Recurrent circuit based neural population codes for stimulus representation and inference

Wen-Hao Zhang[1−3], Si Wu[4], Krešimir Josić[5,6*] Brent Doiron[1−3,7*]

[1]Department of Neurobiology, Grossman Center for Quantitative Biology and Human Behavior, University of Chicago, Chicago, IL, USA.

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

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

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

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

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

[7]Department of Statistics, University of Chicago, Chicago, IL, USA.

*Correspondence should be addressed to either Brent Doiron or Krešimir Josić.

Email: bdoiron@uchicago.edu or josic@math.uh.edu

Abstract

A large part of the synaptic input received by cortical neurons comes from local cortico-cortical connectivity. Despite their abundance, the role of local recurrence in cortical function is unclear, and in simple coding schemes it is often the case that a circuit with no recurrent connections performs optimally. We consider a recurrent excitatory-inhibitory circuit model of a cortical hypercolumn which performs sampling-based Bayesian inference to infer latent hierarchical stimulus features. We show that local recurrent connections can store an internal model of the correlations between stimulus features that are present in the external world. When the resulting recurrent input is combined with feedforward input it produces a population code from which the posterior over the stimulus features can be linearly read out. Internal Poisson spiking variability provides the proper fluctuations for the population to sample stimulus features, yet the resultant population variability is aligned along the stimulus feature direction, producing what are termed *differential correlations*. Importantly, the amplitude of these internally generated differential correlations is determined by the associative prior in the model stored in the recurrent connections, thus providing experimentally testable predictions for how population connectivity and response variability are connected to the structure of latent external stimuli.

Keywords: Sampling-based Bayesian inference, Recurrent circuit dynamics, Population variability, Differential correlations.

Introduction

In an uncertain and changing world, it is imperative for the brain to represent stimulus information reliably. A wrong decision based on a faulty stimulus representation could endanger an animal's life. It is widely believed that the cortex is essential for the representation of the sensory world [1, 2], with individual neurons tuned to a specific stimulus feature so that the population as a whole can code for a richly structured sensory scene. However, there are two central aspects of cortical circuits that remain to be properly integrated into a coherent framework describing how cortex represents stimuli. First, neuronal responses in sensory cortices are often noisy, with the same stimulus evoking starkly different responses across trials (e.g., [3–5]). On the surface such response variability would be detrimental to codes [2, 6], so it is a mystery as to why cortical response is so variable when isolated cortical neurons can respond very reliable [7]. Second, the primary source of synaptic inputs to cortical neurons is not from upstream centers which presumably carry sensory signals, but rather from recurrent pathways between neurons in the same hierarchical layer [8–10]. While such recurrent connections are organized about a stimulus feature axis [11, 12], it is not obvious how combining inputs that have already been processed will improve overall representation. Our study presents an inference coding scheme where these two realities of cortical circuits become essential features needed for optimal representation.

Numerous studies have shown that sensory processing within the cortex can be understood through a Bayesian inference framework (e.g., [13–19]). In such models, sensory cortex infers and synthesizes a description of the possible stimuli that could have given rise to the feedforward inputs received by sensory neurons, i.e. the posterior distribution over the stimulus in the world rather than a single estimate. To perform Bayesian inference, the sensory cortex needs to store a model describing how external stimuli and feedforward inputs are generated. Then, once a feedforward input is received, cortical dynamics inverts this internal model in a process sometimes termed 'analysis-by-synthesis' [15, 20]. The neural response thus represents a posterior distribution over the input stimulus, and unobserved (latent) variables [18, 19]. In our study, we propose that recurrent (local) connections in cortical circuits represent such an internal model (e.g., [21, 22]), and that internally generated variability from recurrent interactions is a source of fluctuations needed to perform sampling-based inference of the posterior distribution (e.g., [19, 23, 24]). The challenge in this approach is to determine how the structure of connectivity and population variability should be set so as to optimally represent the posterior distribution.

To tackle this challenge we consider a hierarchical generative model [15], where inputs are generated by a local stimulus feature and a global context feature. We assume that the local and

global aspects of a scene are correlated, so that knowledge of one provides information about the the other. Examples would be the direction of a breeze (global) and a leaf's movement in that breeze (local), or the orientation of a tree trunk (global) and orientation of bark texture on a section of the trunk (local). We propose a recurrent network model that interprets these inputs by implementing sampling-based inference over the stimulus and context features. In the network, the inferred posterior over the latent features is represented in a subspace of the neuronal population response, and can be linearly read out by a downstream decoder. Importantly, we find that the recurrent connections store the associative prior between stimulus and context features, and that internal Poissonian variability can drive sampling in the feature subspace. To generate the best approximation of the stimulus posterior the strength of recurrent connections increases with the degree of correlation between the global and local sensory inputs. In other words, a strong structure between sensory components of the external world requires a strongly coupled network to best infer that world, whereas no structure would require no recurrent connections.

Our network model provides testable experimental predictions about how the brain represents the structure of the world to interpret sensory information. We propose that a neural signature of network sampling is an internally generated response covariability along the direction of stimulus features, known as differential correlations [2, 25]. The amplitude of these internally generated differential correlations is determined by the associative prior between stimulus and context features in the internal model stored in the recurrent connections. In this way we can use differential noise correlations (which can be measured) as a proxy for strong recurrent connections organized along the stimulus tuning dimension (which are difficult to measure), all of which can be controlled by adjusting the structure of the external world.

Results

Recurrent circuitry does not improve conventional neural codes

Neural coding is often described as the process by which a stimulus feature, s, is encoded in neuronal activity, $\mathbf{r}$, which can then subsequently be decoded to provide an estimate of the feature, $\hat{s}$ (Fig. 1a, top) [26]. In this view, neuronal circuitry should be adapted to provide an accurate feature estimate. However, as we next show, recurrent circuitry degrades such estimates, raising the question of its purpose.

In classical models, stimulus encoding is performed by a neuronal population, where individual neurons are tuned to a preferred stimulus value and the preferred values across the population cover a range of potential stimuli [27–29] (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 the preferred stimuli are distributed on a ring (Fig. 1b, top). Neurons with similar feature preferences tend to

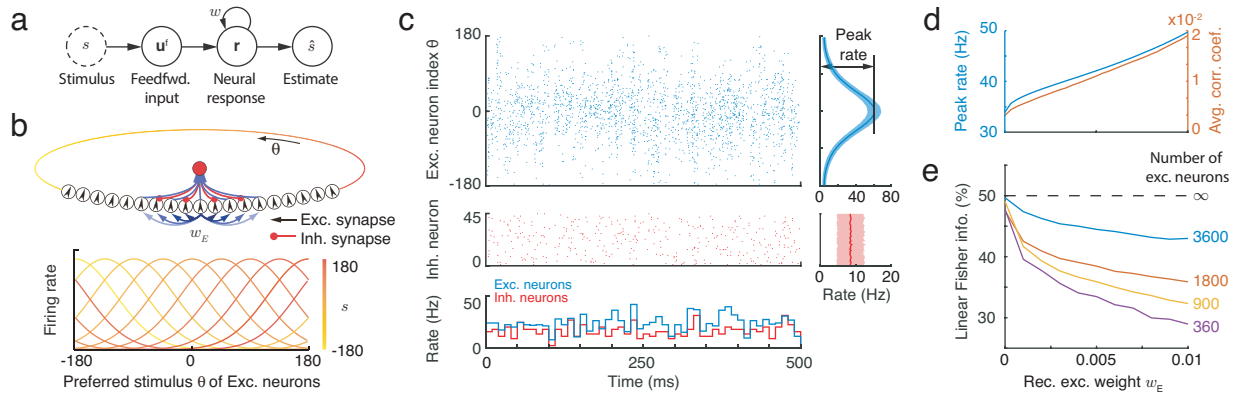

Figure 1: A recurrent network with structured recurrent connections limits Fisher information about external stimuli. (a) A schematic diagram showing how a stimulus, s, is encoded in a neuronal response, $\mathbf{r}$. A stimulus estimate $\hat{s}$ can be decoded from $\mathbf{r}$. (b) A recurrent ring model (top) with population activity of excitatory neurons, $\mathbf{r}$, dependent on a stimulus, s (bottom). Information about the stimulus can be obtained from the activity of E neurons. (c) Population activity of excitatory and inhibitory neurons in the network. Right: the firing rate of neurons averaged over time, with shaded region denoting standard deviation. (d-e) The population responses (d), and linear Fisher information of stimulus, s (e) change with peak excitatory connectivity strength, w_E. Linear Fisher information decreases as excitatory weights increase. Different lines in (e) represent networks with different numbers of excitatory neurons, with the ratio of excitatory to inhibitory neurons kept constant.

be coupled more strongly [11, 12, 30], so that network connectivity aligns with the ring topology (Fig. 1b, top, arrows). We assume structured connectivity between excitatory (E) neurons, and assume inhibition (I) is unstructured, acting only to stabilize network activity [31]. Hence, only E neurons in the network are tuned to a stimulus feature (Fig. 1b, bottom). Further, for simplicity we assume a linear (or linearized) neuronal transfer and synaptic interactions (Eqs. 21-22) and Poisson spike responses (Eqs. 34-38). This canonical, mechanistic network model has been widely used to explore cortical network dynamics and coding [31–33].

We consider the impact of recurrent connectivity on the accuracy of the decoded feature estimate, $\hat{s}$, using a conventional coding framework. The network is presented with a stimulus with feature $s = 0$, and as a result the activity of E neurons tuned to $s = 0$ is elevated (Fig. 1c). 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 (Fig. 1d). We measured the accuracy of the estimate $\hat{s}$ using the linear Fisher Information (LFI) obtained from the population response of excitatory neurons (Eq. 39). LFI provides a bound on the ability of an ideal observer to discriminate the change in a stimulus using a linear decoder [1, 2, 25, 27, 28], and it is closely related to an animal's performance in discrimination tasks widely studied in experiments [2, 27]. The LFI computed from the population response decreases with an increase in recurrent excitatory strength w_E (Fig. 1e), indicating that recurrent coupling degrades the best estimate obtainable from population activity. This is due to the increase in noise correlations coupled with the fact that the decoder has access to only a finite number of neurons. While the decrease in

LFI with excitatory coupling can be mitigated in the limit of large networks (Fig. 1e, dashed line), the information about s available to linear readout of population is then unaffected by recurrent coupling. Rather, the LFI saturates the bound set by the information about s entering the network along the feedforward pathway $\mathbf{u}^{\mathsf{f}}$ [25]. In total, recurrent connectivity does not enhance stimulus representation in terms of increasing Fisher information. Since synaptic coupling is biologically expensive, then a network that most accurately and cheaply represents stimulus information is one that does not have any recurrent connections (i.e., $w_E = 0$).

What is then the purpose of the extensive recurrent connections between neurons [8–10, 12]? There are at least two possibilities. First, while it does degrade stimulus representation, recurrent connectivity might underlie other stimulus processing functions essential for cortical computations. The second possibility is that the coding framework we explored (Fig. 1) is overly simplistic, and recurrent cortical coupling can improve cortical codes within a different framework. Here we develop an extended framework that shows that the second possibility is plausible, but does not exclude the first.

A hierarchical framework of information processing in recurrent circuit

Classical population codes provide a point estimate (i.e., a single value) of a latent stimulus s given input from a feedforward sensory pathway $\mathbf{u}^{\mathsf{f}}$ (Fig. 1a). However, numerous studies suggest that the brain may be performing Bayesian inference (e.g., [13–15, 17, 36, 37]), inferring and synthesizing a full description of every possible latent stimulus. To accomplish this task neural circuits have to approximate a posterior distribution over the latent stimulus from sensory input, $\mathbf{u}^{\mathsf{f}}$, and stored prior information. Such computations can be complex, and require more biological machinery than point estimation. Here we propose a circuit based implementation of Bayesian inference, paying particular attention to the role of recurrent wiring.

We consider a hierarchical generative model describing how a stimulus in the world generates a feedforward input $\mathbf{u}^{\mathsf{f}}$ received by a recurrent cortical circuit (Fig. 2a, from z to $\mathbf{u}^{\mathsf{f}}$). The generative model consists of two parts: an external stage describing the stimulus in the receptive field common to a population of neurons (Fig. 2a, from z to s), and an internal stage describing sensory transmission in the brain (Fig. 2a, from s to $\mathbf{u}^{\mathsf{f}}$). We will show how the recurrent circuit can store an internal model of this generative process, enabling the circuit to compute samples from the posterior distribution over a latent stimulus given sensory input.

In order for the computations in the circuit to be non-trivial, the stimulus has to come from a world with sufficiently rich structure. We therefore assume that the first stage of our model consists of a stimulus feature, s, and a context feature, z, both of which are one dimensional for simplicity. For example, in a V1 hypercolumn the stimulus feature, s, could be the orientation of the visual scene within a classical receptive field (local information) of a hypercolumn of V1 neurons, while

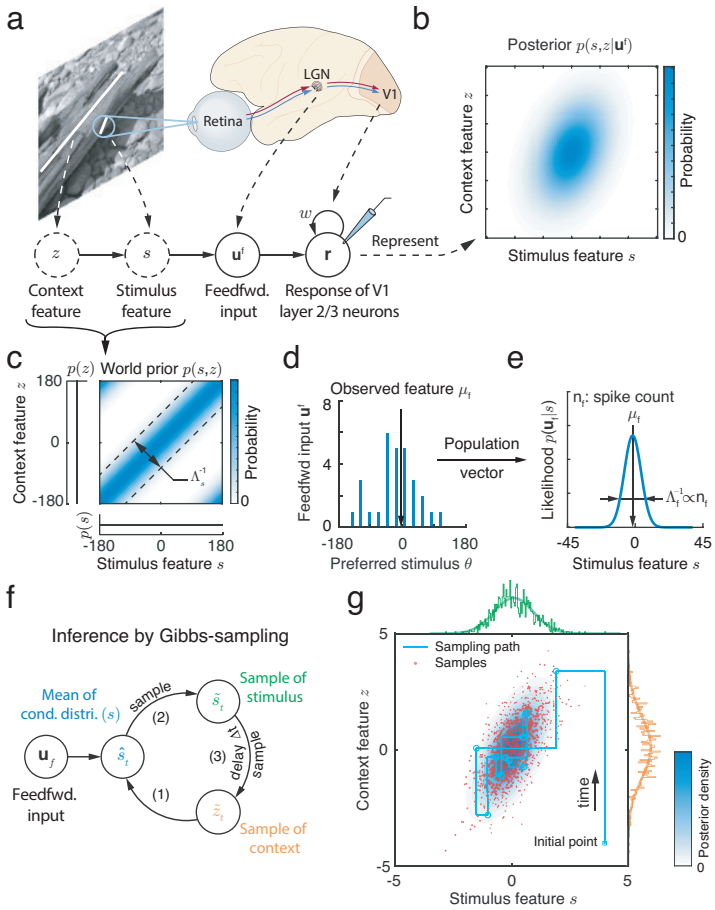

Figure 2: A hierarchical generative model and inference by Gibbs sampling. (a) An example of sensory feedforward input generation: The context feature, z, is the orientation of the tree trunk, and the stimulus feature, s, is the orientation of texture located in the classical receptive field of a V1 hypercolumn. Feedforward input, $\mathbf{u}^{\mathbf{f}}$, conveys stimulus information to V1 neurons which infer and represent the joint posterior of stimulus and context in the world, $p(s, z|\mathbf{u}^{\mathbf{f}})$ in evoked response, $\mathbf{r}$. Solid circles: observations and responses in the brain; dashed circles: latent variables in the world. (Natural image and brain schematic adapted from [34] and [35] respectively). (b) The joint posterior over stimulus and context can be represented by the neuronal response, $\mathbf{r}$. (c) The joint prior over the global context, z, and local stimulus, s, has more mass around the diagonal. The width of this band is controlled by Λ_s, and determined by the correlation between context and stimulus. (d) A realization of feedforward spiking inputs, $\mathbf{u}^{\mathbf{f}}$, defined as independent Poisson random variables with intensity dependent on stimulus, s. (e) The parameters of the likelihood function of s can be linearly read out from $\mathbf{u}^{\mathbf{f}}$ using a population vector (Eq. 12). (f) The posterior over context and stimulus can be obtained via Gibbs sampling (Eqs. 2a-2c) by iteratively generating samples of s and z from their respective conditional distributions. (g) The resulting approximations of the joint and marginal posterior over the latent stimulus feature, s, and context feature, z. Light blue contour: the posterior distribution (Eq. 14); red dots: Samples obtained using Gibbs sampling.

the orientation within a non-classical receptive field of these cells could describe the corresponding context feature z (Fig. 2a). We assume that the conditional distribution of the stimulus feature, for a given context, $p(s|z) = \mathcal{N}(s|z, \Lambda_s^{-1})$, is Gaussian with Λ_s denoting the precision (inverse of the variance), and that the prior distribution over the context feature, $p(z)$, is uniform (Fig. 2c) for simplicity. Although the uniform prior over the context feature, z, implies that the marginal prior

over s, is also uniform, the joint prior over stimulus and context features $p(s, z)$ has a non-trivial structure: As the two features are correlated, the joint prior density is concentrated around the diagonal (Fig. 2c). The width of the band is determined by the precision Λ_s, which measures how strongly the context feature, z, and the stimulus feature, s, are related.

The second stage of the model describes stimulus transmission via the primary sensory pathway (from s to $\mathbf{u}^{\mathsf{f}}$ in Fig. 2a). It is represented by the feedforward input, $\mathbf{u}^{\mathsf{f}}$, received by N_E excitatory (E) neurons in the recurrent circuit. The feedforward input, $\mathbf{u}^{\mathsf{f}}$, is a vector of independent Poisson spike counts (Eq. 10). The mean response of each component of $\mathbf{u}^{\mathsf{f}}$ has a Gaussian tuning profile over the stimulus feature s, i.e., $\langle \mathbf{u}^{\mathsf{f}}_j \rangle \propto \exp[\mathbf{h}_j(s)] = \exp[-(s-\theta_j)^2/2a^2]$. We assume that the preferred stimuli of the feedforward inputs, $\{\theta_j\}_{j=1}^{N_E}$, are uniformly distributed over the stimulus feature range, and that the widths of all tuning curves are equal to a constant, a [27, 38, 39]. We refer to the subspace spanned by the vector $(\theta_1, \theta_2, \ldots, \theta_{N_E})$ as the *stimulus feature subspace*. The Gaussian tuning of the input is consistent with the broad, bell-shaped orientation tuning in V1 (Fig. 1b, bottom), and conditional independence of the Poisson input is consistent with the observation that LGN and V1 simple cells remove redundancy in the input [34, 40].

Combining these two input stages provides a complete description of the feedforward input, $\mathbf{u}^{\mathsf{f}}$, (see details in Eq. 11),

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

Importantly, the inputs, $\mathbf{u}^{\mathsf{f}}$, represent a Gaussian likelihood over the stimulus feature, s [38, 39], i.e., $p(\mathbf{u}^{\mathsf{f}}|s) \propto \mathcal{N}(s|\mu_{\mathsf{f}}, \Lambda_{\mathsf{f}}^{-1})$. This is a consequence of Gaussian tuning, and the preferred stimuli of feedforward inputs are uniformly distributed. The likelihood of the stimulus is determined by the mean, μ_{f}, and precision, Λ_{f}, both of which can be read out linearly from the input, $\mathbf{u}^{\mathsf{f}}$ using a population vector [29, 38] (Eq. 12): The position (center of mass) of $\mathbf{u}^{\mathsf{f}}$ in the stimulus feature subspace (x-axis in Fig. 2e) represents the mean, μ_{f}, which corresponds to a noisy observation of stimulus feature (Fig. 2d-e). The precision, Λ_{f}, is proportional to the sum of input spike counts, $n_{\mathsf{f}} = \sum_j \mathbf{u}^{\mathsf{f}}_j$ (Eq. 12). We next ask how a recurrent neural circuit can infer the joint posterior distribution over the stimulus features s and z given sensory input $\mathbf{u}^{\mathsf{f}}$ (Fig. 2b).

Sampling based inference of the stimulus posterior

Given the generative model with the additional assumptions captured in Eq. 1, the joint posterior over stimulus and context features, $p(s, z|\mathbf{u}^{\mathsf{f}})$ is a multivariate normal distribution (See Eq. 14). In Bayesian theories, this posterior captures an observer's perception of the joint external stimulus [13,

14, 17, 36, 41–43]. Central to understanding such probabilistic computation is how recurrent circuits compute and represent the posterior, $p(s, z|\mathbf{u}^{\mathsf{f}})$ [1, 18, 19].

We propose that Gibbs sampling, a stochastic algorithm which uses sequential samples to approximate a probability distribution [24, 44, 45], can be used to compute the posterior over the features, s and z, given sensory input, $\mathbf{u}^{\mathsf{f}}$. Such sampling algorithms are well suited for cortical circuits given the prominent stochastic dynamics of cortical neurons [19, 23, 46–48]. Gibbs sampling starts with an initial guess for the value of the two latent features, and proceeds by alternately fixing the estimate of one latent feature, and updating the other by drawing a sample from the respective conditional distribution. More precisely, to approximate the posterior distributions over s and z, Gibbs sampling proceeds by generating a sequence of samples, $(\tilde{s}_t, \tilde{z}_t), t = 1, \ldots, T$. Given the generative structure of the stimulus described in the previous section, such sampling can be implemented through a recursive iteration of the following algorithm (Fig. 2f and Eq. (15)),

$$\text{Compute}: \quad p(s|\tilde{z}_t, \mathbf{u}^{\mathsf{f}}) \propto p(\mathbf{u}^{\mathsf{f}}|s)p(s|\tilde{z}_t) \equiv \mathcal{N}(s|\hat{s}_t, \Lambda^{-1}), \tag{2a}$$

$$\text{Sample}: \quad \tilde{s}_t \sim p(s|\tilde{z}_t, \mathbf{u}^{\mathsf{f}}), \tag{2b}$$

$$\text{Sample}: \quad \tilde{z}_{t+\Delta t} \sim p(z|\tilde{s}_t) = \mathcal{N}(z|\tilde{s}_t, \Lambda_s^{-1}), \tag{2c}$$

We assume that samples are generated over time, and Δt is the time step between samples. As the samples are generated by alternately fixing the values of the two variables, the sampling trajectories alternates between horizontal and vertical jumps in parameter space (cyan line in Fig. 2g). The set of sampled values defines a distribution, $q(s, z|\mathbf{u}^{\mathsf{f}}) = T^{-1} \sum_t \delta\left[(s, z)^\top - (\tilde{s}_t, \tilde{z}_t)^\top\right]$, where $\delta(\cdot)$ is the Dirac delta function, T is the number of samples, and $\top$ denotes vector transpose. The distribution $q(s, z|\mathbf{u}^{\mathsf{f}})$ (red dots in Fig. 2g) approximates the true posterior distributions $p(s, z|\mathbf{u}^{\mathsf{f}})$ (blue contour map in Fig. 2g, Eq. 14) [21, 44, 45]. Moreover, if we are only interested in one of the features, e.g., say s, we can record only sequential samples $\tilde{s}_t$ to obtain an approximating distribution $q(s|\mathbf{u}^{\mathsf{f}})$ of the marginal posterior over the stimulus feature, i.e., $p(s|\mathbf{u}^{\mathsf{f}})$ (compare the two green lines at the margin in Fig. 2g).

Bayesian inference in a recurrent network of spiking neurons

If the brain performs Bayesian inference it is important to understand how the requisite probabilistic computations are implemented in cortical circuits. We will show that the Gibbs algorithm can be naturally implemented in a recurrent network with internally generated variability providing samples of stimulus feature, s, and context feature, z in alternation. To demonstrate this we consider a simple linear network model consisting of only excitatory neurons (Fig. 3a, Eqs. 21-22). The responses of E neurons, $\mathbf{r}_t$, are number of spikes independently sampled from Poisson distribution with parameters given by the firing rates $\boldsymbol{\lambda}_t$. We assume that the firing rates are a

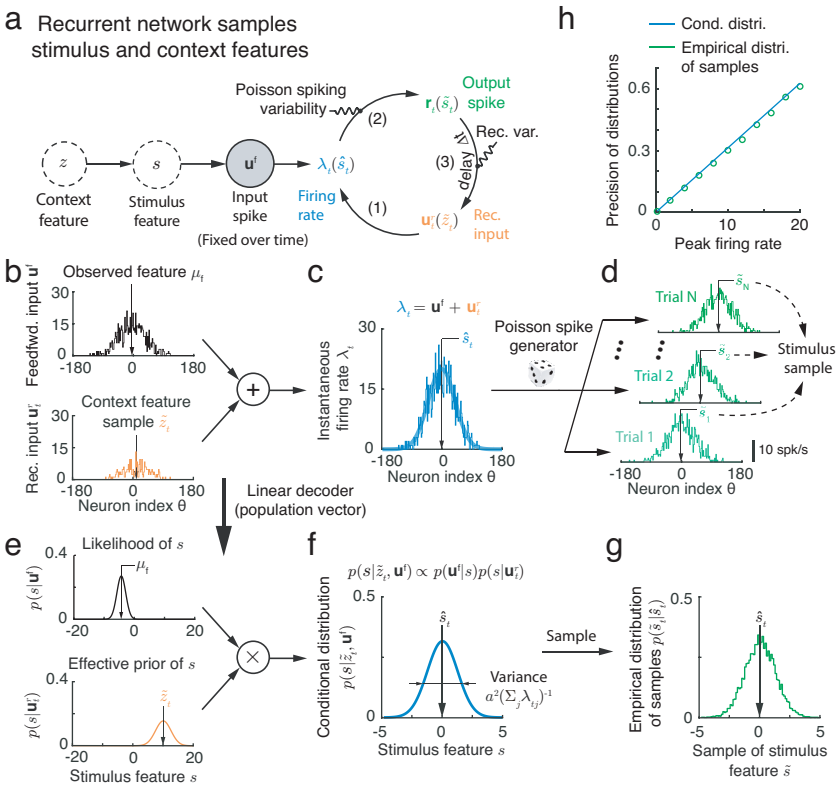

Figure 3: A recurrent circuit with internal Poisson variability can implement latent feature sampling. (a) A schematic of recurrent circuit dynamics, in which stimulus and context features are encoded respectively by the population response, $\mathbf{r}$, and recurrent inputs $\mathbf{u}^{\mathrm{r}}$. Poissonian spiking and recurrent variability drive the fluctuations needed for sampling stimulus and context features, respectively. (b-d) The sum of feedforward and recurrent inputs (b) yields the instantaneous firing rate (c), which determines the mean of independent Poisson spike counts (d). (e-g) The distributions and samples of stimulus features linearly read out from corresponding population inputs and responses shown in (b-d). The empirical distribution of samples (g) from Poisson spike generation in (d) is consistent with the conditional distribution (f) represented by instantaneous firing rate (c). (h) The peak firing rate determines the precision of conditional distribution of stimulus feature (f) and the variability of stimulus feature samples.

sum of the sensory feedforward input, $\mathbf{u}^{\mathrm{f}}$ (Eq. 1), and recurrent inputs $\mathbf{u}^{\mathrm{r}}_t$.

A biologically realistic implementation of Gibbs sampling in a recurrent network needs to resolve several fundamental questions. How does the recurrent circuit represent the generative model that needs to be inverted to infer the latent variables (Eq. 1)? How are the conditional distributions, and samples from the posterior over the features (Eqs. 2a-2c) represented in neural circuitry and activity? Random sampling requires internally generated variability whose statistics should be determined by conditional distributions that change with each sample (Eq. 2b-2c). Is there a universal and robust circuit mechanism to generate such variability? We address these questions next.

Representing the posterior in the recurrent network

We first show how recurrent connections in the network can produce an internal representation of the generative model: A sample of the stimulus feature, $\tilde{s}_t$, will be represented by the response of E neurons in the recurrent network, $\mathbf{r}_t$, while a sample of the context feature, $\tilde{z}_t$, will be represented by the recurrent inputs, $\mathbf{u}^{\mathsf{r}}_t$ (note the similarity between Fig. 2f and 3a). We assume that given the spiking activity $\mathbf{r}_{t-\Delta t}$, the recurrent inputs, $\mathbf{u}^{\mathsf{r}}_t$, are independent, and have Poisson-like variability (the variance equals the mean, see Eq. 23). We further assume the mean recurrent input, $\langle \mathbf{u}^{\mathsf{r}}_t \rangle$, averaged over trials when fixing the emitted spikes $\mathbf{r}_{t-\Delta t}$, has a Gaussian tuning profile centered on the stimulus feature sample $\tilde{s}_{t-\Delta t}$ represented by $\mathbf{r}_{t-\Delta t}$, i.e., $\langle \mathbf{u}^{\mathsf{r}}_{tj} \rangle \propto \exp[\mathbf{h}_j(\tilde{s}_{t-\Delta t})] = \exp[-(\tilde{s}_{t-\Delta t} - \theta_j)^2/2a^2]$. With above two assumptions, the conditional distribution of stimulus feature, s, given a particular recurrent input, $\mathbf{u}^{\mathsf{r}}_t$, will be a normal distribution,

$$p(s|\mathbf{u}^{\mathsf{r}}_t) \propto \exp[\mathbf{h}(s)^\top \mathbf{u}^{\mathsf{r}}_t] \propto p(s|\tilde{z}_t) = \mathcal{N}(s|\tilde{z}_t, \Lambda_s^{-1}), \tag{3}$$

whose mean, $\tilde{z}_t$, and precision, Λ_s, can be linearly read out from $\mathbf{u}^{\mathsf{r}}_t$ (Fig. 3b and e, bottom-left; Eq. 25). By construction, we see the recurrent and feedforward inputs represent conditional distributions of stimulus feature, s, with the same basis function (logarithm of tuning), $\mathbf{h}(s)$, so that summing the inputs to a neuron corresponds to multiplying the likelihood $p(\mathbf{u}^{\mathsf{f}}|s)$ and the conditional distribution $p(s|\tilde{z}_t)$ (Eq. 3; Fig. 3c and f),

$$\begin{aligned} &\ln p(s|\tilde{z}_t, \mathbf{u}^{\mathsf{f}}) = \ln p(\mathbf{u}^{\mathsf{f}}|s) + \ln p(s|\tilde{z}_t), \\ \Leftrightarrow \quad &\mathbf{h}(s)^\top \boldsymbol{\lambda}_t = \mathbf{h}(s)^\top \mathbf{u}^{\mathsf{f}} + \mathbf{h}(s)^\top \mathbf{u}^{r}_t. \end{aligned} \tag{4}$$

The firing rate, $\boldsymbol{\lambda}_t$, thus represents the instantaneous conditional distribution, $p(s|\tilde{z}_t, \mathbf{u}^{\mathsf{f}})$, and has the same basis function, $\mathbf{h}(s)$, as that of the feedforward and recurrent inputs (Eq. 2a). Table 2 shows how inputs and instantaneous firing rate represent the distributions over the stimulus features.

Internal Poisson variability drives sampling in stimulus feature subspace

A neural circuit that implements sampling must generate specific level of variability in the feature subspace (see Eq. 2b-2c). A sample from the conditional distribution (Eq. 2b) can be written as $\tilde{s}_t = \hat{s}_t + \Lambda^{-1/2}\xi_t$, where $\hat{s}_t$ and Λ are the mean and precision of the conditional distribution $p(s|\tilde{z}_t, \mathbf{u}^{\mathsf{f}})$ represented by $\boldsymbol{\lambda}_t$ (Eqs. 2a and 4), and ξ_t are independent samples from a standard normal distribution (See Eq. 17). If spiking responses, $\mathbf{r}_t$, represent stimulus samples, $\tilde{s}_t$, then the recurrent network must generate neuronal responses whose projection onto the stimulus feature subspace follows a Gaussian distribution with variance Λ^{-1}.

Poisson-like spiking is observed across cortex [5, 27, 49], and is thus a natural candidate for

the requisite variability. We can show that when the instantaneous firing rate, $\boldsymbol{\lambda}_t$, represents the conditional distribution, $p(s|\tilde{z}_t, \mathbf{u}^{\mathsf{f}})$ in Eq. (4), then independent Poisson spike generation produces samples with the correct statistics in feature subspace (Fig. 3d and g, details in Methods),

$$\mathbf{r}_{tj} \quad \sim \quad \text{Poisson}(\boldsymbol{\lambda}_{tj}) \quad \Rightarrow \quad \tilde{s}_t \sim \mathcal{N}(\tilde{s}_t|\hat{s}_t, \Lambda^{-1}). \tag{5}$$

A sample from the conditional stimulus feature distribution, $\tilde{s}_t$, and the mean of this conditional distribution, $\hat{s}_t$, can be read out respectively from the response, $\mathbf{r}_t$, and firing rate, $\boldsymbol{\lambda}_t$, by using the same linear decoder (population vector), i.e., $\tilde{s}_t = \sum_j \mathbf{r}_{tj}\theta_j / \sum_j \mathbf{r}_{tj}$, $\hat{s}_t = \sum_j \boldsymbol{\lambda}_{tj}\theta_j / \sum_j \boldsymbol{\lambda}_{tj}$. The weights in this decoder are determined by the basis functions (logarithm of tuning), $\mathbf{h}(s)$, as in Eq. (4). Importantly, the variability of stimulus features samples, Λ^{-1}, is determined by the sum of the instantaneous firing rates, $\sum_j \boldsymbol{\lambda}_{tj}$. To confirm this observation we simulate our network with a fixed instantaneous firing rate $\boldsymbol{\lambda}_t$ (Fig. 3c), to generate independent Poisson input over trials (Fig. 3d). In each trial, a sample of $\tilde{s}_t$ was read out by using a population vector (Fig. 3d, Eq. 5). The empirical distribution of the samples, $\tilde{s}_t$, collected over trials (Fig. 3g) matched the conditional distribution represented by the instantaneous firing rate, $\boldsymbol{\lambda}_t$ (Fig. 3c and f, Eq. 4).

We have thus shown that internally generated Poisson variability can serve as a source of fluctuations for sampling from the the conditional distribution over the stimulus feature. Further, we confirm that the overall firing rate of the neural population determines the variability of the stimulus feature samples (Fig. 3h). An advantage of this implementation is that sampling variability, Λ^{-1}, is determined by the firing rate, and can thus change on short-time scales.

Recurrent connections represent the relation between stimulus and context features

To generate samples from the correct joint distribution of stimulus and context features, the recurrent input, $\mathbf{u}^{\mathsf{r}}$, to neurons in the circuit must satisfy Eq. (3), which constrains the mean and variability of recurrent input. First, the input $\mathbf{u}^{\mathsf{r}}$, and the population response, $\mathbf{r}$, should both have Gaussian tuning with the same tuning width over the stimulus feature s (See Eqs. 23 and 26). Second, the mean and variance of the recurrent input conditioned on $\mathbf{r}_t$ need to be equal.

We show that the first requirement is satisfied when the weights between neurons in the recurrent circuit are homogeneous with value (See Eq. 28),

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

where the last equality is obtained from that the sum of $\mathbf{r}$ and $\mathbf{u}^{\mathsf{r}}$ represent the precision of the conditional distribution $p(s|z, \mathbf{u}^{\mathsf{f}})$ and the prior $p(s|z)$, respectively (Table 2). The recurrent weight, w_E^*, thus encodes the relation between the stimulus feature, s, and the context feature, z. We see that on one hand, an increased correlation between s and z (reflected by a narrower diagonal band

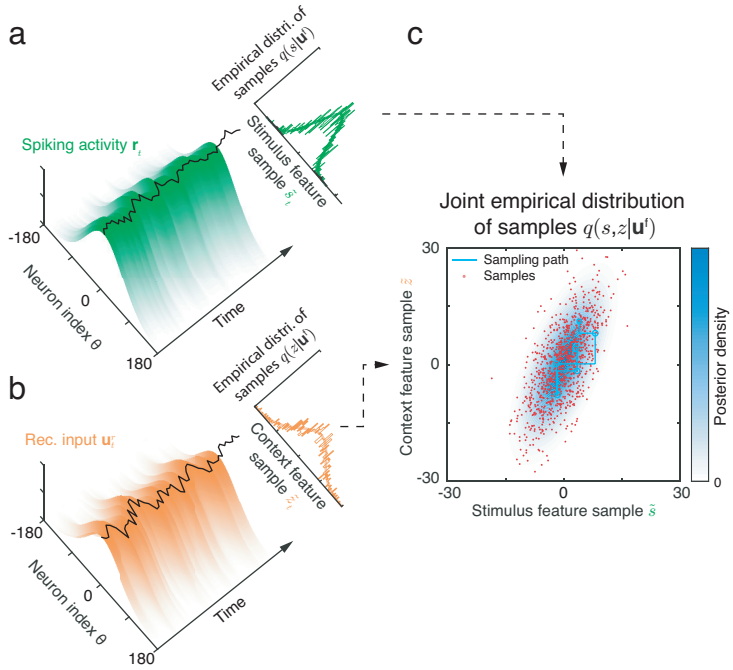

Figure 4: Reading out the posterior over latent features from activity and recurrent inputs in a model circuit. (a-b) Spiking activity, $\mathbf{r}_t$ (a), and recurrent input $\mathbf{u}_t^r$ (b) evolve stochastically. Their projection onto the feature subspace (black curve overlaid on the top of the surface), can be read out linearly by population vector and interpreted as samples of stimulus and context features from their joint posterior distribution (Eqs. 2b-2c). Top right insets: the empirical marginal distributions of samples and marginal posteriors of features (smooth lines). (c) The joint value (red dot) of instantaneous samples of stimulus feature (black curve on the gsurface in a), and context feature (black curve on surface in b) represent samples from the joint posterior of stimulus and context. The true joint posterior distribution is shown in blue.

in Fig. 2c) requires an increase in the recurrent weight w_E^* to satisfy Eq. (3). On the other hand, when context provides no information about the stimulus ($\Lambda_s = 0$), then recurrent interactions are not needed (i.e., w_E^*=0). This is equivalent to the example we chose to motivate our approach, where linear Fisher information was maximized in the absence of recurrent connectivity (Fig. 1e).

To satisfy the second requirement, recurrent interactions in the network should exhibit Poisson-like variability (Fig. 3a, step 3; Eq. 21). This allows the noisy recurrent input, $\mathbf{u}_t^r$, to drive the sampling from the conditional context feature (z) distribution (compare step 3 in Fig. 2d to the recurrent interactions in Fig. 3a). This sampling mechanism is equivalent to the one we proposed for generating samples from the stimulus feature distribution (Eq. 5).

A stochastic E-I network implements sampling in feature subspace

To confirm that a biologically realistic recurrent neural circuit can implement Gibbs sampling, we modeled the proposed mechanism in a network consisting of both excitatory and inhibitory Poisson neurons with synaptic interactions (See details in Eqs. 34-38). The feedforward and recurrent circuitry of our network was identical to the one used to probe the effect of recurrent coupling on a

linear estimation of stimulus s (Fig. 1). We chose the response properties to satisfy representation requirements we described previously: The mean (tuning) of the recurrent inputs $\mathbf{u}^{\mathsf{r}}$ and neuronal response $\mathbf{r}$ both had the same Gaussian profile as the feedforward input $\mathbf{u}^{\mathsf{f}}$ (Eq. 4). The recurrent interactions in the network exhibited independent Poisson-like variability given the emitted spiking responses (Eq. 36) so as to sample the context feature. Thus, we included three sources of variability in our model (Fig. 4a): External variability in feedforward inputs $\mathbf{u}^{\mathsf{f}}$, internal variability due to spike generation, and noisiness in the recurrent interactions. We next explore the contributions of these sources of variability to the inference of stimulus and context features.

Population response of a model network generates feature samples

To show that the model recurrent network produces samples from the joint posterior over stimulus and context features, we simulated its response to a fixed feedforward input $\mathbf{u}^{\mathsf{f}}$ generated following Eq. (1). Therefore all response variability was due to Poissonian spiking and noise in recurrent interactions. Fixing $\mathbf{u}^{\mathsf{f}}$ separates external and internal noise sources in the network, allowing us to directly compare samples generated by the network with the posterior conditioned on a fixed input $\mathbf{u}^{\mathsf{f}}$ (Eq. 14; Fig. 1b). Moreover, since only E neurons are selective for the stimulus and context feature in the E-I network (Fig. 1c), we only read out the samples of features from E neurons. To simplify the notation, we suppress the subscript of neuronal type in the main text, and denote the responses and recurrent inputs of E neurons by $\mathbf{r}_t$ and $\mathbf{u}^{\mathsf{r}}_t$ respectively.

By construction, samples of stimulus and context features are represented respectively by the position of the instantaneous spiking responses $\mathbf{r}_t$, and the recurrent inputs $\mathbf{u}^{\mathsf{r}}_t$ in feature subspace (black curves overlaid on the top of green and orange population responses in Fig. 4a and b). Hence the samples of stimulus and context features can be read out separately from $\mathbf{r}_t$ and $\mathbf{u}^{\mathsf{r}}_t$ using a population vector (Eqs. 5 and 25; Fig. 3b-g). The resulting trajectory of samples in the direct product of the stimulus and context feature subspace exhibits the characteristics of Gibbs sampling (Fig. 4c, cyan line) as the spiking activity and recurrent inputs are updated iteratively in the network. The empirical distribution produced by the collection of samples, $q(s, z|\mathbf{u}^{\mathsf{f}}) = T^{-1} \sum_t \delta \left[(s, z)^\top - (\tilde{s}_t, \tilde{z}_t)^\top \right]$, (Fig. 4c, red dots) provides a good approximation of the joint posterior over stimulus and context features (Fig. 4c, blue contour).

Optimal inference by the network with the correct recurrent weight

To produce correct samples from the joint posterior over stimulus and context features, the recurrent connectivity must support network activity which correctly captures the generative model of the stimulus (Eq. 6). We next verified that setting the recurrent weights to an optimal value allows for sampling from the correct posterior distribution, and that these optimal weights are determined by the prior precision, Λ_s, as suggested by the theory (Eq. 6).

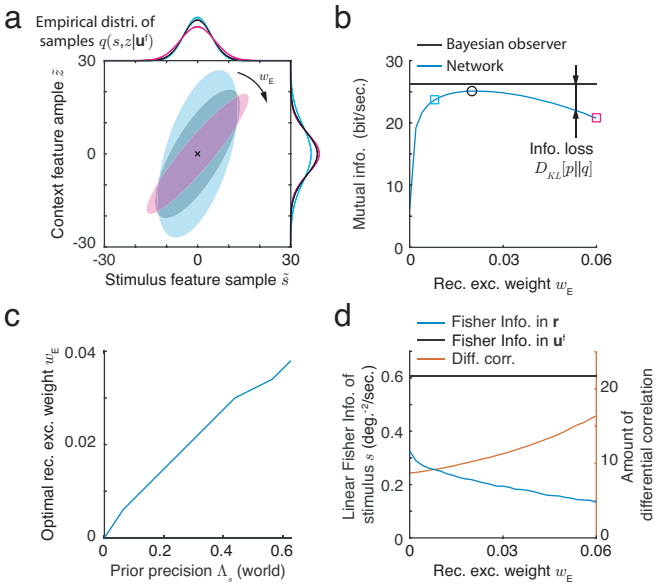

Figure 5: Generating samples from the joint posterior over latent features requires that a network stores an appropriate internal model in recurrent weights. (a) The joint empirical distribution of samples read out from the population responses for different values of peak recurrent excitatory weight, w_E. The ratio of excitatory and inhibitory weights was fixed. Ellipses capture 3 standard deviation from the mean of the joint distribution of samples. The color denotes the value of w_E, shown as symbols with the corresponding color in panel b. (b) The mutual information between latent features, s and z, and feedforward inputs in an ideal Bayesian observer (black) and the network model (blue). The difference between two lines is the KL divergence between the posterior, $p(s, z|\mathbf{u}^\mathbf{f})$ and the distribution of samples generated by the network, $q(s, z|\mathbf{u}^\mathbf{f})$, which quantifies the loss of mutual information in approximating p by q. The loss of information is minimized when the the weights in the recurrent network, w_E, are set so that q best matches p (black circle). The three dots correspond to the parameters used in panel (a). (c) The optimal recurrent weight, w_E^*, that maximizes the lower bound on the mutual information in the network's activity increases with prior precision Λ_s (Fig. 2c). (d) The linear Fisher information of stimulus, s, in the population response, $\mathbf{r}$, decreases with recurrent weight w_E (blue), which is upper bounded by the Fisher information in feedforward inputs $\mathbf{u}^\mathbf{f}$ (black).

Towards this end, we fixed the parameters in the generative model (Eq. 1) and changed the recurrent weights in the network. For simplicity, we only varied the peak excitatory strength, w_E (Eq. 37), and maintained stable activity by fixing the ratio of excitatory and inhibitory synaptic weights. The joint distributions of samples produced by the recurrent network varied considerably as we varied the peak recurrent weights, w_E (Fig. 5a). Changing w_E did not change the mean, but did affect the variance and covariance of the distribution of samples, and hence the inferred distribution over the features (Fig. 5a). Stronger weights, w_E, increased the variance in the marginal distribution over stimulus features, while the variance over context feature samples decreased. Moreover, the orientation of the joint distribution rotated clockwise, increasing the correlations between the samples $\tilde{s}$ and $\tilde{z}$.

We quantified the distance between the empirical distribution of samples, $q(s, z|\mathbf{u}^\mathbf{f})$, generated by the network and the true posterior, $p(s, z|\mathbf{u}^\mathbf{f})$ (Eq. 14), with Kullback-Leibler (KL) divergence

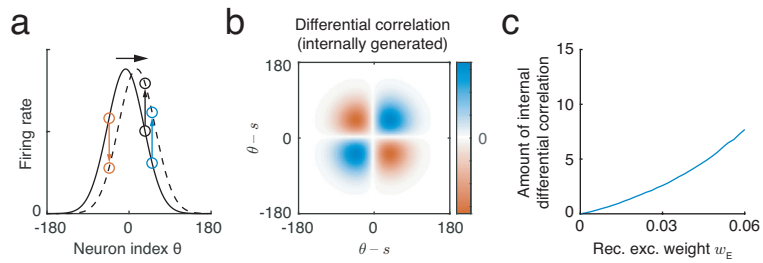

Figure 6: Sampling over stimulus feature subspace is reflected in internally generated differential correlations. (a) When the network generates samples in the stimulus feature subspace, the population firing rate fluctuates along the feature subspace. Neuron pairs on the same side (the black and blue circles) with respect to the mean of stimulus samples ($s = 0$ in this example), have positively correlated activity, while neuron pairs on opposite sides (the black the red circles) with respect to the mean of stimulus samples are negatively correlated. Fluctuations in the stimulus feature subspace generate correlations proportional to the product of the derivatives of the population firing rates with the stimulus. (b) The pattern of internally generated differential correlation which is resulted from the sampling in stimulus feature subspace. (c) The internal differential correlation increases with recurrent excitatory weight w_E which increases with the correlation between stimulus and context in the prior, Λ_s.

between q and p. KL divergence quantifies the loss of mutual information, measured in bits, between latent features, i.e., s and z, and the feedforward inputs $\mathbf{u}^{\mathbf{f}}$ due to approximating the true posterior distribution, p, with the distribution q (Eqs. 18-19, [45]). KL divergence is always non-negative: A zero KL divergence indicates q equals p, while a positive KL divergence reflects a deviation of q from p which causes information loss. The recurrent connections in the network can be tuned to minimize the information lost in approximating the posterior, p, by the distribution of samples, q.

Our simulations showed that the loss of mutual information 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). In contrast, the linear Fisher information of stimulus feature, s, obtained from the neuronal responses, $\mathbf{r}$, decreases monotonically with w_E (Fig. 1d and 5d, left vertical axis). This is due to the increase in correlated variability with stronger recurrent interactions (Fig. 5d, right vertical axis), and a resulting decreases in discriminability of the external stimulus feature s from the neuronal response, $\mathbf{r}$.

To confirm that the optimal recurrent weights, w_E, increase with prior precision (Eq. 6), we simulated the network response at different values of Λ_s in the generative model, and found the weights that minimized the loss of mutual information. The results confirm our theoretical observations (Fig. 5c): When $\Lambda_s = 0$, so that the context feature and stimulus features are uncorrelated, a network with no interactions performs best ($w_E^* = 0$). For small Λ_s (relative to $\Lambda_{\mathbf{f}}$) the optimal weight w_E^* increases roughly linearly with Λ_s, as predicted by our theory.

Neural signature of sampling in feature subspace: internally generated differential correlations

We next asked what measurable signatures in neural data would characterize sampling of the type we describe. Such sampling would cause instantaneous firing rates, $\boldsymbol{\lambda}_t$, to fluctuates along the stimulus feature subspace over time (Fig. 6a), and this would be reflected in covariations of the responses of pairs of neurons within the population. To understand the structure of such covariability, we considered population response fluctuations along the stimulus feature subspace, and ignored the fluctuations along other directions. Suppose the mean population response projected onto the feature subspace is centered at $\hat{s} = 0°$ (Fig. 6a). When the population response fluctuates around its mean, the responses of neurons with stimulus preference on the same side of the projected mean position will tend to be positively correlated (the black and blue neurons in Fig. 6a). In contrast, neurons on opposite sides of the mean will tend to be negatively correlated (the black and red neurons in Fig. 6a). This description generalizes to how sampling dynamics structures pairwise correlations in stimulus feature subspace across the entire network (Fig. 6b). Our theory suggests that sampling in stimulus feature subspace results in covariability proportional to the product of the derivative of mean population firing rate, i.e., $\mathbf{f}'_s\mathbf{f}'^{\top}_s$ where $\mathbf{f}'_s$ denotes the derivative of mean population rate $\mathbf{f}$ over stimulus feature, s. Such correlations have been referred to as differential correlations in previous studies [2, 25], and they limit the information about s that is available to a decoder as measured by linear Fisher information (see Fig. 1e).

In our model, a component of differential correlations is internally generated through recurrent network interactions. Consequently, their magnitude is determined by the strength of recurrent connections. We computed the covariance matrix of the population response, $\Sigma(\mathbf{r}|\mathbf{u}^{\mathsf{f}})$, (See Eq. 31, e.g., [50]), and found that the amplitude of differential correlations is proportional to the recurrent weight w_E^* (Fig. 6c),

$$\Sigma(\mathbf{r}|\mathbf{u}^{\mathsf{f}}) = \mathrm{diag}[\mathbf{f}(s)] + a^2 n_{\mathsf{f}}^{-1} w_E^* \mathbf{f}'_s \mathbf{f}'^{\top}_s + (\text{terms independent of } s), \quad \text{where} \quad s = \mu_{\mathsf{f}}. \tag{7}$$

In Eq. (7), the second term on the right hand side represents internally generated differential correlation with amplitude $a^2 n_{\mathsf{f}}^{-1} w_E^*$, where a is the tuning width and n_{f} is the total number of feedforward input spikes received by all E neurons (Table 1), and $\mathrm{diag}[\mathbf{f}(s)]$ is a diagonal matrix capturing the individual Poisson spiking variability, with $\mathbf{f}(s)$ the tuning of E neurons over s. Moreover, since the optimal recurrent weight, w_E^*, is determined by the correlation between stimulus and context (Fig. 5c, Λ_s in Eq. 6), the amplitude of differential correlations in a network implementing optimal sampling is determined by the same quantity (Fig. 5d and Eq. 32). We confirmed this conclusion in network simulations which showed that differential correlations increased with excitatory recurrent weight, w_E (Fig. 6c), and hence with an increase in prior precision (Fig. 5c).

To understand why internally generated differential correlations are shaped by the structure

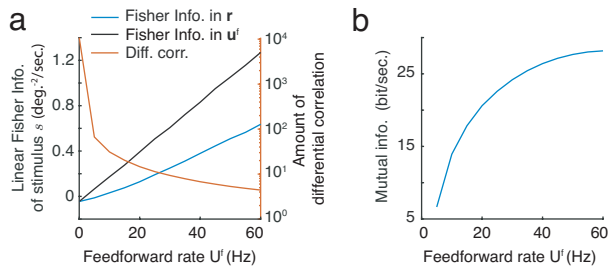

Figure 7: The external differential correlations, inherited from feedforward inputs, impair information representation. (a) The feedforward rate determines the input gain, and the reliability of sensory transmission and the amount of external differential correlation inherited from feedforward input. (b) The mutual information in the distribution of samples generated by the network with optimized recurrent weight w_E^* monotonically increases with feedforward rate.

of the generative model, we consider the case of uncorrelated context and stimulus ($\Lambda_s = 0$), as we did when examining linear Fisher information (Fig. 1a). In this case, an optimal network should have no recurrent connectivity, and instantaneous firing rates, $\boldsymbol{\lambda}_t$, are fixed in response to a fixed feedforward input. Therefore, the variabilitiy in the population response, $\mathbf{r}$, is only due to independent Poisson variability in spike generation (the 1st term on the right hand side of Eq. 7). Increasing the correlations between stimulus and context (Fig. 6c) results in larger recurrent weights, w_E^* (Fig. 6c), which increase recurrent variability (Fig. 1d, red), and hence increase internally generated differential correlations.

Distinguishing external and internal differential correlations

Unlike in our simulations, in a typical neurophysiology experiment a stimulus feature, s, is fixed, while the feedforward input, $\mathbf{u}^{\mathsf{f}}$, fluctuates (Fig. 4a). Hence in experimentally recorded neuronal response, fluctuations in the stimulus feature subspace reflect a combination of sensory transmission noise, and any internally generated variability. Differential correlations in the evoked population activity, $\mathbf{r}$, in response to a fixed stimulus, s, can be decomposed into external correlations inherited from fluctuations in $\mathbf{u}^{\mathsf{f}}$, and the correlations generated from recurrent interactions. In this case the total magnitude of differential correlation in the evoked response is $a^2 n_{\mathsf{f}}^{-1}(w_E^* + 1)\mathbf{f}_s'\mathbf{f}_s'^{\top}$ (Eq. 33), where the new term, $a^2 n_{\mathsf{f}}^{-1}\mathbf{f}_s'\mathbf{f}_s'^{\top}$, reflects differential correlations inherited from the input (comparing with Eq. 7).

Although the two sources of differential correlations are intertwined, we found that they impact information processing differently. This difference can potentially provide a way to distinguish them, and verify whether the network indeed produces samples in the joint stimulus and context feature subspace. Internally generated differential correlations have magnitude proportional to the recurrent weight, w_E^* (Eq. 7). Furthermore, the non-monotonic relation between w_E and the mutual information (Fig. 5b) implies a similar non-monotonic relation between internal differential

correlations and the mutual information (Fig. 8a). In contrast, an increase in feedforward input gain (the peak firing rate of feedforward input, Eq. 10) leads to a decrease in differential correlations (Fig. 7a), while the mutual information keeps increasing (Fig. 7b). We therefore have a monotonic, decreasing relationship between externally generated differential correlations and the mutual information (Fig. 8b). Thus such inherited correlations always impair information processing, as observed previously [2, 25]. Therefore, the two sources of differential correlations could be distinguished by the non-monotonic and the monotonic relation with the mutual information in the network (see Discussion and Fig. 8).

Discussion

In this study we propose a new role for recurrent connections in stimulus representation, and demonstrate their function in a circuit model of a cortical hypercolumn that implements sampling-based Bayesian inference of the joint posterior over a local stimulus and a global context feature. There is strong evidence that Bayesian inference underlies perceptual processing [14, 15, 17, 36], such as multisensory cue integration [13], and sensorimotor learning [37]. In our circuit model, the recurrent connections store an associative prior between the stimulus and context features, and the internal Poisson variability provides proper variability for sampling stimulus features. Finally, the stimulus and context features are embedded in a subspace in the neural population responses from which the posterior of features can be read out linearly.

A neural signature of network sampling in stimulus feature subspace is the covariability of population responses along the stimulus feature subspace, i.e., the appearance of internally generated differential correlations. Such correlations are due to recurrent interactions in the network, and not inherited from sensory transmission noise in feedforward inputs (e.g., [25, 51]). We have shown that the amplitude of internally generated differential correlations in networks that are tuned to accurately sample from the posterior distribution are determined by the associative prior between stimulus and context features. This observation could be used to verify whether cortical circuits sample stimulus features, as we discuss below. While external and internal differential correlations are necessarily intertwined in any recorded neuronal responses, we found that they impact information representation in the circuit differently. The mutual information in the network is maximized at a unique amplitude of internal differential correlations (Fig. 5 and 6), while external differential correlations always decreased the mutual information and hence impaired information processing (Fig. 7b).

The proposed circuit sampling mechanism produces experimentally testable predictions. First, it can be tested whether the amount of internal differential correlations is determined by the prior correlation between stimulus and context (Fig. 6b-c). This could be done using a perceptual learning task in which a subject is repeatedly presented with the same set of stimuli (e.g., [52]).

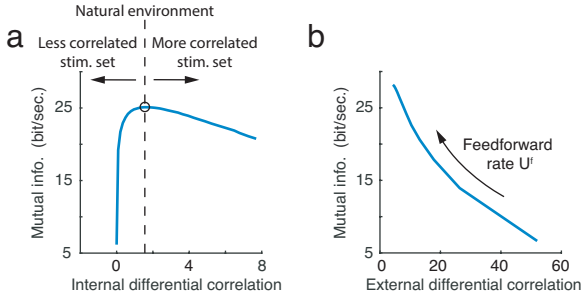

Figure 8: External and internal differential correlations have a distinct impact on information representation in recurrent circuit model. While we find a non-monotonic relation between the mutual information in recurrent network and internal differential correlation (a), the mutual information decreases with external differential correlations (b).

The experimenter could change the correlation between global context and the local stimulus in the stimulus set, and test whether the differential correlations increase during and after learning as suggested by Fig. 5c-d. Indirect evidence from previous studies supports this possibility: It has been shown that neurons become less correlated with perceptual learning with a simple stimulus set [53], and spontaneous population response patterns, including correlation structure, become more similar to evoked response patterns during development [54]. Second, to verify the impact of internal differential correlations on information representation, one could test whether there exists a non-monotonic relation between the amount of internal differential correlation and the mutual information estimated from population responses (Fig. 8a). For example, suppose an animal is fully adapted to the natural environment. Training the animal with a correlated stimulus set would then alter the amplitude of internal differential correlations. Regardless of whether the differential correlations increase or decrease, our network model predicts the mutual information in the population responses will decrease (Fig. 8a).

The recurrent circuit model we presented includes novel mechanisms for inference and representation, which combine features from previously proposed neural coding frameworks, including parametric representations characteristic of probabilistic population codes (PPCs) [18, 38], and features of sampling-based codes (SBCs) e.g., [19, 23, 46, 48, 55]: In our network, the conditional distribution of the stimulus feature is represented by instantaneous firing rates using a parametric code similar to PPC (Eqs. 3 and 4), i.e., the linear exponential family with sufficient statistics. The network produces samples in the stimulus feature subspace with sampling variability generated internally (Fig. 3a). On the other hand, there are also some fundamental distinctions between our network and PPCs, as well as conventional SBCs: PPCs are generally implementd in deterministic networks that sum Poissonian inputs, and variability of neuronal responses is inherited from feedforward inputs. In contrast, the network model we proposed is doubly stochastic and includes both internal Poisson variability to drive sampling in the joint stimulus and context feature subspace, as well as variability in feedforward inputs to the network (Fig. 3a). Second, in PPCs the posterior

of stimulus feature is represented parametrically in neuronal responses, while in the proposed network the joint posterior over stimulus and context features is approximated by sequential samples. Most conventional SBCs rely on the assumption that the profile of averaged population responses is directly proportional to the posterior over a stimulus feature (e.g., [19, 23, 24, 46, 48, 55]), i.e., $\langle \mathbf{r} \rangle \propto p(s|\mathbf{u}^{\mathsf{f}})$, while in our network the posterior is approximated by samples in the stimulus feature subspace in the space of neuronal population responses. The samples of stimulus and context features can be read out linearly from the instantaneous spiking activity, $\mathbf{r}_t$, and recurrent input $\mathbf{u}_t^{\mathsf{r}}$ respectively (Fig. 4). Consequently, in the proposed network the sampling space has dimension of two, while in conventional SBCs the dimension of sampling space is the number of neurons in the network. A benefit of sampling in feature subspace is that the small dimension of the sampling space making sampling more efficient.

On the network level, the recurrent connections are translation-invariant in the stimulus feature subspace, an assumption widely used in continuous attractor networks (CAN, e.g, [32, 56–58]). Translation-invariant connections are a mathematical simplification, but are not required for a circuit to implement sampling. CANs have been shown to achieve maximal likelihood estimation via template matching [18, 57, 59]. Here we have shown that a network with CAN-like structure and internally generated variability is able to perform sampling in stimulus feature subspace to approximate the posterior over stimulus features, rather than providing a point estimate as in most deterministic CANs. As Poissonion spiking is ubiquitous in cortex, real cortical circuits may use it to sample stimulus features. Biophysically, Poisson variability can emerge due to a balance between excitation and inhibition (e.g., [60–62]).

The proposed network mechanism implementing sampling-based inference can be extended to more general cases, and more complex generative models. For example, a similar network can generate samples from other types of distributions belonging to the linear exponential family [38, 45]: Eq. 4 suggests that the basis function (logarithm of tuning) $\mathbf{h}(s)$ determines the type of distribution from which the network samples. These basis functions can be changed by changing the tuning of feedforward inputs, along with the recurrent connection profiles. The network could also produce samples from a high-dimensional stimulus feature distribution, $(s_1, s_2, \ldots, s_m)$. In this case, features could be inferred jointly by coupled networks consisting of neurons with internal Poisson spiking variability, with each subnetwork inferring the marginal posterior of a feature. The recurrent connections in the coupled networks then need to be set to make the mean recurrent inputs received by each network have the same profile as the feedforward inputs received by that network. Our recent preliminary study demonstrates this possibility [63].

In conclusion, we showed how recurrent connections in cortical circuit can store information about the structure of the world, and how internal Poisson variability can drive sampling from the posterior over directly and indirectly observed features. Internally generated differential correlation are a signature of the sampling of stimulus features, and the size of these correlations is determined

by the prior stored in recurrent circuits.

Methods

The probabilistic generative model

We assume that the statistics of stimuli in the world and the feedforward inputs received by neurons can be described by a hierarchical generative model. For illustration, we use a hypercolumn of V1 neurons as a concrete example to explain the underlying idea (Fig. 2a), although our generative model is general. We denote by z the one dimensional context feature (e.g., orientation) of a big object in the visual field, which represents the contextual information and may cover the receptive fields of multiple V1 hypercolumns. Although the orientation distribution of natural images is not uniform and even bimodal [64], we consider z satisfies a uniform distribution for simplicity (Fig. 2c, marginal plots).

$$p(z) = \mathcal{U}(-180^\circ, 180^\circ). \tag{8}$$

This uniform assumption simplifies our model significantly, and can be linked with the homogeneous neural population in a network model as considered in Eqs. (10-11), but it doesn't substantially change our main results. Moreover, we use s to represent the one dimensional stimulus feature (e.g., orientation) of the local stimulus falling into the central receptive field (cRF) of a single V1 hypercolumn. Due to the variations between the local (stimulus) and global (context) aspects of the senory scene, the stimulus feature s distributes around contextual feature z, and is assumed to obey a normal distribution,

$$p(s|z, \Lambda_s) = \mathcal{N}\left(s|z, \Lambda_s^{-1}\right). \tag{9}$$

Here $\mathcal{N}\left(s|z, \Lambda_s^{-1}\right)$ denotes a normal distribution over s with mean z and precision (inverse of variance) Λ_s.

Given the stimulus feature s located in the center receptive field of the V1 hypercolumn, we let the feedforward inputs $\mathbf{u}^{\mathsf{f}}$ received by N_E excitatory neurons in V1 be independent Poisson spikes, with input tuning (mean input) over s modeled as a Gaussian function,

$$\begin{aligned} &p(\mathbf{u}^{\mathsf{f}}|s) = \textstyle\prod_{j=1}^{N_E} \text{Poisson}\left[\mathbf{u}_j^{\mathsf{f}} \middle| \langle \mathbf{u}_j^{\mathsf{f}}(s) \rangle\right], \\ &\langle \mathbf{u}_j^{\mathsf{f}}(s) \rangle = U^{\mathsf{f}} \exp[\mathbf{h}_j(s)] = U^{\mathsf{f}} \exp[-(\theta_j - s)^2/2a^2], \end{aligned} \tag{10}$$

where $\mathbf{u}_j^{\mathsf{f}}$ denotes the feedforward input received by j-th excitatory neuron. $\langle \mathbf{u}_j^{\mathsf{f}}(s) \rangle$ is the tuning of the feedforward input, with θ_j denoting the preferred stimulus feature of $\mathbf{u}_j^{\mathsf{f}}$, and U^{f} and a characterize respectively the peak rate and width of the tuning. This setup of input has been widely used in previous neural coding studies, e.g., the probabilistic population code (PPC) [18, 38, 39].

With the assumption that the preferred stimulus feature $\{\theta_j\}_{j=1}^{N_E}$ of all feedforward inputs are

uniformly distributed in the stimulus feature s' space, Eq. (10) suggests that the likelihood of s based on a single observation of $\mathbf{u}^{\mathsf{f}}$ can be written as the form of the exponential family of distributions with linear sufficient statistics [18, 38, 39],

$$\begin{aligned} p(\mathbf{u}^{\mathsf{f}}|s) &= \prod_{j=1}^{N_E} \frac{\langle \mathbf{u}_j^{\mathsf{f}}(s)\rangle^{\mathbf{u}_j^{\mathsf{f}}}}{\mathbf{u}_j^{\mathsf{f}}!} \exp[-\langle \mathbf{u}_j^{\mathsf{f}}(s)\rangle], \\ &= \phi(\mathbf{u}^{\mathsf{f}}) \exp\left[\mathbf{h}(s)^\top \mathbf{u}^{\mathsf{f}}\right], \\ &\propto \exp\Big[-\textstyle\sum_j \mathbf{u}_j^{\mathsf{f}} \frac{(\theta_j - s)^2}{2a^2}\Big], \\ &\propto \mathcal{N}\big(s|\mu_{\mathsf{f}}, \Lambda_{\mathsf{f}}^{-1}\big). \end{aligned} \tag{11}$$

The 2nd line in above equation denotes the standard form of distributions belonging to a linear exponential family, where $\phi(\mathbf{u}^{\mathsf{f}})$ is a function which only depends on Poisson spike input $\mathbf{u}^{\mathsf{f}}$. Here $\mathbf{h}(s)$ is called the basis function [18], which determines the profile of input tuning (Eq. 10), as well as the type of likelihood distribution (Eq. 11). Since the input tuning has a Gaussian profile, the likelihood of s given the $\mathbf{u}^{\mathsf{f}}$ can be normalized into a Gaussian distribution (Eq. 11), with mean μ_{f} and precision Λ_{f} that can be linearly read out from $\mathbf{u}^{\mathsf{f}}$,

$$\mu_{\mathsf{f}} = n_{\mathsf{f}}^{-1} \textstyle\sum_j \mathbf{u}_j^{\mathsf{f}} \theta_j, \quad \Lambda_{\mathsf{f}} = a^{-2} n_{\mathsf{f}} = a^{-2} \textstyle\sum_j \mathbf{u}_j^{\mathsf{f}}. \tag{12}$$

Here the integer n_{f} is the feedforward spike count. The mean μ_{f} represents the position of $\mathbf{u}^{\mathsf{f}}$ in stimulus feature subspace, and the precision Λ_x is proportional to the sum of input spike count, n_{f} (Fig. 2d-e). All together, the whole generative model is then

$$\begin{aligned} p(\mathbf{u}^{\mathsf{f}}, s, z) &= p(\mathbf{u}^{\mathsf{f}}|s) p(s|z) p(z), \\ &\propto \textstyle\prod_{j=1}^{N_E} \text{Poisson}\big[\mathbf{u}_j^{\mathsf{f}} \big| \langle \mathbf{u}_j^{\mathsf{f}}(s)\rangle\big] \mathcal{N}\left(s|z, \Lambda_s^{-1}\right), \\ &\propto \mathcal{N}\big(s|\mu_{\mathsf{f}}, \Lambda_{\mathsf{f}}^{-1}\big) \mathcal{N}\big(s|z, \Lambda_s^{-1}\big). \end{aligned} \tag{13}$$

Approximate Bayesian inference by Gibbs sampling

Given a feedforward input $\mathbf{u}^{\mathsf{f}}$, a neural circuit interprets the outside world by inferring the latent features s and z. Our neural circuit is assumed to infer the whole posterior distribution over latent stimuli, instead of making a single point estimate [14, 36]. The simple generative model (Eq. 13) yields an analytical solution of the posterior distribution,

$$\begin{aligned} p(s, z|\mathbf{u}^{\mathsf{f}}) &= \mathcal{N}\big[(s,z)^\top | \boldsymbol{\mu}_p, \mathbf{K}_p^{-1}\big], \\ \boldsymbol{\mu}_p = (\mu_{\mathsf{f}}, \mu_{\mathsf{f}})^\top, &\quad \mathbf{K}_p = \begin{pmatrix} \Lambda_{\mathsf{f}} + \Lambda_s & -\Lambda_s \\ -\Lambda_s & \Lambda_s \end{pmatrix}. \end{aligned} \tag{14}$$

Although the posterior has an analytical solution in the proposed generative model (Eq. 14), we are more interested in a general inference algorithm which can be implemented in the cortical circuit. As such, Eq. (14) will be mainly used as a criteria to evaluate the algorithm. Since s and z are organized hierarchically in the generative model, their inference can be conducted iteratively, which has the potential to be implemented by a recurrent network. Moreover, to make use of the internal variability originated inside a recurrent network, we consider a Monte Carlo method to infer the posterior [15, 19, 23, 24, 46, 55, 65]. Combined together, this motivates us to consider a Gibbs sampling algorithm to compute the posterior distribution [45, 66], which is a stochastic process iteratively estimating the latent hierarchical variables.

In Gibbs sampling we suppose a sample $\tilde{z}_t$ which serves as an estimate for context z at time t. The sample $\tilde{z}_t$ is combined with feedforward input $\mathbf{u}^{\mathsf{f}}$ to update the conditional distribution of stimulus feature s (step 1 in Fig. 2f),

$$\begin{aligned} p(s|\tilde{z}_t, \mathbf{u}^{\mathsf{f}}) \propto p(\mathbf{u}^{\mathsf{f}}|s)p(s|\tilde{z}_t) \propto \mathcal{N}\left(s|\hat{s}_t, \Lambda^{-1}\right), \\ \hat{s}_t = \frac{\Lambda_{\mathsf{f}}\mu_{\mathsf{f}} + \Lambda_s \tilde{z}_t}{\Lambda_{\mathsf{f}} + \Lambda_s}, \quad \Lambda = \Lambda_{\mathsf{f}} + \Lambda_s. \end{aligned} \tag{15}$$

Here $\hat{s}_t$ and Λ denote the mean and precision (inverse of variance) of the conditional distribution $p(s|\tilde{z}_t, \mathbf{u}^{\mathsf{f}})$. Subsequently, a sample of $\tilde{s}_t$ will be drawn from the conditional distribution $p(s|\tilde{z}_t, \mathbf{u}^{\mathsf{f}})$ (step 2 in Fig. 2f),

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

Then the conditional distribution of z will be updated according to the sample $\tilde{s}_t$ and then a new sample $\tilde{z}_{t+\Delta t}$ will be drawn from it (step 3 in Fig. 2f),

$$\tilde{z}_{t+\Delta t} \sim p(z|\tilde{s}_t) = \mathcal{N}(z|\tilde{s}_t, \Lambda_s^{-1}). \tag{16}$$

The above three steps in Gibbs sampling (Eqs. 15-16) will be performed iteratively until reaching an equilibrium state. In equilibrium the distribution of samples of $\tilde{s}_t$ and $\tilde{z}_t$ approximate the true posterior distribution accurately (Fig. 2g; compare the red dots with the blue contour map). Gibbs sampling (Eqs. 15-16) can also be written as a discrete equation,

$$\begin{aligned} \hat{s}_t &= \frac{\Lambda_{\mathsf{f}}\mu_{\mathsf{f}} + \Lambda_s \tilde{z}_t}{\Lambda_{\mathsf{f}} + \Lambda_s}, \\ \tilde{s}_t &= \hat{s}_t + (\Lambda_{\mathsf{f}} + \Lambda_s)^{-1/2}\xi_t, \\ \tilde{z}_{t+\Delta t} &= \tilde{s}_t + \Lambda_s^{-1/2}\epsilon_t, \end{aligned} \tag{17}$$

where ϵ_t and ξ_t are two independent Gaussian white noises.

The inference in the information-theoretic point of view

The inference of latent hierarchical features $\Theta = \{s, z\}$ based on the feedforward input $\mathbf{u}^{\mathsf{f}}$ can be explained from an information theoretic perspective [66, 67]. Inferring the latent stimuli Θ corresponds to maximizing the mutual information between $\mathbf{u}^{\mathsf{f}}$ and Θ, defined as,

$$\begin{aligned} I(\Theta; \mathbf{u}^{\mathsf{f}}) &= H(\Theta) - H(\Theta|\mathbf{u}^{\mathsf{f}}), \\ &= -\mathbb{E}_{p(\Theta)}[\log p(\Theta)] + \mathbb{E}_{p(\Theta,\mathbf{u}^{\mathsf{f}})}[\log p(\Theta|\mathbf{u}^{\mathsf{f}})], \end{aligned} \tag{18}$$

where log is the logarithm function with base 2, and mutual information has the units of bits. In above equation, $\mathbb{E}_{p(X)}[f(\cdot)] = \int p(X) f(\cdot) dX$ calculates the expectation of any function $f(\cdot)$ under the distribution $p(X)$. $H(\Theta)$ and $H(\Theta|\mathbf{u}^{\mathsf{f}})$ measure the entropy of the prior $p(s, z)$ (Eqs. 8 and 9) and the posterior distribution $p(s, z|\mathbf{u}^{\mathsf{f}})$ (Eq. 14), respectively. The mutual information between $\mathbf{u}^{\mathsf{f}}$ and Θ is the reduction in entropy of Θ due to knowledge of the feedforward input $\mathbf{u}^{\mathsf{f}}$, and is determined by the parameters of generative models, e.g., the precision Λ_{f} and Λ_s in the generative model (Eq. 13).

When an estimator (sampler or a recurrent network) uses an internal model (prior) which does not exactly match the structure between the real world Θ, the distribution approximated by the estimator, $q(\Theta|\mathbf{u}^{\mathsf{f}})$, will differ from the true posterior $p(\Theta|\mathbf{u}^{\mathsf{f}})$ (Eq. 14). In this case the information of Θ extracted by the estimator will be decreased,

$$I(\Theta, \mathbf{u}^{\mathsf{f}}) \geq -\mathbb{E}_{p(\Theta)}[\log p(\Theta)] + \mathbb{E}_{p(\Theta,\mathbf{u}^{\mathsf{f}})}[\log q(\Theta|\mathbf{u}^{\mathsf{f}})] \equiv I_q(\Theta; \mathbf{u}^{\mathsf{f}}), \tag{19}$$

where the true mutual information $I(\Theta, \mathbf{u}^{\mathsf{f}})$ is an upper bound for the information extracted by the estimator, $I_q(\Theta; \mathbf{u}^{\mathsf{f}})$. Indeed, 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., $D_{KL}[p||q] = I(\Theta, \mathbf{u}^{\mathsf{f}}) - I_q(\Theta, \mathbf{u}^{\mathsf{f}}) = \mathbb{E}_p(\ln p - \ln q)$ (which is always non-negative). The equality in Eq. (19) can be satisfied only if the approximated distribution q matches the posterior p exactly.

The mutual information of the estimator $I_q(\Theta; \mathbf{u}^{\mathsf{f}})$ can be computed analytically when the approximated distribution $q(\Theta|x) = \mathcal{N}\left(\Theta|\boldsymbol{\mu}_q, \mathbf{K}_q^{-1}\right)$ is a bivariate normal distribution (substituting Eqs. 13 and 14 into Eq. 19),

$$I_q(\Theta; \mathbf{u}^{\mathsf{f}}) = \log L + \frac{1}{2}[1 - \log(2\pi)] + \frac{1}{2}\log\frac{|\mathbf{K}_q|}{\Lambda_s} - \frac{1}{2}\left[\mathrm{tr}(\mathbf{K}_p^{-1}\mathbf{K}_q) + (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^\top \mathbf{K}_q(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)\right], \tag{20}$$

where $L = 360^\circ$ is the length of the stimulus feature subspace. Here $\boldsymbol{\mu}_p$ and $\mathbf{K}_p$ are the mean and the precision matrix of the posteiror distribution (Eq. 14). When the distribution q matches

exactly with the posterior p the estimator reaches the mutual information,

$$I(\Theta; \mathbf{u}^{\mathsf{f}}) = \log L - \frac{1}{2}[1 + \log(2\pi)] + \frac{1}{2}\log \Lambda_{\mathsf{f}}.$$

As expected we see that the mutual information increases with the sensory transmission reliability Λ_{f} (Eq. 7b).

In comparison, the Fisher information, widely used in neural coding studies (e.g., [27, 68]), is defined as the curvature of the log-likelihood function,

$$\mathcal{I}(\Theta|\mathbf{u}^{\mathsf{f}}) = \mathbb{E}_{p(\mathbf{u}^{\mathsf{f}}|\Theta)}\left[-\frac{\partial^2}{\partial\Theta^2}\ln p(\mathbf{u}^{\mathsf{f}}|\Theta)\right]$$

The inverse of Fisher information, called the Cramér-Rao bound, measures the lower bound of the mean squared error (MSE) of any unbiased estimator [67],

$$\mathbb{E}_{\Theta}[(\hat{\Theta}(\mathbf{u}^{\mathsf{f}}) - \Theta)^2] \geq \mathcal{I}(\Theta|\mathbf{u}^{\mathsf{f}})^{-1},$$

where $\hat{\Theta}(\mathbf{u}^{\mathsf{f}})$ is an unbiased estimator of Θ based on feedforward input $\mathbf{u}^{\mathsf{f}}$. The unbiased estimator which reaches the Cramér-Rao bound is the maximum-likelihood estimator (MLE). With the MSE as loss function, the MLE outputs the mean of the likelihood function

$$\hat{\Theta}_{Fisher}(\mathbf{u}^{\mathsf{f}}) = \mathbb{E}_{p(\mathbf{u}^{\mathsf{f}}|\Theta)}(\Theta) = \int \Theta p(\mathbf{u}^{\mathsf{f}}|\Theta) d\mathbf{u}^{\mathsf{f}}.$$

It is worthy to note that the MLE estimator is a point estimator, i.e., which only outputs a single value of estimate rather than a whole posterior distribution (Eq. 14). In a probabilistic framework (i.e., Eq. 19), the approximated "distribution" of the MLE based on a $\mathbf{u}^{\mathsf{f}}$ is,

$$q_{Fisher}(\Theta|\mathbf{u}^{\mathsf{f}}) = \mathcal{N}\big[\Theta|\hat{\Theta}(\mathbf{u}^{\mathsf{f}}), \delta\big(\Theta - \hat{\Theta}(\mathbf{u}^{\mathsf{f}})\big)\big],$$

where $\delta(\cdot)$ is the Dirac delta function. It is clear to see the MLE discards the uncertainty information, in that the precision in q_{Fisher} is fixed and is blind to the precision in the generative model (Eq. 13). In other words, the structure of the world Θ does not impact the quality of the estimate of Θ.

Towards a stochastic recurrent network sampling latent features

We consider a generic recurrent network consisting of N_E excitatory Poisson neurons and attempt to enact the sampling of latent features through the network dynamics. For simplicity, neuronal activity is described by a Hawkes process (a type of inhomogeneous Poisson process, e.g., [69]). The instantaneous firing rate of N_E excitatory Poisson neurons at time t, $\boldsymbol{\lambda}_t$, obey the following

Table 1: Notations in the model

Symbol	Description
z	Context feature (scalar variable)
s	Stimulus feature (scalar variable)
Λ_s	The prior precision
$\mathbf{r}_t$	Neuronal spikes at time t (vector)
$\boldsymbol{\lambda}_t$	Firing rate at time t (vector)
$\mathbf{u}^{\mathsf{f}}$	Feedforward spiking inputs (vector)
$\mathbf{u}^{\mathsf{r}}_t$	Recurrent input at time t (vector)
U^{f}	Peak feedforward input rate
U^{r}	Peak value of recurrent input
n_{f}	The total spike count of $\mathbf{u}^{\mathsf{f}}$ to all neurons (scalar variable)
n_{r}	The sum of recurrent input (scalar variable)
w_E	The peak recurrent weight between excitatory neurons
θ_j	The preferred stimulus feature of j-th excitatory neuron
a	Tuning width

recurrent equations:

$$\boldsymbol{\lambda}_t = \mathbf{u}^{\mathsf{f}} + \mathbf{u}^{\mathsf{r}}_t = \mathbf{u}^{\mathsf{f}} + \left(w_E \mathbf{r}_{t-\Delta t} + \sigma_r \boldsymbol{\xi}_t\right), \tag{21}$$

$$\mathbf{r}_t \sim \prod_{j=1}^{N_E} \text{Poisson}(\boldsymbol{\lambda}_{tj}), \tag{22}$$

where $\mathbf{u}^{\mathsf{f}}$ is the feedforward Poisson spike input (Eq. 10), $\mathbf{u}^{\mathsf{r}}_t$ is the continuous valued recurrent input at time t. At each time, the neurons independently generate Poisson distributed spike counts $\mathbf{r}_t$ (over the interval $[t-\Delta t, t]$) from the instantaneous rate $\boldsymbol{\lambda}_t$. The parameters w_E and σ_r denotes the excitatory recurrent weight and recurrent variability, respectively. To embed Gibbs sampling of stimulus and context features in the recurrent network we need to address two critical issues. First, we need to determine how the probability distributions (Eqs. 15-16) are represented in neuronal responses. Second, we need to determine how the requisite sampling could be implemented in the network, i.e., what response variability provides the correct variability for sampling stimulus feature?

Representation in neuronal activity

We require the network to generate and maintain a representation, in which the recurrent input $\mathbf{u}^{\mathsf{r}}$ represents the context feature z, and the spiking response $\mathbf{r}$ represents the stimulus feature s. Specifically, we assume that when fixing the spiking activity $\mathbf{r}_{t-\Delta t}$ which represents stimulus feature sample $\tilde{s}_{t-\Delta t}$, the recurrent inputs, $\mathbf{u}^{\mathsf{r}}_t$, have independent Poisson-like variability (the variance equals

the mean), and its mean, $\langle \mathbf{u}^{\mathsf{r}}_t \rangle$, has a Gaussian profile over $\tilde{s}_{t-\Delta t}$,

$$\begin{aligned} p(\mathbf{u}^{\mathsf{r}}_t|\tilde{s}_{t-\Delta t}) &= \textstyle\prod_{j=1}^{N_E} \mathcal{N}[\mathbf{u}^{\mathsf{r}}_{tj}|\langle \mathbf{u}^{\mathsf{r}}_{tj}|\tilde{s}_{t-\Delta t}\rangle, \langle \mathbf{u}^{\mathsf{r}}_{tj}|\tilde{s}_{t-\Delta t}\rangle], \\ \langle \mathbf{u}^{\mathsf{r}}_{tj}|\tilde{s}_{t-\Delta t}\rangle &= U^{\mathsf{r}} \exp[\mathbf{h}_j(\tilde{s}_{t-\Delta t})] = U^{\mathsf{r}} \exp[-(\theta_j - \tilde{s}_{t-\Delta t})^2/2a^2], \end{aligned} \tag{23}$$

Eq. (23) also assumes $\mathbf{u}^{\mathsf{r}}$ has the same tuning profile as the feedforward input (Eq. 10) (by sharing the same basis function $\mathbf{h}(s)$), and hence the recurrent and feedforward inputs received by the same j-th neuron have the same preferred stimulus θ_j. It is important to point out that the independent assumption of $\mathbf{u}^{\mathsf{r}}$ given stimulus feature sample $\tilde{s}_{t-\Delta t}$ doesn't mean an independent $\mathbf{u}^{\mathsf{r}}$ conditioned on an external stimulus feature s as observed in experiment. Due to the nature sampling dynamics the recurrent input $\mathbf{u}^{\mathsf{r}}$ fluctuates in the stimulus feature subspace, correlating activity through an external stimulus feature (Eq. 7).

Since the considered network is linear (Eq. 21), if we only consider the linear information of s contained in $\mathbf{u}^{\mathsf{r}}$ (regardless of the nonlinear information [70]), then the likelihood of s given a recurrent input $\mathbf{u}^{\mathsf{r}}$ can be normalized into a Gaussian distribution as follows,

$$\begin{aligned} p(s|\mathbf{u}^{\mathsf{r}}) &\propto \exp\left[\mathbf{h}(s)^\top \mathbf{u}^{\mathsf{r}}\right] = \exp\Big[-\textstyle\sum_j \mathbf{u}^{\mathsf{r}}_j \frac{(\theta_j - s)^2}{2a^2}\Big], \\ &= \mathcal{N}(s|z, \Lambda_s^{-1}). \end{aligned} \tag{24}$$

Comparing Eqs. (24) and (9) we see that the recurrent input $\mathbf{u}^{\mathsf{r}}$ can represent a conditional distribution of stimulus feature s, whose mean and precision of the prior can be linearly read out from $\mathbf{u}^{\mathsf{r}}$,

$$\Lambda_s = a^{-2} n_r = a^{-2} \textstyle\sum_j \mathbf{u}^{\mathsf{r}}_j, \quad z = n_r^{-1} \textstyle\sum_j \mathbf{u}^{\mathsf{r}}_j \theta_j. \tag{25}$$

Because the recurrent and feedforward inputs represent the distributions in the same way by design (see Eqs. 11 and 24), a convenient consequence for neural implementation is that linear sum of feedforward and recurrent input to give instantaneous firing rate $\boldsymbol{\lambda}_t$ (Eq. 21) computes the conditional distribution $p(s|\tilde{z}_t, \mathbf{u}^{\mathsf{f}})$ (Eq. 15),

$$\begin{aligned} \ln p(s|\tilde{z}_t, \mathbf{u}^{\mathsf{f}}) &= \ln p(x|s) + \ln p(s|\tilde{z}_t), \\ \ln \mathcal{N}\left(s|\hat{s}_t, \Lambda^{-1}\right) &= \mathbf{h}(s)^\top \mathbf{u}^{\mathsf{f}} + \mathbf{h}(s)^\top \mathbf{u}^{\mathsf{r}}_t = \mathbf{h}(s)^\top \boldsymbol{\lambda}_t. \end{aligned}$$

Thus, the instantaneous firing rate $\boldsymbol{\lambda}_t$ represents the conditional distribution $p(s|\tilde{z}_t, \mathbf{u}^{\mathsf{f}})$, with the precision and the mean of the conditional distribution being similarly read out as $\Lambda = a^{-2} \sum_j \boldsymbol{\lambda}_{tj}$ and $\hat{s}_t = \sum_j \boldsymbol{\lambda}_{t,j} \theta_j / \sum_j \boldsymbol{\lambda}_{t,j}$. Table 2 lists how the population inputs, and instantaneous firing rate in the recurrent network represent the parameters of the distributions. Meanwhile, the average

Table 2: Representation in neural population inputs and responses.

Population input/response	$\boldsymbol{\lambda}_t$	$\mathbf{u}^{\mathsf{f}}$	$\mathbf{u}^{\mathsf{r}}_t$
Distribution	$p(s\|\tilde{z}_t, x)$	$p(x\|s)$	$p(s\|\tilde{z}_t)$
Mean	$\hat{s}_t$	μ_{f}	$\tilde{z}_t$
Precision	Λ	Λ_{f}	Λ_s

(tuning) of $\boldsymbol{\lambda}$ can be computed as,

$$\begin{aligned}\langle \boldsymbol{\lambda}_j(s)\rangle &= \langle \mathbf{u}^{\mathsf{f}}_j(s)\rangle + \langle \mathbf{u}^{\mathsf{r}}_j(s)\rangle,\\ &= (U^{\mathsf{f}} + U^{\mathsf{r}})\exp[\mathbf{h}_j(s)],\\ &= R\exp[-(s-\theta_j)^2/2a^2].\end{aligned} \tag{26}$$

Poisson spike generation samples stimulus feature subspace

We demonstrate that the internal Poisson spike generation effectively samples stimulus feature s from the conditional distribution $p(s|\tilde{z}_t, \mathbf{u}^{\mathsf{f}})$, as long as the instantaneous firing rate $\boldsymbol{\lambda}_t$ represents $p(s|\tilde{z}_t, \mathbf{u}^{\mathsf{f}})$ as in Eq. (4). Specifically, we will show the empirical distribution of stimulus sample $\tilde{s}_t$ read out from $\mathbf{r}_t$ is consistent with $p(s|\tilde{z}_t, \mathbf{u}^{\mathsf{f}})$ represented in $\boldsymbol{\lambda}_t$. For simplicity, suppose the instantaneous firing rate $\boldsymbol{\lambda}_t$ is close to its trial average, i.e., a smooth Gaussian profile (Eq. 26). This can be satisfied with a sufficient long time window or sufficiently large neural population, so that $\boldsymbol{\lambda}_{tj}(\hat{s}_t) \approx \langle \boldsymbol{\lambda}_{tj}(\hat{s}_t)\rangle = R\exp[-(\hat{s}_t-\theta_j)^2/2a^2]$. When generating Poisson spikes $\mathbf{r}_t$ from $\boldsymbol{\lambda}_t(\hat{s}_t)$ (Eq. 22), the likelihood of an emitted spike $\mathbf{r}_t$ given $\hat{s}_t$ is proportional to a Gaussian distribution,

$$p(\mathbf{r}_t|\hat{s}_t) = \prod_{j=1}^{N} \text{Poisson}\left[\mathbf{r}_{tj}|\boldsymbol{\lambda}_{tj}(\hat{s}_t)\right] \propto \mathcal{N}\left(\tilde{s}_t|\hat{s}_t, \Lambda_{\mathbf{r}}^{-1}\right), \tag{27}$$

with precision $\Lambda_{\mathbf{r}} = a^{-2}n_r = a^{-2}\sum_j \mathbf{r}_{tj}$ and $\tilde{s}_t = n_{\mathbf{r}}^{-1}\sum_j \mathbf{r}_{tj}\theta_j$. Comparing the empirical distribution of stimulus feature sample $\tilde{s}_t$ collected over trials (Eq. 27), and the conditional distribution $p(s|\tilde{z}, \mathbf{u}^{\mathsf{f}})$ (Eq. 4), we see they are both Gaussian distributions with the same mean $\hat{s}_t$ but different precision ($\Lambda_{\mathbf{r}}$ versus Λ). Nevertheless, the precision of the empirical distribution, $\Lambda_{\mathbf{r}}$, is asymptotically equal to the precision of conditional distribution, Λ, in the limit of large neural population or in a long time window (see Supplementary Information for a rigorous math derivation),

$$\Lambda_{\mathbf{r}} = a^{-2}\sum_j \mathbf{r}_{tj} \approx \Lambda = a^{-2}\sum_j \boldsymbol{\lambda}_{tj}.$$

Combined together, this suggests the Poisson spike generation asymptotically sample a stimulus feature in the limit of large neural population of long time window, which is also confirmed by simulation (Fig. 3h).

Recurrent connections store the association between stimulus and context

The representation of context feature sample in recurrent input $\mathbf{u}^{\mathsf{r}}$ (Eqs. 23 and 24) put constraints on the recurrent weight w_E and recurrent variability σ_r. Because the stimulus tuning of $\mathbf{u}^{\mathsf{r}}_t$ and $\mathbf{r}_t$ have the same width a (Eqs. 23 and 26), every E neuron could be only self-connected, otherwise the tuning of $\mathbf{u}^{\mathsf{r}}$ would be larger than that of $\mathbf{r}$. Combining Eqs. (23 and 26) and Table 2, it can be checked when every E neuron is self-connected with the same weight w_E, an optimal weight satisfying the self-consistent requirement is

$$w_E^* = \frac{\mathbf{u}^{\mathsf{r}}_j}{\mathbf{r}_j} = \frac{U^{\mathsf{r}} e^{-(\theta_j - s)^2/2a^2}}{R e^{-(\theta_j - s)^2/2a^2}} = \frac{\Lambda_s}{\Lambda_x + \Lambda_s}, \tag{28}$$

which yields Eq. (6) in the main text. Note that the simplification to strictly self-connections results from the simplified network model only consisting of E neurons (Eq. 21). While this provides theoretical insight it does not need to be held in a network having both excitatory and inhibitory neurons (as considered in Fig. 1), as will be shown later.

Another assumption in recurrent connections is that $\mathbf{u}^{\mathsf{r}}$ has independent Poisson-like variability (the variance is the same as mean) given spiking responses $\mathbf{r}_{t-\Delta t}$ emitted at last time step, which constrains the recurrent variability σ_r (Eq. 21).

$$\mathbf{u}^{\mathsf{r}}_t = \hat{\mathbf{u}}_{rt} + \sqrt{\hat{\mathbf{u}}_{rt}}\boldsymbol{\xi}_t, \quad \hat{\mathbf{u}}_{rt} = w_E^* \mathbf{r}_{t-\Delta t}, \tag{29}$$

where $\boldsymbol{\xi}_t$ is a N_E dimensional Gaussian white noise with $\langle \boldsymbol{\xi}_t(i)\boldsymbol{\xi}_{t'}(j)\rangle = \delta_{ij}\delta(t-t')$, where δ_{ij} and $\delta(t-t')$ are Kronecker and Dirac delta functions respectively.

Analysis of neuronal response conditional distributions

In this section we calculate the distribution of neuronal responses $\mathbf{r}$ given an external stimulus feature s; this can be compared directly with experimental observations [27]. For a fixed external stimulus s, the neuronal response $\mathbf{r}$ fluctuates over time/trial due to both sensory transmission noise described by $p(\mathbf{u}^{\mathsf{f}}|s)$ (Eq. 13 and Fig. 2b), as well as the internal variability in the network described by $p(\mathbf{r}|\mathbf{u}^{\mathsf{f}})$ (Fig. 3a). Taken together, the recorded distribution of $\mathbf{r}$ in response to an external stimulus s is then constructed as

$$p(\mathbf{r}|s) = \int p(\mathbf{r}|\mathbf{u}^{\mathsf{f}}) p(\mathbf{u}^{\mathsf{f}}|s) d\mathbf{u}^{\mathsf{f}}. \tag{30}$$

To compute the response distribution from internal variability $p(\mathbf{r}|\mathbf{u}^{\mathsf{f}})$, we decompose it into the Poisson spiking variability $p(\mathbf{r}|\boldsymbol{\lambda})$ (Eq. 22) and the variability of instantaneous firing rate, $p(\boldsymbol{\lambda}|\mathbf{u}^{\mathsf{f}})$,

yielding

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

For simplicity, we only compute the covariability of $p(\mathbf{r}|\boldsymbol{\lambda})$ along the stimulus feature space, since how stimuli are being represented in population responses is our main interest. Through approximating Poisson spiking variability $p(\mathbf{r}|\boldsymbol{\lambda})$ by a multivariate normal distribution (Eq. 22), and considering the limit of weak fluctuations in $\boldsymbol{\lambda}$ along the stimulus feature subspace, the response distribution with a fixed feedforward input $\mathbf{u}^{\mathsf{f}}$ can be computed approximately as (see math details in Supplementary Information),

$$p(\mathbf{r}|\mathbf{u}^{\mathsf{f}}) \approx \mathcal{N}\big[\mathbf{r}\big|\boldsymbol{\lambda}(s), \mathrm{diag}(\boldsymbol{\lambda}(s)) + V(\hat{s}|\mu_{\mathsf{f}})\boldsymbol{\lambda}_s'^{\top}\boldsymbol{\lambda}_s'\big], \quad \text{where } s = \mu_{\mathsf{f}}. \tag{31}$$

The covariance structure of the neuronal response includes two terms. The first term, $\mathrm{diag}(\boldsymbol{\lambda}(s))$, is a diagonal matrix whose entries are the vector $\boldsymbol{\lambda}$ denoting the (independent) Poisson spiking variability (Eq. 13). The second term, $V(\hat{s}|\mu_{\mathsf{f}})\boldsymbol{\lambda}_s'^{\top}\boldsymbol{\lambda}_s'$, captures the covariability due to firing rate fluctuations along the stimulus feature subspace, where $\boldsymbol{\lambda}_s' = d\boldsymbol{\lambda}_s/ds$ is the derivative of $\boldsymbol{\lambda}_s$ over the stimulus feature s. The covariance $\boldsymbol{\lambda}_s'^{\top}\boldsymbol{\lambda}_s'$ is often termed differential correlations [2, 25]. With the Gaussian tuning of $\boldsymbol{\lambda}_s$ (Eq. 26), $\boldsymbol{\lambda}_s'^{\top}\boldsymbol{\lambda}_s'$ exhibits anti-symmetric structure (Fig. 6b), which has been previously remarked (e.g., [25, 32, 71–73]).

In Eq. (31), $V(\hat{s}|\mu_{\mathsf{f}})$ characterizes the fluctuation of instantaneous firing rate $\boldsymbol{\lambda}$ along the stimulus feature subspace stemming from the internal variability in the network, i.e., the amount of internal differential correlations, where $\hat{s}$ and μ_{f} denote the position of $\boldsymbol{\lambda}$ and $\mathbf{u}^{\mathsf{f}}$ on the stimulus feature subspace respectively (Table 2). It can be solved from the dynamics of Gibbs sampling (Eq. 17) and the network structure (Eq. 6) as,

$$V(\hat{s}|\mu_{\mathsf{f}}) = \frac{\Lambda_s}{\Lambda_{\mathsf{f}}(\Lambda_{\mathsf{f}} + \Lambda_s)} = a^2 n_{\mathsf{f}}^{-1} w_E^*. \tag{32}$$

It is noteworthy that $V(\hat{s}|\mu_{\mathsf{f}})$ is constrained by the network connections, in that it increases with the prior precision Λ_s stored by the recurrent weight w_E^* in the network. Consequently, any internally generated differential correlations can be regarded as a neural signature of network performing sampling in stimulus feature subspace.

Finally, by substituting Eqs. (31) and (13) into Eq. (30), the response distribution conditioned on external stimulus feature s, $p(\mathbf{r}|s)$, is computed as,

$$p(\mathbf{r}|s) \approx \mathcal{N}\big[\mathbf{r}\big|\boldsymbol{\lambda}(s), \mathrm{diag}(\boldsymbol{\lambda}(s)) + V(\hat{s}|s)\boldsymbol{\lambda}_s'^{\top}\boldsymbol{\lambda}_s'\big].$$

Here the variance $V(\hat{s}|s)$ on the stimulus feature subspace is a mixture of internal variability and

sensory noise, and can be solved according to Eq. (17) as,

$$V(\hat{s}|s) = V(\hat{s}|\mu_{\mathsf{f}}) + V(\mu_{\mathsf{f}}|s) = \frac{\Lambda_s}{\Lambda_{\mathsf{f}}(\Lambda_{\mathsf{f}} + \Lambda_s)} + \frac{1}{\Lambda_{\mathsf{f}}} = a^2 n_{\mathsf{f}}^{-1}(w_E^* + 1). \tag{33}$$

It is useful to note that $V(\mu_{\mathsf{f}}|s)$ denotes the fluctuations of neuronal activity on the stimulus feature subspace due to the sensory transmission noise (Eq. 13).

A spiking network model with excitatory and inhibitory Poisson neurons

The goal of the network presented above (Eqs. 21-22) is to provide a biologically plausible mechanism to implement Gibbs sampling. To test the validity of the proposed mechanism we simulate a recurrent network model which is consistent with many past cortical models (e.g., [31–33]). The network consists of N_E excitatory (E) and N_I inhibitory (I) spiking neurons, with each neuron modeled as a Hawkes process [69, 74]. At time t, the spiking response of neuron j in population $a = \{E, I\}$, $\mathbf{r}_j^a(t)$, is drawn from a Poisson distribution with instantaneous firing rate $\boldsymbol{\lambda}_j^a(t)$,

$$\mathbf{r}_j^a(t) \sim \text{Poisson}[\boldsymbol{\lambda}_j^a(t)]. \tag{34}$$

The firing rate $\boldsymbol{\lambda}_j^a(t)$ is the sum of feedforward input $\mathbf{u}_j^{a\mathsf{f}}(t)$ and recurrent input $\mathbf{u}_j^{a\mathsf{r}}(t)$,

$$\boldsymbol{\lambda}_j^a(t) = \mathbf{u}_j^{a\mathsf{f}}(t) + \mathbf{u}_j^{a\mathsf{r}}(t).$$

The feedforward inputs are filtered spikes from upstream neurons,

$$\mathbf{u}_j^{ab}(t) = \sum_n \eta\big(t - t_{jn}^{\mathsf{f}}\big),$$

where t_{jn}^{f} is the time of the n-th spike received by neuron j of type a from feedforward inputs. Here $\eta(t)$ is the synaptic input profile which is modeled as

$$\eta(t) = \exp(-t/\tau_d)/\tau_d, \quad \text{for} \quad t > 0. \tag{35}$$

Throughout, we set the synaptic time constant $\tau_d = 2\text{ms}$. The recurrent input received by neuron j in population a is defined by

$$\begin{aligned} \mathbf{u}_j^{a\mathsf{r}}(t) &= \bar{\mathbf{u}}_j^{a\mathsf{r}}(t) + \sqrt{[\bar{\mathbf{u}}_j^{a\mathsf{r}}(t)]_+}\,\xi(t), \\ \bar{\mathbf{u}}_j^{a\mathsf{r}}(t) &= \sum_{b=\{E,I\}} \sum_{k=1}^{N_b} \frac{J_{jk}^{ab}}{\sqrt{N}} \sum_n \eta(t - t_{kn}^b), \end{aligned} \tag{36}$$

where $\bar{\mathbf{u}}_j^{a\mathsf{r}}(t)$ is the mean recurrent input at time t given the neuronal activities of the presynaptic neurons. It is noteworthy that the recurrent input in current network is explicitly corrupted by a noise whose variance is the same as the mean of recurrent input. Noise in the recurrent input is a necessary condition for implementing sampling-based inference in network dynamics as derived in Eq. (29). In reality, the recurrent noise may be generated by the chaotic state in network dynamics [60, 62] or synaptic noise [75]. In the above equation the function $[\cdot]_+$ rectifies the negative input, and $\xi(t)$ is a standard Gaussian white noise. 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 nth 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 [76]. Finally, the synaptic input profile of the recurrent input $\eta(t)$ is the same as the one in feedforward input (Eq. 35) for convenience.

To model the coding of a circular stimulus feature such as orientation, the excitatory neurons are arranged on a ring [32, 71]. Each excitatory neuron j has its preferred stimulus θ_j, and the preferred stimuli of all excitatory neurons are uniformly distributed from $-180°$ to $180°$, consistent with the range of latent features (Eq. 8). The inhibitory neurons do not have a preference over stimulus orientation, and their role is to keep the stability of network responses. The recurrent connections between E neurons are modeled as a Gaussian function decaying with the distance between two neurons on the ring,

$$J_{jk}^{EE} = \frac{w_{EE}L}{\sqrt{2\pi}a} \exp\left[-\frac{(\theta_j - \theta_k)^2}{2a^2}\right], \tag{37}$$

where a periodic boundary is imposed on the Gaussian function. The weight w_{EE} denotes the average connection strength of all E to E connections. The parameter $a = 40°$ is the footprint of connectivity in feature space (i.e the ring). Here $L = 360°$ is the length of the ring manifold (Eq. 8); the multiplication with L in the above equation sets the sum of all E to E connection strengths to be $N_E w_{EE}$. Moreover, the E neurons and I 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 connections of the same type have the same weight, i.e., $J_{jk}^{EI} = w_{EI}$, $J_{jk}^{IE} = w_{EE}$ and $J_{jk}^{II} = w_{II}$. To simplify the network parameter 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 $w_{II} = w_{EI} \equiv w_I$.

The feedforward inputs applied to E neurons are independent Poisson spikes as described by Eq. (10), with input firing rate $\langle \mathbf{u}_j^{E\mathsf{f}}(s)\rangle = U^{\mathsf{f}} e^{-(s-\theta_j)^2/(4a^2)}$. Meanwhile, the inhibitory neurons also receive feedforward inputs of independent Poisson spikes. The firing rate of feedforward spikes applied to every I neuorn is proportional to the overall feedforward rate applied to E neurons, in

order to keep the baseline activity of E population not increase with stimulus strength U^{f},

$$\langle \mathbf{u}_j^{I\mathsf{f}} \rangle = \frac{w_{I\mathsf{f}}}{N_I} \sum_{j=1}^{N_E} \langle \mathbf{u}_j^{E\mathsf{f}}(s) \rangle. \tag{38}$$

In the simulations shown in Fig. (4-8), the network has $N_E = 180$ E neurons and $N_I = 45$ I neurons, while the number of neurons are increased by a fixed factor in Fig. 1e. The ratio between the average connection from I neurons and the one from E neurons is kept fixed with $w_I/w_E = 5$. The feedforward weight to I neurons $w_{I\mathsf{f}} = 0.8$. The network dynamics are simulated by using an Euler method with a time step of 0.5ms. Given the stimulus orientation s and strength U^{f}, the feedforward input rate is fixed and the feedforward spiking input is drawn from the Poisson distribution at each time step.

Estimating the information content in network simulation

To estimate the information context contained in network responses, we fix an external stimulus feature s in the world, generate independent Poissonian feedforward inputs on every time step, and simulate the network dynamics accordingly (Eqs. 34-38). For simplicity, the external stimulus feature s is set as zero without loss of generality, since the feedforward inputs and network responses are translation-invariant with the value of stimulus feature s.

In estimating the mutual information in network responses, we simulate the spiking network (Eqs. 34-38) for 2000 seconds. For each time step, the samples of the stimulus feature and context feature are read out by a population vector from the instantaneous spiking activities of E neurons, $\mathbf{r}^E$, and recurrent input of E neurons, $\mathbf{u}^{E\mathsf{r}}$ respectively (Table 2),

$$\tilde{s}_t = \textstyle\sum_j \mathbf{r}_j^E(t)\theta_j / \sum_j \mathbf{r}_j^E(t), \qquad \tilde{z}_t = \sum_j \mathbf{u}_j^{E\mathsf{r}}(t)\theta_j / \sum_j \mathbf{u}_j^{E\mathsf{r}}(t).$$

Then the joint distribution of samples $\tilde{s}$ and $\tilde{z}$ is approximated as a bivariate normal distribution, and its mean, $\boldsymbol{\mu}_q$, and covariance matrix, $\mathbf{K}_q$, are computed. Note that the posterior (Eq. 14) is defined on a given feedforward input $\mathbf{u}^{\mathsf{f}}$ rather than a given external stimulus feature s. The fluctuation of $\mathbf{u}^{\mathsf{f}}$ over time will bring extra variability in the distribution of samples generated by the network compared with posterior. If a network optimally samples posterior, the mean and precision matrix of samples read out from the network in response to a fixed external stimulus feature s are,

$$\boldsymbol{\mu}_q = (\langle \mu_{\mathsf{f}} \rangle, \langle \mu_{\mathsf{f}} \rangle)^\top = (s, s)^\top, \quad \mathbf{K}_q = \begin{pmatrix} \frac{1}{2}\Lambda_{\mathsf{f}} + \Lambda_s & -\Lambda_s \\ -\Lambda_s & \Lambda_s \end{pmatrix}.$$

Comparing above equation with Eq. (14), we see the temporal fluctuation of $\mathbf{u}^{\mathsf{f}}$ only changes the first element in precision matrix $\mathbf{K}_q$. To cancel its effect, we add the first element of $\mathbf{K}_q$ with $\Lambda_{\mathsf{f}}/2$,

where Λ_{f} is read out from feedforward inputs (Eq. 12). Then the mean and corrected precision matrix of samples are substituted into Eq. (20) to get the mutual information contained in the population activities about $\Theta = \{s, z\}$.

The linear Fisher information of stimulus feature s in responses of E neurons, $\mathbf{r}^E$, is estimated by using an bias-correction algorithm [51]. A pair of stimuli $s_\pm = \pm 1^\circ$ are individually presented to the network for $T = 1 \times 10^4$ trials with each trial lasting 200ms, and the empirical mean and covariance of E neurons' responses $\mathbf{r}^E$ under each stimulus are recorded. Then the bias-correction estimate of linear Fisher information is,

$$\mathcal{I}_{bc}(s) = \frac{d\langle \mathbf{r}^E \rangle}{ds}^\top \Sigma(\mathbf{r}^E)^{-1} \frac{d\langle \mathbf{r}^E \rangle}{ds} \frac{2T - N - 3}{2T - 2} - \frac{2N}{T ds^2}, \tag{39}$$

where $d\langle \mathbf{r} \rangle = \langle \mathbf{r}^E_+ \rangle - \langle \mathbf{r}^E_- \rangle$, $ds = s_+ - s_-$, and $\langle \cdot \rangle$ represents the average over trials. $\Sigma(\mathbf{r}^E)$ is the covariance matrix of E neurons' responses. N is the number of neurons in the network.

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.).

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

[3] Michael N Shadlen and William T Newsome. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *Journal of neuroscience*, 18(10):3870–3896, 1998.

[4] 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.

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

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

[7] Zachary F Mainen and Terrence J Sejnowski. Reliability of spike timing in neocortical neurons. *Science*, 268(5216):1503–1506, 1995.

[8] 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.

[9] 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.

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

[11] L Federico Rossi, Kenneth D Harris, and Matteo Carandini. Excitatory and inhibitory intracortical circuits for orientation and direction selectivity. *bioRxiv*, page 556795, 2019.

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

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

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

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

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

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

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

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

[20] David Mumford. Neuronal architectures for pattern-theoretic problems. *Large-scale neuronal theories of the brain*, pages 125–152, 1994.

[21] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

[22] Wen-Hao Zhang, Aihua Chen, Malte J Rasch, and Si Wu. Decentralized multisensory information integration in neural systems. *The Journal of Neuroscience*, 36(2):532–547, 2016.

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

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

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

[26] Donald H Perkel and Theodore H Bullock. Neural coding. *Neurosciences Research Program Bulletin*, 1968.

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

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

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

[30] William H Bosking, Ying Zhang, Brett Schofield, and David Fitzpatrick. Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *Journal of neuroscience*, 17(6):2112–2127, 1997.

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

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

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

[34] 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.

[35] 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.

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

[37] Konrad P Körding, Ulrik Beierholm, Wei Ji Ma, Steven Quartz, Joshua B Tenenbaum, and Ladan Shams. Causal inference in multisensory perception. 2007.

[38] 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.

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

[40] Yang Dan, Joseph J Atick, and R Clay Reid. Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *Journal of Neuroscience*, 16(10):3351–3362, 1996.

[41] Robert A Jacobs. Optimal integration of texture and motion cues to depth. *Vision Research*, 39(21):3621–3629, 1999.

[42] David Alais and David Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14(3):257–262, 2004.

[43] Yong Gu, Dora E Angelaki, and Gregory C DeAngelis. Neural correlates of multisensory cue integration in macaque mstd. *Nature Neuroscience*, 11(10):1201–1210, 2008.

[44] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

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

[46] 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.

[47] 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.

[48] 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.

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

[50] John P Cunningham and M Yu Byron. Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, 17(11):1500–1509, 2014.

[51] 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.

[52] Robert L Goldstone. Perceptual learning. *Annual review of psychology*, 49(1):585–612, 1998.

[53] Amy M Ni, Douglas A Ruff, Joshua J Alberts, Jen Symmonds, and Marlene R Cohen. Learning and attention reveal a general relationship between population activity and behavior. *Science*, 359(6374):463–465, 2018.

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

[55] 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.

[56] 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.

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

[58] 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.

[59] 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.

[60] Carl Van Vreeswijk and Haim Sompolinsky. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274(5293):1724–1726, 1996.

[61] Michael London, Arnd Roth, Lisa Beeren, Michael Häusser, and Peter E Latham. Sensitivity to perturbations in vivo implies high noise and suggests rate coding in cortex. *Nature*, 466(7302):123–127, 2010.

[62] Ashok Litwin-Kumar and Brent Doiron. Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nature neuroscience*, 15(11):1498, 2012.

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

[64] J Hans Van Hateren and Arjen van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1394):359–366, 1998.

[65] Thomas L Griffiths, Edward Vul, and Adam N Sanborn. Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, 21(4):263–268, 2012.

[66] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[67] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

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

[69] Volker Pernice, Benjamin Staude, Stefano Cardanobile, and Stefan Rotter. How structure determines correlations in neuronal networks. *PLoS computational biology*, 7(5):e1002059, 2011.

[70] Si Wu, Shun-ichi Amari, and Hiroyuki Nakahara. Information processing in a neuron ensemble with the multiplicative correlation structure. *Neural Networks*, 17(2):205–214, 2004.

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

[72] 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.

[73] 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.

[74] 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.

[75] 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.

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