|s) &\approx \mathcal{N}\big[\mathbf{r}\big|\mathbf{f}(s), \mathrm{diag}(\mathbf{f}(s)) + V(\bar{s}|s)\mathbf{f}'_s\mathbf{f}'^{\top}_s\big], \\ V(\bar{s}|s) &= V(\bar{s}|\mu_{\mathsf{f}}) + V(\mu_{\mathsf{f}}|s) = \frac{\Lambda_s}{\Lambda_{\mathsf{f}}(\Lambda_{\mathsf{f}}+\Lambda_s)} + \frac{1}{\Lambda_{\mathsf{f}}} = a^2 n_{\mathsf{f}}^{-1}(w_E^*+1). \end{aligned} \tag{46}$$

Here the variance, $V(\bar{s}|s)$, in the stimulus feature subspace is a mixture of internal variability, $V(\bar{s}|\mu_{\mathsf{f}})$, and sensory noise, $V(\mu_{\mathsf{f}}|s)$ (Eq. 23). The neuronal response distribution in coupled networks (Fig. 6A) can be obtained similarly (see the Supplemental Information).

A spiking network model with excitatory and inhibitory Poisson neurons

To test the proposed inference mechanisms in a network consisting of E neurons (Eqs. 10-37), we simulated a well studied recurrently coupled cortical model [21, 22]. The network consisted of N_E excitatory (E) and N_I inhibitory (I) spiking neurons, with the activity of each neuron modeled as a Hawkes process [65]. At time t, we represent the response of neuron j in population $a = \{E, I\}$, $\mathbf{r}^a_{tj}$, as a spike count drawn from a Poisson distribution with instantaneous firing rate, $\boldsymbol{\lambda}^a_{tj}$,

$$\mathbf{r}^a_{tj} \sim \mathrm{Poisson}\left[\boldsymbol{\lambda}^a_{tj}\right]. \tag{47}$$

Each neuron has a refractory period of 2ms after emitting a spike. The firing rate $\boldsymbol{\lambda}^a_{tj}$ is the sum of feedforward input $\mathbf{u}^{af}_{tj}$ and recurrent input $\mathbf{u}^{ar}_{tj}$, so that $\boldsymbol{\lambda}^a_{tj} = \mathbf{u}^{af}_{tj} + \mathbf{u}^{ar}_{tj}$. The feedforward inputs are filtered spikes from upstream neurons, $\mathbf{u}^{af}_{tj} = \sum_n \eta\big(t - t^{\mathsf{f}}_{jn}\big)$, where t^{f}_{jn} is the time of the n^{th} spike received by neuron j of population a from the feedforward inputs. Here $\eta(t)$ is the synaptic input profile which is modeled as $\eta(t) = \exp(-t/\tau_d)/\tau_d$, $(t > 0)$. Throughout, we set the synaptic time constant $\tau_d = 2$ms. To mimic the Poisson-like variability to sample a context in a hierarchical generative model (Eqs. 23 and 31), the recurrent input received by neuron j in population a is defined by

$$\begin{aligned} \mathbf{u}^{ar}_{tj} &= \bar{\mathbf{u}}^{ar}_{tj} + \sqrt{[\bar{\mathbf{u}}^{ar}_{tj}]_+}\xi_t, \\ \bar{\mathbf{u}}^{ar}_{tj} &= \sum_{b=\{E,I\}} \sum_{k=1}^{N_b} \frac{J^{ab}_{jk}}{\sqrt{N}} \sum_n \eta(t - t^b_{kn}), \end{aligned} \tag{48}$$

where $\bar{\mathbf{u}}^{ar}_{tj}$ is the mean recurrent input at time t given the neuronal activities of the presynaptic neurons. The recurrent input in the network is corrupted by noise whose variance equals the mean

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of the recurrent input. In a physiological network, recurrent noise may be generated by the chaotic state in network dynamics [70] or synaptic noise [66, 67]. In Eq. (48) the function $[\cdot]_+$ rectifies the negative input, and ξ_t is a random variable following a standard Gaussian distribution. The coefficient J_{ij}^{ab} is the synaptic weight from neuron j in population b to neuron i in population a. The time t_{kn}^b is the time of the n^{th} spike fired by neuron k in population b. The parameter $N = N_E + N_I$ is the total number of neurons in the network. The scaling of the synaptic weights by $1/\sqrt{N}$ is standard in networks where excitation is balanced by recurrent inhibition [70]. Finally, the synaptic input profile of the recurrent input, $\eta(t)$, is the same as the one we chose for the feedforward input for convenience. Note that the rectification in Eq. (48) on recurrent inputs will introduce errors resulting in deviations of the sampling distribution from the true posterior, and hence we chose the recurrent weights to be small (Fig. 5). The rectification only arises when using (continuous) recurrent inputs to sample the context variable, and doesn't impact the generality of sampling by (discrete) Poisson spiking variability.

To model the coding of a circular stimulus such as orientation, the excitatory neurons are arranged on a ring [22, 68]. The preferred stimuli, θ_j, of the excitatory neurons are equally spaced on the interval $(-180°, 180°]$, consistent with the range of latent features (Eq. 21). Inhibitory neurons are not tuned to stimulus, and their role is to stabilize network responses. Note that the recurrent connections between E neurons are modeled using a Gaussian function decaying with the distance between the stimuli preferred by the two cells, rather than only self-connection in the simple network with only E neurons (Eqs. 30),

$$J_{jk}^{EE} = \frac{w_{EE} L}{\sqrt{2\pi} a} \exp\left[-\frac{(\theta_j - \theta_k)^2}{2a^2}\right], \tag{49}$$

We imposed periodic boundaries on the Gaussian function to avoid boundary effect in simulations. Although in the generative model we assumed non-periodic feature variables (Eq. 3), as long as the variance of the associated distributions are smaller than the width of the feature space, the network model with periodic boundaries on the recurrent connection (Eq. 49) provides a good approximation of the non-periodic Gaussian posterior (Eq. 24). The weight w_{EE} denotes the average connection strength of all E to E connections. The parameter $a = 40°$ defines the footprint of connectivity in feature space (i.e the ring), and $L = 360°$ is the length of the ring manifold (Eq. 21); Multiplication by L in Eq. (49) sets the sum of all E to E connection strengths equal to $N_E w_{EE}$. Moreover, the excitatory and inhibitory neurons are all-to-all connected with each other (similar for I to I connections). For simplicity, we consider the E to I, I to I and I to E connections all to be unstructured (in feature space) and assume that connections of the same type have equal weight, i.e., $J_{jk}^{EI} = w_{EI}$, $J_{jk}^{IE} = w_{EE}$ and $J_{jk}^{II} = w_{II}$. To simplify the network further, we consider the connections from the same population of neurons to have the same average weight, i.e., $w_{EE} = w_{IE} \equiv w_E$ and $w_{II} = w_{EI} \equiv w_I$. For the feedforward network model shown in Fig. 2,

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

we only remove the E recurrent connections between E neurons, i.e., $w_{EE} = 0$, while keeping other connections, including w_{EI}, w_{II}, and w_{IE}, the same as the recurrent network.

The feedforward inputs applied to E neurons consist of independent Poisson spike counts as described by Eq. (18), with rate $\langle \mathbf{u}_j^{E\mathsf{f}}(s)\rangle = U^{\mathsf{f}} e^{-(s-\theta_j)^2/(4a^2)}$. The inhibitory neurons also receive feedforward indpendent Poissonian inputs. The firing rate of the input received by every I neuorn is proportional to the overall feedforward rate of input to E neurons, in order to keep the excitatory and inhibitory balance of neuronal activities in the network,

$$\langle \mathbf{u}_j^{I\mathsf{f}}\rangle = \frac{w_{I\mathsf{f}}}{N_I}\sum_{j=1}^{N_E}\langle \mathbf{u}_j^{E\mathsf{f}}(s)\rangle. \tag{50}$$

In the simulations, we started with a network of $N_E = 180$ excitatory and $N_I = 45$ inhibitory neurons, and increased the number of neurons by a fixed factor in Fig. 1D. The ratio between the average connection from I neurons and the one from E neurons was kept fixed with $w_I/w_E = 5$. We set the feedforward weight of input to I neurons to $w_{I\mathsf{f}} = 0.8$. We simulated the dynamics of the model network using the Euler method with a time step of 0.1ms. The typical parameters used in simulation can be found in Table 1 in Supplemental Information. Further details about the simulations and numerical estimates of mutual information and linear Fisher information are also presented in Supplemental Information. The code of network simulation was written in MATLAB 2018b, and can be found at GitHub (https://github.com/wenhao-z/Sampling_PoissSpk_Neuron).

A spiking network model of coupled neural circuits

In the coupled neural circuits used to infer latent variables organized in parallel (Fig. 6A) the two networks are copies of each other, i.e., the two networks have the same intrinsic parameters. Each network is equivalent to the one described in the previous section, except that there is no recurrent connections between E neurons in the same network, and no variability in recurrent interactions (no noise in Eq. 48). The absence of recurrent connections between E neurons in the same network is due to the uniform marginal prior of stimulus. Nevertheless, in the same network the E and I neurons are connected using the same connection profile as above to keep network activity stable. Between the two networks, there are only E connections which target both E and I neurons. The connections between E neurons across networks have the same pattern as that given described by Eq. (49) with the peak connection strength from network n to network m denoted as w_{EE}^{mn}. The connections from E neurons in one network to I neurons in the other is set to the same as the peak strength of E connections across networks for simplicity, i.e., $w_{IE}^{mn} = w_{EE}^{mn}$. To simplify the network model further, we set the inter-network connections to be symmetric, which means $w_{EE}^{mn} = w_{EE}^{nm}$. In the simulations w_{EE}^{mn} was adjusted to determine how the sampling distribution is affected (Fig. 7A).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Comparing the sampling distribution with posterior in coupled neural circuits

We read out the samples from the posterior distribution of each stimulus, $\tilde{s}_{mt}$, individually from the spiking activities of E neurons, $\mathbf{r}_{mt}$, in network m in every time window of 20ms by using a population vector. We used this collection of samples to estimate the mean, $\langle \tilde{\mathbf{s}} \rangle = (\langle \tilde{s}_1 \rangle, \langle \tilde{s}_2 \rangle)^\top$, and covariance matrix, $\mathbf{\Sigma_s}$, of the sampling distribution. Meanwhile, the mean $\boldsymbol{\mu}_\mathbf{f}$ and precision matrix $\mathbf{\Lambda}_\mathbf{f}$ of the likelihood are linearly read out from the feedforward inputs fed into the network model (Eq. 33).

If the sampling distribution is comparable with the posterior, the sampling mean $\langle \tilde{\mathbf{s}} \rangle$ and covariance $\mathbf{\Sigma_s}$ should satisfy Eq. (34). We use the actual sampling covariance and the likelihood parameters to predict the sampling mean, i.e., $\langle \tilde{\mathbf{s}} \rangle_{\text{pred}} = \mathbf{\Sigma_s \Lambda_f \mu_f}$, and compare it with the actual $\langle \tilde{\mathbf{s}} \rangle$ (Fig. 7D-F). To obtain the posterior precision matrix, given the sampling mean $\langle \tilde{\mathbf{s}} \rangle$ and the likelihood parameters, we vary the single parameter of prior precision Λ_s to minimize the KL divergence from the prediction of posterior by using the value of Λ_s, and the actual sampling distribution. Given this value of Λ_s, the prediction of posterior precision is computed as $\mathbf{K}_{\text{pred}} = \mathbf{\Lambda}_s + \mathbf{\Lambda}_\mathbf{f}$ (Eq. 34) which is then compared with actual sampling precision matrix ($\mathbf{\Sigma_s}^{-1}$; see Fig. 7C-G). The prior precision, Λ_s, is a *subjective* prior, which reflects the prior stored in the recurrent network and may change with input (see Discussion). More details of network simulation and parameters can be found in Supplemental Information.

Acknowledgements

National Institutes of Health grants 1R01MH115557 (K.J.), 1U19NS107613-01 (B.D.), R01EB026953 (B.D.); National Science Foundation grant NSF-DBI-1707400 (K.J.). Vannevar Bush faculty fellowship N00014-18-1-2002 (B.D); Simons Foundation Collaboration on the Global Brain (B.D.).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

[1] Alexandre Pouget, Peter Dayan, and Richard S Zemel. Inference and computation with population codes. *Annual Review of Neuroscience*, 26(1):381–410, 2003.

[2] Brent Doiron, Ashok Litwin-Kumar, Robert Rosenbaum, Gabriel K Ocker, and Krešimir Josić. The mechanics of state-dependent neural correlations. *Nature neuroscience*, 19(3):383–393, 2016.

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

[4] Adam Kohn, Ruben Coen-Cagli, Ingmar Kanitscheider, and Alexandre Pouget. Correlations and neuronal population information. *Annual review of neuroscience*, 39:237–256, 2016.

[5] Julie A Harris, Stefan Mihalas, Karla E Hirokawa, Jennifer D Whitesell, Hannah Choi, Amy Bernard, Phillip Bohn, Shiella Caldejon, Linzy Casal, Andrew Cho, et al. Hierarchical organization of cortical and thalamic connectivity. *Nature*, 575(7781):195–202, 2019.

[6] Seung Wook Oh, Julie A Harris, Lydia Ng, Brent Winslow, Nicholas Cain, Stefan Mihalas, Quanxin Wang, Chris Lau, Leonard Kuan, Alex M Henry, et al. A mesoscale connectome of the mouse brain. *Nature*, 508(7495):207–214, 2014.

[7] Rodney J Douglas and Kevan AC Martin. Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.*, 27:419–451, 2004.

[8] L Federico Rossi, Kenneth D Harris, and Matteo Carandini. Spatial connectivity matches direction selectivity in visual cortex. *Nature*, 588(7839):648–652, 2020.

[9] Kenneth D Harris and Thomas D Mrsic-Flogel. Cortical connectivity and sensory coding. *Nature*, 503(7474):51–58, 2013.

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

[11] Alan Yuille and Daniel Kersten. Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7):301–308, 2006.

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

[13] Konrad P Körding and Daniel M Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–247, 2004.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

964 [14] David C Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural
965 coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, 2004.

966 [15] Alexandre Pouget, Jeffrey M Beck, Wei Ji Ma, and Peter E Latham. Probabilistic brains:
967 knowns and unknowns. *Nature neuroscience*, 16(9):1170, 2013.

968 [16] József Fiser, Pietro Berkes, Gergő Orbán, and Máté Lengyel. Statistically optimal perception
969 and learning: from behavior to neural representations. *Trends in cognitive sciences*, 14(3):119–
970 130, 2010.

971 [17] Rubén Moreno-Bote, Jeffrey Beck, Ingmar Kanitscheider, Xaq Pitkow, Peter Latham, and
972 Alexandre Pouget. Information-limiting correlations. *Nature neuroscience*, 17(10):1410, 2014.

973 [18] Peter Dayan and Laurence F Abbott. *Theoretical neuroscience*, volume 806. Cambridge, MA:
974 MIT Press, 2001.

975 [19] Bruno B Averbeck, Peter E Latham, and Alexandre Pouget. Neural correlations, population
976 coding and computation. *Nature reviews neuroscience*, 7(5):358, 2006.

977 [20] Apostolos P Georgopoulos, Andrew B Schwartz, and Ronald E Kettner. Neuronal population
978 coding of movement direction. *Science*, 233(4771):1416–1419, 1986.

979 [21] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. The stabilized supralinear
980 network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*,
981 85(2):402–417, 2015.

982 [22] R Ben-Yishai, R Lev Bar-Or, and H Sompolinsky. Theory of orientation tuning in visual
983 cortex. *Proceedings of the National Academy of Sciences*, 92(9):3844–3848, 1995.

984 [23] David C Somers, Sacha B Nelson, and Mriganka Sur. An emergent model of orientation
985 selectivity in cat visual cortical simple cells. *Journal of Neuroscience*, 15(8):5448–5465, 1995.

986 [24] Chengcheng Huang, Alexandre Pouget, and Brent David Doiron. Internally generated popu-
987 lation activity in cortical networks hinders information transmission. *bioRxiv*, 2020.

988 [25] Daniel Kersten, Pascal Mamassian, and Alan Yuille. Object perception as bayesian inference.
989 *Annu. Rev. Psychol.*, 55:271–304, 2004.

990 [26] Kenji Doya, Shin Ishii, Alexandre Pouget, and Rajesh PN Rao. *Bayesian brain: Probabilistic*
991 *approaches to neural coding*. MIT press, 2007.

992 [27] Patrik O Hoyer and Aapo Hyvärinen. Interpreting neural response variability as monte carlo
993 sampling of the posterior. In *Advances in neural information processing systems*, pages 293–
994 300, 2003.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[28] Lars Buesing, Johannes Bill, Bernhard Nessler, and Wolfgang Maass. Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS computational biology*, 7(11):e1002211, 2011.

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

[30] 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.

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

[32] Rodrigo Echeveste, Laurence Aitchison, Guillaume Hennequin, and Máté Lengyel. Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *bioRxiv*, page 696088, 2020.

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

[34] Kenneth H Britten, Michael N Shadlen, William T Newsome, and J Anthony Movshon. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *Journal of Neuroscience*, 12(12):4745–4765, 1992.

[35] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[36] 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.

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

[38] 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, 1996.

[39] Eric R Kandel, James H Schwartz, Thomas M Jessell, Department of Biochemistry, Molecular Biophysics Thomas Jessell, Steven Siegelbaum, and AJ Hudspeth. *Principles of neural science*, volume 4. McGraw-hill New York, 2000.

[40] Michael S Lewicki and Terrence J Sejnowski. Bayesian unsupervised learning of higher order structure. *Advances in neural information processing systems*, pages 529–535, 1997.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[41] Agnieszka Grabska-Barwinska, Jeffrey M Beck, Alexandre Pouget, and Peter E Latham. Demixing odorsfast inference in olfaction. *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, 2013.

[42] David J Field, Anthony Hayes, and Robert F Hess. Contour integration by the human visual system: evidence for a local association field. *Vision research*, 33(2):173–193, 1993.

[43] Wilson S Geisler, Jeffrey S Perry, BJ Super, and DP Gallogly. Edge co-occurrence in natural images predicts contour grouping performance. *Vision research*, 41(6):711–724, 2001.

[44] Lee Cossell, Maria Florencia Iacaruso, Dylan R Muir, Rachael Houlton, Elie N Sader, Ho Ko, Sonja B Hofer, and Thomas D Mrsic-Flogel. Functional organization of excitatory synaptic strength in primary visual cortex. *Nature*, 518(7539):399–403, 2015.

[45] Ingmar Kanitscheider, Ruben Coen-Cagli, Adam Kohn, and Alexandre Pouget. Measuring fisher information accurately in correlated neural populations. *PLoS Comput Biol*, 11(6):e1004218, 2015.

[46] Tai Sing Lee. The visual system's internal model of the world. *Proceedings of the IEEE*, 103(8):1359–1378, 2015.

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

[48] Sabyasachi Shivkumar, Richard Lange, Ankani Chattoraj, and Ralf Haefner. A probabilistic population code based on neural samples. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[49] Ingmar Kanitscheider, Ruben Coen-Cagli, and Alexandre Pouget. Origin of information-limiting noise correlations. *Proceedings of the National Academy of Sciences*, 112(50):E6973–E6982, 2015.

[50] 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.

[51] Si Wu, KY Michael Wong, CC Alan Fung, Yuanyuan Mi, and Wenhao Zhang. Continuous attractor neural networks: candidate of a canonical model for neural information representation. *F1000Research*, 5, 2016.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[52] 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.

[53] Richard D Lange and Ralf M Haefner. Task-induced neural covariability as a signature of approximate bayesian learning and inference. *bioRxiv*, page 081661, 2020.

[54] Kechen Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *The Journal of Neuroscience*, 16(6):2112–2126, 1996.

[55] Sophie Deneve, Peter E Latham, and Alexandre Pouget. Reading population codes: a neural implementation of ideal observers. *Nature Neuroscience*, 2(8):740–745, 1999.

[56] Si Wu, Shun-ichi Amari, and Hiroyuki Nakahara. Population coding and decoding in a neural field: a computational study. *Neural Computation*, 14(5):999–1026, 2002.

[57] Misha Tsodyks, Klaus Pawelzik, and Henry Markram. Neural networks with dynamic synapses. *Neural computation*, 10(4):821–835, 1998.

[58] Eric Schulz, Joshua B Tenenbaum, David Duvenaud, Maarten Speekenbrink, and Samuel J Gershman. Compositional inductive biases in function learning. *Cognitive psychology*, 99:44–79, 2017.

[59] Richard D Lange, Ankani Chattoraj, Jeffrey Beck, Jacob Yates, and Ralf Haefner. A confirmation bias in perceptual decision-making due to hierarchical approximate inference. *bioRxiv*, page 440321, 2021.

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

[61] 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.

[62] Gaby Maimon and John A Assad. Beyond poisson: increased spike-time regularity across primate parietal cortex. *Neuron*, 62(3):426–440, 2009.

[63] Wenhao Zhang, Tai Sing Lee, Brent Doiron, and Si Wu. Distributed sampling-based bayesian inference in coupled neural circuits. *bioRxiv*, 2020.

[64] Deep Ganguli and Eero P Simoncelli. Implicit encoding of prior probabilities in optimal neural populations. *Advances in neural information processing systems*, 2010:658, 2010.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.26.477877; this version posted January 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[65] James Trousdale, Yu Hu, Eric Shea-Brown, and Krešimir Josić. Impact of network structure and cellular response on spike time correlations. *PLoS computational biology*, 8(3):e1002408, 2012.

[66] Dmitri A Rusakov, Leonid P Savtchenko, and Peter E Latham. Noisy synaptic conductance: Bug or a feature? *Trends in Neurosciences*, 2020.

[67] Robert Rosenbaum, Jonathan Rubin, and Brent Doiron. Short term synaptic depression imposes a frequency dependent filter on synaptic information transfer. *PLoS Comput Biol*, 8(6):e1002557, 2012.

[68] Si Wu, Kosuke Hamaguchi, and Shun-ichi Amari. Dynamics and computation of continuous attractors. *Neural Computation*, 20(4):994–1025, 2008.

[69] Klaus Wimmer, Duane Q Nykamp, Christos Constantinidis, and Albert Compte. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature neuroscience*, 17(3):431, 2014.

[70] C van Vreeswijk and Haim Sompolinsky. Chaotic balanced state in a model of cortical circuits. *Neural computation*, 10(6):1321–1371, 1998.