bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Recurrent 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.

[*]Equal contribution.

Correspondence should be addressed to: 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 recurrent connections 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 *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. This provides experimentally testable predictions for how population connectivity and response variability are related to the structure of latent external stimuli.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 reliably represent and interpret stimuli. A wrong decision based on a faulty stimulus representation could endanger an animal's life. The sensory cortex is essential for the representation of the sensory world[1;2], with individual neurons tuned to stimulus features so that the population as a whole can code for a richly structured sensory scene. However, two central cortical circuit features remain to be properly integrated into a coherent representation framework. First, neuronal responses in sensory cortices are often noisy, with the same stimulus evoking starkly different responses across trials[3;4]. If such response variability is unrelated to sensory stimuli, it would be detrimental to their representation in cortical circuits[2]. Why cortical responses are so variable when isolated cortical neurons can respond reliably remains a mystery[5]. 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[6–8]. While such recurrent connections are organized about a stimulus feature axis[9;10], it is not obvious whether and how their presence improves overall representation. Here we propose a biologically motivated inference coding scheme where these two cortical circuit features are essential for optimal stimulus representation.

Numerous studies have shown that cortical sensory processing can be understood in terms of Bayesian inference[11–17]. In such models, sensory cortex infers and synthesizes a description of the possible stimuli that could have given rise to the sensory inputs, rather than generating a single stimulus estimate. To perform Bayesian inference, the sensory cortex needs to store a model describing how external stimuli and the resulting sensory inputs are generated. Once a sensory input is received, cortical dynamics inverts this internal model in a process termed 'analysis-by-synthesis'[13]. The neural response represents a posterior distribution over the input stimulus and unobserved (latent) variables[16;17]. We propose that recurrent (local) connections in cortical circuits represent such an internal model[18], and that internally generated variability is a source of fluctuations needed for generating samples from the posterior distribution,[17;19;20]. The challenge in this approach is to determine how the structure of connectivity and population variability should be set so as to accurately represent the posterior distribution.

To tackle this challenge we consider a hierarchical generative model[13], where sensory inputs are generated from a local stimulus feature and a global, latent context feature. We assume that the local and global aspects of a scene are correlated, so that knowledge of one provides information about 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. 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. The best approximation of the posterior over the stimulus and context features is achieved in a network with recurrent connections whose strength increases with the degree of correlation between context and local sensory inputs. In other words, a strongly coupled network is needed for optimal inference in a well-structured world, whereas an unconnected network would work best in a world without structure.

We provide 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;21]. 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. Thus measurable differential noise correlations can be used as a proxy for strong recurrent connections organized along the stimulus tuning dimension. These correlations can be controlled by adjusting the structure of the external world.

Results

Recurrent circuitry does not improve conventional neural codes

Neural coding is the process by which a stimulus feature, s, is encoded in neuronal activity, $\mathbf{r}$, which is subsequently decoded to provide an estimate of the feature, $\hat{s}$ (Fig. 1a, top)[22]. In this view, it is reasonable to expect that neuronal circuitry is adapted to provide an accurate feature estimate. However, we next show that in simple coding schemes recurrent connectivity at best does not affect stimulus estimation, and at worst degrades such estimates, raising the question of its purpose.

In classical models, stimulus encoding is performed by a neuronal population, with individual neurons tuned to a preferred stimulus value. The preferred values across the population cover a range of potential stimuli[22–24] (Fig. 1b, bottom). If the stimulus, s, ranges over a periodic domain (such as the orientation of a bar in a visual scene, or the direction of an arm reach) then it is commonly assumed that the neurons' preferred stimuli are distributed on a ring (Fig. 1b, top). Neurons with similar feature preferences tend to be coupled more strongly[9;10], so that network connectivity aligns with the ring topology (Fig. 1b, top, arrows). We implemented such a model with structured connectivity between excitatory (E) neurons, and unstructured inhibition (I) that acts only to stabilize network activity[25]. Hence, only E neurons in the network are tuned to a

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

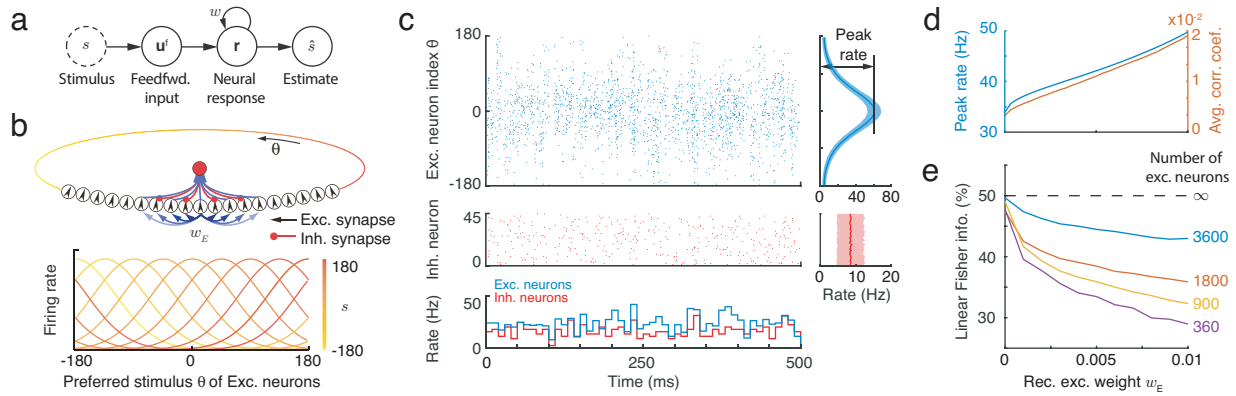

Figure 1: A network with structured recurrent connections limits linear Fisher Information (LFI) about external stimuli. (a) A schematic diagram showing how a stimulus, s, is encoded in a neuronal response, $\mathbf{r}$. A stimulus estimate $\hat{s}$ can be decoded from $\mathbf{r}$. (b) A recurrent 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) Example realization of population spiking activity of excitatory and inhibitory neurons in the network. Right: the firing rate of neurons averaged over time, with shaded region encompassing one standard deviation from the mean. (d) The population responses measured in peak firing rate (left axis) and spike count correlations averaged over excitatory neurons (right axis). (e) For finite size networks (colored lines; ratio of excitatory to inhibitory neurons kept constant) the LFI decreases as w_E increases. In the limit of infinite network size LFI does not depend on w_E (dashed line).

stimulus feature (Fig. 1b, bottom). Further, for simplicity we assumed a linear (or linearized) neuronal transfer and synaptic interactions (Eqs. 19-20), and Poisson spike responses (Eqs. 31-34). We presented the network with a stimulus with feature $s = 0$, resulting in elevated activity of E neurons with the corresponding preference (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[3] (Fig. 1d). This canonical, mechanistic network model has been widely used to explore cortical network dynamics and coding[25–27].

We considered the impact of recurrent connectivity on the accuracy of the decoded feature estimate, $\hat{s}$, using a conventional coding framework. Towards this end, we measured the linear Fisher Information (LFI) of s in the population response of E neurons (see details in Eq. S24 in Supplementary Information). The inverse of the LFI provides a lower bound on the squared error between s and $\hat{s}$ made by a linear decoder[1;2;21–23;28], and it is closely related to an animal's performance in discrimination tasks widely studied in experiments[2;22]. We found that for networks with large numbers of decoded neurons LFI is not affected by the peak recurrent strength w_E (Fig. 1e, dashed line), and converges to the information about s provided by the feedforward input[21]. For the simple network and stimulus world we consider (Fig. 1a,b) this result can be readily understood: Since in our network a neuron's mean response is a linear function of its feedforward and recurrent inputs, the entire network acts as a linear filter of the stimulus s. An (invertible) linear filter can neither augment nor degrade the stimulus information available to be decoded by a linear readout (as LFI measures; see Eq. S15 in Supplementary Information). Thus,

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

LFI does not depend on the details of the filter, or in this case on the strength of recurrent coupling w_E. However, networks with a finite number of neurons produce internal noise which can degrade information about the stimulus, and hence reduce the LFI (Fig. 1e, colored lines)[21;28]. For such finite networks this noise is shared between neurons and thus an increase with coupling strength, w_E, in turn reducing LFI. In sum, recurrent connectivity does not improve stimulus representation in terms of increasing LFI. Since synaptic coupling is biologically expensive, 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[6–8;10]? There are at least three possibilities. First, while recurrent connectivity does degrade stimulus representation, it might underlie other stimulus processing functions essential for cortical computations. Second, LFI and the linear network interactions we assumed are overly simplistic, and nonlinear network interactions[28] or nonlinear decoders[29] may be required to properly model population coding, and the performance of neural codes may be optimized at nonzero coupling strengths. The final possibility is that it is the stimulus world we used to probe neural coding (Fig. 1a) that is overly simplistic, and recurrent cortical coupling in our linear network can improve cortical codes that represent more complex stimulus worlds. We next develop an extended framework that shows that the last possibility is plausible, but does not exclude the first two.

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, transmitted via a feedforward sensory pathway (Fig. 1a). However, numerous studies suggest that the brain is performing Bayesian inference[11–13;15;32;33], and thus synthesizes and estimates the probability of a range of possible latent stimuli. To accomplish this task neural circuits must approximate a posterior distribution over the latent stimulus from the sensory input, $\mathbf{u}^{\mathbf{f}}$, and any 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.

In order for the circuit computation to be non-trivial, the stimulus has to come from a world with sufficiently rich structure. We therefore consider a hierarchical generative model (Fig. 2a) that describes how a feedforward input, $\mathbf{u}^{\mathbf{f}}$, received by the cortical circuit is generated from stimulus and context features in the outside world. We 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. The structure of the world is described by the joint distribution, $p(s, z)$. In the example of a V1 hypercolumn the stimulus feature, s, could be the orientation of the visual texture within a classical receptive field (local information) of a hypercolumn of V1 neurons, while the orientation within

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

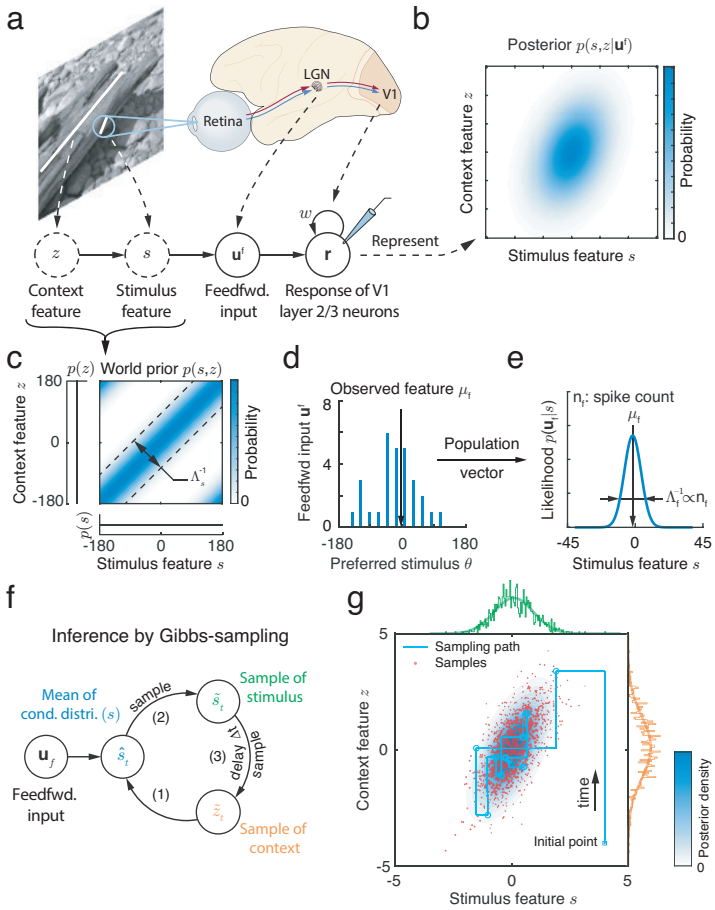

Figure 2: A hierarchical generative model and inference by Gibbs sampling. (a) An example of sensory feedforward input generation: The context feature, z, is the orientation of the tree trunk, and the stimulus feature, s, is the orientation of the bark texture located in the classical receptive field of a V1 hypercolumn. Feedforward input, $\mathbf{u}^f$, conveys stimulus information to V1 neurons. The evoked response, $\mathbf{r}$, represents the joint posterior over stimulus and context, $p(s, z|\mathbf{u}^f)$. Solid circles: observations and responses in the brain; dashed circles: latent variables in the world. (Natural image and brain schematic adapted from[30] and[31] respectively). (b) The joint posterior $p(s, z|\mathbf{u}^f)$ 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, $\mathbf{u}^f$, 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 $\mathbf{u}^f$ using a population vector (Eq. 12). (f) The posterior over context and stimulus can be approximated 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.

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 precision (inverse of the variance) equal to Λ_s. For simplicity, we assume that the prior distribution over the context feature, $p(z)$, is uniform (Fig. 2c). As we show in the Discussion, this assumption is not essential for our main conclusions. While the uniform

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

prior over the context feature, z, implies that the marginal prior over s, is also uniform, the joint prior over stimulus and context features, $p(s, z)$, can have non-trivial structure. Specifically, if the two features are positively 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, 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). The feedforward input, $\mathbf{u}^{\mathsf{f}}$, is represented by a vector of independent Poisson spike counts (Eq. 10) that are received by N_E excitatory (E) neurons in the recurrent circuit. The mean intensity of each component of $\mathbf{u}^{\mathsf{f}}$ has a Gaussian tuning profile over the stimulus feature, s, so that $\langle \mathbf{u}^{\mathsf{f}}_j \rangle \propto \exp[\mathbf{h}_j(s)] = \exp[-(s-\theta_j)^2/2a^2]$. 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 equal a constant, a[22;34;35]. We refer to the subspace in the space of sensory inputs 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[30].

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

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

The input vector, $\mathbf{u}^{\mathsf{f}}$, represents a Gaussian likelihood over the stimulus feature, s[34;35], i.e., $p(\mathbf{u}^{\mathsf{f}}|s) \propto \mathcal{N}(s|\mu_{\mathsf{f}}, \Lambda_{\mathsf{f}}^{-1})$. This is a consequence of the Gaussian tuning, and the assumption that the preferred stimuli of feedforward inputs are uniformly distributed. The likelihood of the stimulus, s, is thus 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[24;34] (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). Further, 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 approximate the joint posterior distribution over the stimulus features s and z from sensory input, $\mathbf{u}^{\mathsf{f}}$ (Fig. 2b).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sampling based inference of the joint 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 of perception, the computed approximation of this posterior captures an observer's perception of stimulus and context[11;12;14;16;32;33]. We next show how recurrent circuit dynamics can be used to approximately compute, and represent this posterior distribution[1;16;17].

We propose that Gibbs sampling, a stochastic algorithm which generates samples from a probability distribution[36], can be used to approximate the posterior over the features s and z given sensory input, $\mathbf{u}^{\mathsf{f}}$. This algorithm requires a mechanism for generating random numbers, and is thus well suited for implementation in cortical circuits given the prominent stochasticity of neuronal activity in cortex[13;17;19;37–40]. Gibbs sampling starts with an initial guess for the value of the two latent features, and proceeds by alternately fixing a sampled value of one 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, $(\tilde{s}_t, \tilde{z}_t)$ indexed by time t. We consider discretized time with $t = \Delta t, 2\Delta t, \ldots, T$, where Δt is the time increment between successive samples, T is the sampling duration, and thus $T/\Delta t$ are the number of samples. This discrete construction merely simplifies exposition; extensions to continuous time sampling are straightforward. 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)),

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

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

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

The samples are generated by alternately fixing the values of the two variables, so that sampling trajectories alternate between horizontal and vertical jumps in parameter space (cyan line in Fig. 2g). The set of sampled values (red dots in Fig. 2g) 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 and $\top$ denotes vector transpose. The distribution $q(s, z|\mathbf{u}^{\mathsf{f}})$ approximates the true posterior distributions $p(s, z|\mathbf{u}^{\mathsf{f}})$ (blue contour map in Fig. 2g, Eq. 14)[36]. Moreover, to approximate the marginal posterior distribution over one of the features, say $p(s|\mathbf{u}^{\mathsf{f}})$, we can keep only samples $\tilde{s}_t$ to obtain an approximating distribution $q(s|\mathbf{u}^{\mathsf{f}})$ (compare the two green lines at the margin in Fig. 2g).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

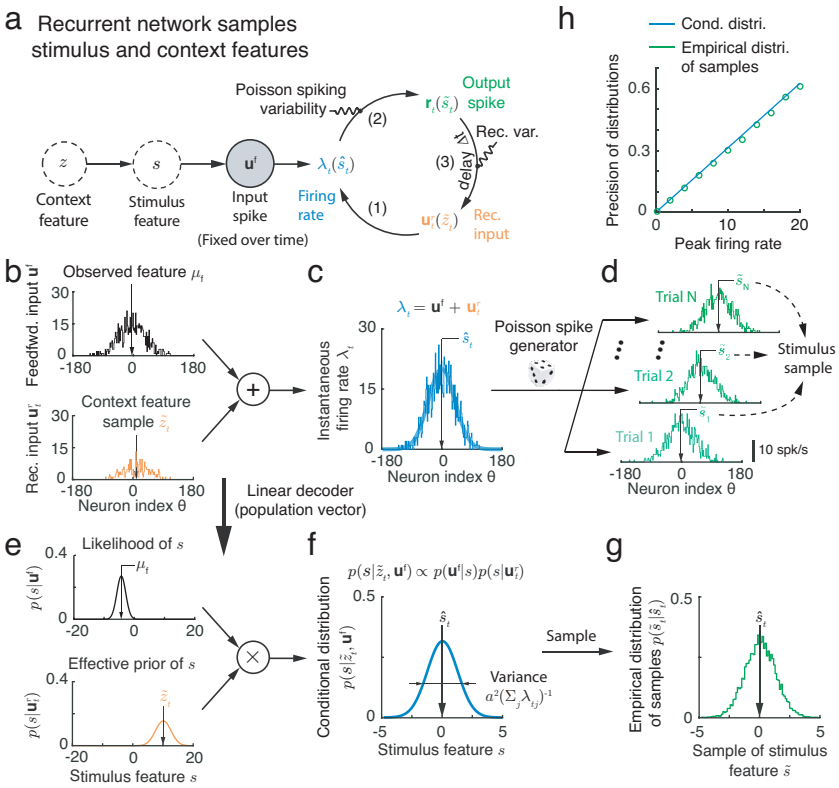

Figure 3: A recurrent circuit with internal Poisson variability can implement latent feature sampling. (a) Schematic of recurrent circuit dynamics, in which stimulus, s, and context, z, features are encoded respectively by the population response, $\mathbf{r}$, and recurrent inputs $\mathbf{u}^{\mathsf{r}}$. Poissonian spiking and recurrent variability produce 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 are 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 the conditional distribution of stimulus feature (f) and the variability of stimulus feature samples.

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 next show how the proposed Gibbs algorithm for alternate sampling of the stimulus feature, s, and the context feature, z can be naturally implemented in a recurrent network with internally generated variability. We first explain the mechanism using a linear network model consisting of only E neurons (Fig. 3a, Eqs. 19-20), and later demonstrate how it can be implemented in a network consisting of both E and I neurons (Fig. 1). The response of the E population at time t, $\mathbf{r}_t$, is a vector of spike counts, with each entry sampled from a Poisson distribution with parameters given by the vector of firing rates $\boldsymbol{\lambda}_t$. We assume that these firing rates are a sum of the sensory feedforward input, $\mathbf{u}^{\mathsf{f}}$ (Eq. 1), and recurrent inputs $\mathbf{u}^{\mathsf{r}}_t$.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We next need 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 a recurrent network can store an internal representation of the generative model by comparing Gibbs sampling (Fig. 2f) to recurrent network dynamics (Fig. 3a): 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, $\mathbf{u}^{\mathsf{r}}_t$. In the network the sum of feedforward and recurrent inputs (Fig. 3, step 1) is used to compute the conditional distribution of the stimulus feature (Eq. 2a). For this to hold, we make two further assumptions: First, we assume the mean recurrent input, $\langle \mathbf{u}^{\mathsf{r}}_t \rangle_{\mathbf{r}_{t-\Delta t}}$, averaged over realizations (trials) conditioned on the response $\mathbf{r}_{t-\Delta t}$, has the same tuning profile as the feedforward inputs (Eq. 1) but centered at the stimulus feature sample $\tilde{s}_{t-\Delta t}$ represented by $\mathbf{r}_{t-\Delta t}$, i.e., $\langle \mathbf{u}^{\mathsf{r}}_{tj} \rangle \propto \exp[\mathbf{h}_j(\tilde{s}_{t-\Delta t})] = \exp[-(\tilde{s}_{t-\Delta t} - \theta_j)^2/2a^2]$. Second, given the response $\mathbf{r}_{t-\Delta t}$, we assume that the recurrent inputs, $\mathbf{u}^{\mathsf{r}}_t$, are independent, and have Poisson-like variability (the variance equals the mean, see Eq. 21). Under these assumptions, the stimulus feature, s, given recurrent input $\mathbf{u}^{\mathsf{r}}_t$, follows 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 read out linearly from $\mathbf{u}^{\mathsf{r}}_t$ (Fig. 3b and e, bottom-left; Eq. 23). Since the recurrent input encodes a sampled value of context feature, we refer the subspace in the space of recurrent inputs spanned by the vector $(\theta_1, \theta_2, \ldots, \theta_{N_E})$ as the *context feature subspace*. By construction, the recurrent and feedforward inputs represent conditional distributions of the stimulus feature, s, with the same basis function, $\mathbf{h}(s)$ (logarithm of the tuning curve). Therefore summing the feedforward and recurrent 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),

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The firing rate of the neurons in the recurrent circuit, $\boldsymbol{\lambda}_t$, thus represents the instantaneous conditional distribution, $p(s|\tilde{z}_t, \mathbf{u}^{\mathbf{f}})$, and has the same basis function, $\mathbf{h}(s)$, as that of the feedforward and recurrent inputs (Eq. 2a). Table 2 summarizes 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 stimulus 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 $p(s|\tilde{z}_t, \mathbf{u}^{\mathbf{f}})$ encoded in the mean firing rates, $\boldsymbol{\lambda}_t$ (Eqs. 2a and 4), and ξ_t are independent samples from a standard normal distribution. 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[4;22], and is thus a natural candidate for the requisite variability. We show that when the instantaneous firing rate, $\boldsymbol{\lambda}_t$, represents the conditional distribution, $p(s|\tilde{z}_t, \mathbf{u}^{\mathbf{f}})$ 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, $\tilde{s}_t$, can be read out from the population response, $\mathbf{r}_t$, using a linear decoder, $\tilde{s}_t = \sum_j \mathbf{r}_{tj}\theta_j / \sum_j \mathbf{r}_{tj}$. The posterior mean, $\hat{s}_t$, is related to the mean firing rates, $\boldsymbol{\lambda}_t$, in the same way. 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 feature samples, Λ^{-1}, is determined by the sum of the instantaneous firing rates, $\sum_j \boldsymbol{\lambda}_{tj}$.

To confirm that Poisson spike generation leads to sampling from the correct stimulus feature distribution, we fixed a realization of instantaneous firing rates, $\boldsymbol{\lambda}_t$ (Fig. 3c), which corresponds to fixing the conditional distribution of the stimulus feature (Eq. 4), and then generate independent Poisson spikes 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 provide 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 the sampling variability, Λ^{-1}, is determined by firing rates, and can thus change over short times scales.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Recurrent connections represent the relation between stimulus and context features

To generate samples from the correct joint distribution of stimulus and context features, the mean and variance of the recurrent input, $\mathbf{u}^{\mathsf{r}}$, to neurons in the circuit must satisfy Eq. (3). 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. 21 and 24). Second, the mean and variance of the recurrent input conditioned on the population responses, $\mathbf{r}_t$, need to be equal.

We show that the first requirement is satisfied when the weights between neurons in the recurrent circuit are self-connected with homogeneous connection strength (see Eq. 26). The optimal recurrent weight to correctly sample the posterior is

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

where Λ_s is the precision of the stimulus prior and Λ_{f} is the precision of the feedforward input distribution $p(\mathbf{u}^{\mathsf{f}}|s)$. The recurrent weight, w_E^*, thus encodes the relation between the stimulus feature, s, and the context feature, z. An increase in correlation between s and z (reflected by a narrower diagonal band in Fig. 2c) requires an increase in the recurrent weight w_E^* to satisfy Eq. (3). However, when context and stimulus are uncorrelated ($\Lambda_s = 0$), then recurrent interactions are not needed (i.e., w_E^*=0). This is in agreement with the example we chose to motivate our approach, where LFI was maximized in the absence of recurrent connectivity (Fig. 1e).

To fulfill the second requirement, the recurrent interactions in the network need to exhibit Poisson-like variability (Fig. 3a, step 3; Eq. 19), which could be the result of noisy synaptic transmission[41;42]. The resulting recurrent input, $\mathbf{u}_t^{\mathsf{r}}$, introduces the variability needed to generate samples 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 neural circuit can carry out Gibbs sampling, we implemented the proposed algorithm in a recurrent network consisting of E and I Poisson neurons with synaptic interactions (See details in Eqs. 31-34). The feedforward and recurrent circuitry of this network was identical to the one used to probe the effect of recurrent coupling on a linear estimation of stimulus s (Fig. 1). We assumed that the response properties of the individual neurons satisfy the representation requirements we described previously: The tuning of the recurrent inputs, $\mathbf{u}^{\mathsf{r}}$, and neuronal responses, $\mathbf{r}$, followed the same Gaussian profile as the feedforward input, $\mathbf{u}^{\mathsf{f}}$ (Eq. 4). The recurrent inputs in the network exhibited independent Poisson-like variability given a spiking response (Eq. 32). Thus, the model included three sources of variability (Fig. 4a): The external

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

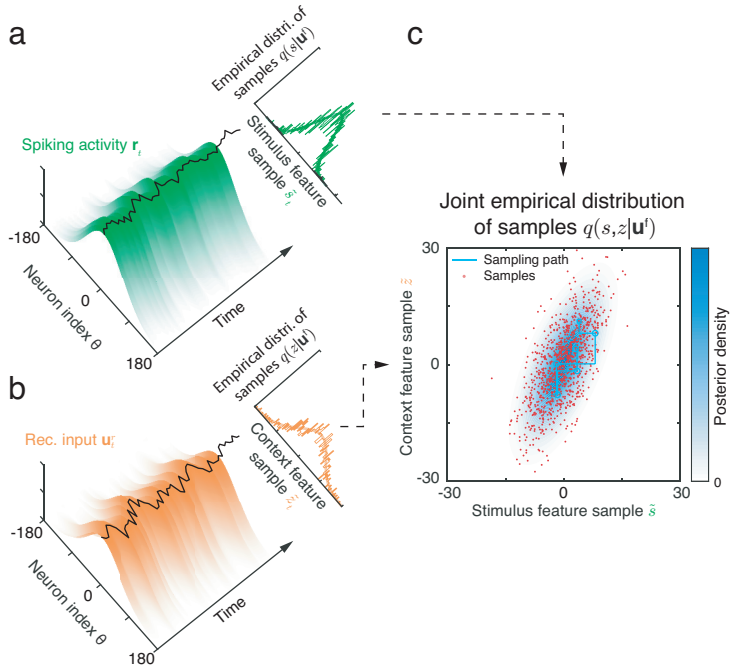

Figure 4: Reading out the posterior over latent features from spiking activity and recurrent inputs in a model cortical circuit. (a-b) Spiking activity, $\mathbf{r}_t$ (a), and recurrent input $\mathbf{u}_t^r$ (b) evolve stochastically. Their projection onto the feature subspace (black curve overlaid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

variability in feedforward inputs $\mathbf{u}^f$ which determines the likelihood function of stimulus feature (Eq. 1), internal variability due to spike generation driving sampling from the stimulus feature distribution, and variability in the recurrent interactions driving the sampling from the context feature distribution. We next demonstrate that with these sources of variability the network is able to infer stimulus and context features.

Population responses 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 $\mathbf{u}^f$ generated according to Eq. (1). Response variability was due to the combination of Poissonian spiking and noisy recurrent interactions. Fixing $\mathbf{u}^f$ 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, $\mathbf{u}^f$ (Eq. 14; Fig. 1b). Since inhibition is unstructured (see Methods) only E neurons are selective for the stimulus and context features in the E-I network (Fig. 1c), implying that we

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

only need to read out the samples of features from E neurons. To simplify notation, we therefore suppress subscripts denoting neuronal type in the main text, and denote responses and recurrent inputs to 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 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 23; 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 we developed earlier (Eq. 6).

Towards this end, we fixed all parameters in the generative model (Eq. 1) and changed only recurrent weights in the model network. For simplicity, we only varied the peak excitatory strength, w_E (Eq. 33), 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, $q(s, z|\mathbf{u}^{\mathsf{f}})$, generated by the network and the true posterior, $p(s, z|\mathbf{u}^{\mathsf{f}})$ (Eq. 14), using the Kullback-Leibler (KL) divergence. The KL divergence quantifies the loss of mutual information, measured in bits, between latent features, i.e., s and z, and the feedforward inputs $\mathbf{u}^{\mathsf{f}}$ due to approximating the true posterior distribution, p, with the sampling distribution, q (Eq. 17)[36]. KL divergence is always non-negative: A zero KL divergence indicates q equals p, while a positive KL divergence reflects a deviation of q

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

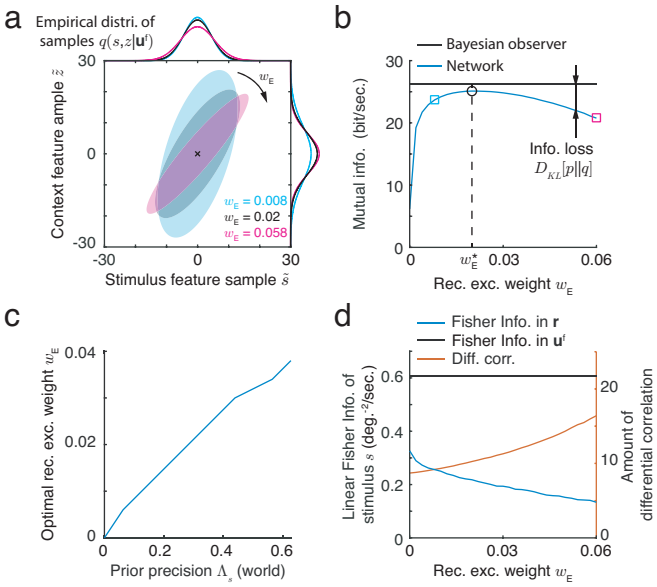

Figure 5: Generating samples from the joint posterior over latent features requires that a network stores an appropriate internal model in its recurrent weights. (a) The joint empirical distribution of samples read out from the population responses for different values of peak recurrent excitatory weight, 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. Different colors correspond to the three values of w_E, denoted by different symbols in panel b. (b) The mutual information between latent features, s and z, and feedforward inputs for ideal Bayesian observer (black) and our network models (blue). The difference between the 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 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 (LFI) of stimulus, s, in the population response, $\mathbf{r}$, decreases with recurrent weight w_E (blue), which is bounded above by the Fisher information in feedforward inputs $\mathbf{u}^\mathbf{f}$ (black). The decrease in LFI is due (in part) to a concomitant increase in differential noise correlations (orange) with w_E.

from p which causes information loss. 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, in 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), resulting in a decrease in discriminability of the external stimulus feature s from the neuronal response, $\mathbf{r}$.

To confirm that the optimal recurrent weight, w_E^*, increases 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 $\Lambda_s = 0$, and the context and stimulus features are uncorrelated, a network

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

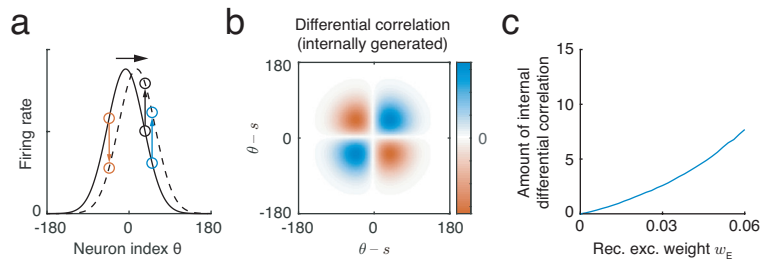

Figure 6: Sampling along the stimulus feature subspace is reflected in internally generated differential correlations. (a) When the network generates samples in the stimulus feature subspace, the population firing rate fluctuates along the feature subspace. Neuron pairs on the same side (the black and 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 is a result of sampling in the stimulus feature subspace. (c) Internally generated differential correlations increase with recurrent excitatory weight, w_E, which increases with the correlation between stimulus and context in the prior, Λ_s, in a network that samples optimally.

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 based inference of the type we describe. Such sampling would cause instantaneous firing rates, $\boldsymbol{\lambda}_t$, to fluctuates along the stimulus feature subspace (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 here describe population response fluctuations along the stimulus feature subspace, since we are mainly interested in how stimulus features are represented in this subspace. For illustration, 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). A similar argument shows how sampling dynamics structures pairwise correlations in stimulus feature subspace across the entire network (Fig. 6b). We can make these arguments more precise: 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., $\mathbf{f}_s'\mathbf{f}_s'^\top$ where $\mathbf{f}_s'$ denotes the derivative of mean population rate $\mathbf{f}(s)$ over stimulus feature, s. Such correlations have been referred to as differential correlations in previous studies[2;21], and they limit the information about s that is available to a

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

decoder as measured by linear Fisher Information (see Fig. 1e).

Moreover, in our model, the amplitude of the internally generated differential correlation as a consequence of sampling based inference is determined by the strength of recurrent connections (Fig. 6c). We computed the covariance matrix of the population response, $\Sigma(\mathbf{r}|\mathbf{u}^{\mathsf{f}})$, (See Eq. 28), and found that the amplitude of internally generated differential correlations is proportional to the recurrent weight w_E^*,

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

The second term on the right hand side in above equation represents internally generated differential correlations with amplitude $a^2 n_{\mathsf{f}}^{-1} w_E^*$, where a is the tuning width, n_{f} is the total number of feedforward input spikes received by all E neurons (Table 1). $\mathrm{diag}[\mathbf{f}(s)]$ is a diagonal matrix capturing the individual Poisson spiking variability, with $\mathbf{f}(s)$ the tuning of E neurons over s. Since the optimal recurrent weight, w_E^*, is determined by the correlation between stimulus and context in the prior (Fig. 5c, Λ_s in Eq. 6), the amplitude of differential correlations in a network implementing optimal sampling is determined by the correlations between stimulus and context features, Λ_s (Fig. 5d and Eq. 29). We confirmed this conclusion in network simulations which showed that differential correlations increased with excitatory recurrent weight, w_E (Fig. 6c), and hence with Λ_s (Fig. 5c).

To understand why internally generated differential correlations are shaped by the structure 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, $\boldsymbol{\lambda}_t$, will be fixed in response to a fixed feedforward input. We find that in this case the variability 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 increases recurrent variability (Fig. 1d, red), and hence increases internally generated differential correlations, thus reducing linear Fisher Information.

Distinguishing external and internal differential correlations

In many neurophysiology experiments a stimulus feature, s, is fixed, while the feedforward input, $\mathbf{u}^{\mathsf{f}}$, fluctuates (Fig. 4a). Hence in experimentally recorded neuronal responses, fluctuations in the stimulus feature subspace reflect a combination of sensory transmission noise, and any internally generated variability. Differential noise 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 internally generated from recurrent network interactions. In this case the total magnitude of differential correlations in the evoked response is $a^2 n_{\mathsf{f}}^{-1} w_E \mathbf{f}_s' \mathbf{f}_s'^{\top} + a^2 n_{\mathsf{f}}^{-1} \mathbf{f}_s' \mathbf{f}_s'^{\top}$

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

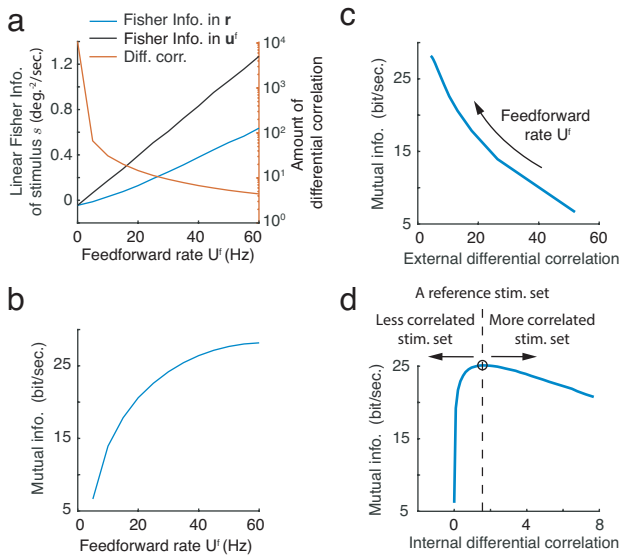

Figure 7: (a-b) Externally generated differential correlations, inherited from feedforward inputs, impair information representation. (a) The feedforward rate determines the input gain, and the reliability of sensory transmission and the amount of external differential correlation inherited from feedforward input. (b) Mutual information (the information between the feedforward inputs and the latent features sampled by the network with optimized recurrent weight w_E^*) monotonically increases with feedforward rate. (c-d) External and internal differential correlations have a distinct impact on information representation in the recurrent circuit model. While we find a monotonic relation between mutual information in a recurrent network and external differential correlations (c), there is a non-monotonic relation between mutual information and internal differential correlations (d). The direction of the arrow denotes the change of the parameters that lead to the change in differential correlations.

(Eq. 30), where the new term, $a^2 n_{\mathbf{f}}^{-1} \mathbf{f}_s' \mathbf{f}_s'^{\top}$, reflects differential correlations inherited from the input (compare with Eq. 7). Although the two sources of differential correlations are intertwined, we found that they impact information processing differently. This difference offers a potential way to distinguish between them, and determine whether the network produces samples in the joint stimulus and context feature subspace.

Externally imposed differential correlations decrease with the strength (i.e. contrast) of the feedforward input (the peak firing rate of feedforward input, Eq. 10) (Fig. 7a, orange curve). Consequently, the LFI of s in the observations of either the feedforward input, $\mathbf{u}^{\mathbf{f}}$, or the population activity, r, increase with the strength of the feedforward input (Fig. 7a, black and blue curves, respectively). Similarly, the mutual information (the information between feedforward inputs $\mathbf{u}^{\mathbf{f}}$ and the latent features, i.e., s and z, sampled by the network, Eq. 17 and Fig. 5b) also increases with feedforward input intensity (Fig. 7b). We therefore have a monotonic, decreasing relationship between externally generated differential correlations and mutual information (Fig. 7c). This is expected since such inherited correlations always impair information processing, as observed previously[2;21]. The magnitude of internally generated differential correlations is proportional to the recurrent excitatory weight, w_E (Eq. 7). The non-monotonic relation between w_E and the mutual

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

information between $\mathbf{u}^{\mathbf{f}}$ and s and z (Fig. 5b) creates a similar non-monotonic relation between internally generated differential correlations and mutual information (Fig. 7d). Therefore, the impact of external and internal differential correlations on stimulus coding can be distinguished by their respective monotonic and non-monotonic relation with the mutual information in the network (see Discussion).

Discussion

In this study we propose a new role for recurrent connections in neural circuits that represent complex sensory scenes. We demonstrated their function in a model cortical hypercolumn that implements a sampling-based approximation of the joint posterior over a local stimulus and a global context feature. There is strong evidence that Bayesian inference underlies perceptual processing[12;13;15;32], such as multisensory cue integration[11] and sensorimotor learning[14]. In our circuit model, the recurrent connections store an associative prior between the stimulus and context features, and thus reflect the statistical structure of the world learned (or assumed) by the animal. Internally generated Poisson variability, reflecting the known trial-to-trial and dynamic variability in neural dynamics, provides the stochasticity needed for implementing the sampling algorithm. Finally, the stimulus and context features are embedded in a subspace in the neural population responses from which the posterior of features can be linearly read out.

A neural signature of network sampling in stimulus feature subspace is the covariability of population responses along the stimulus feature subspace, which is a reflection of internally generated differential noise correlations. Such correlations are due to recurrent interactions in the network, and not inherited from sensory transmission noise in feedforward inputs[21]. 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 produce samples from a posterior distribution over the stimulus features, as we discuss below. While externally and internally generated differential correlations are necessarily intertwined in any recorded neuronal responses, we found that they impact information representation in the circuit differently. The mutual information between feedforward inputs and latent features sampled by the network is maximized at a unique amplitude of internal differential correlations (Fig. 5 and 6), while external differential correlations always decreased mutual information and hence impaired information processing (Fig. 7c).

The proposed circuit sampling mechanism thus produces experimentally testable predictions. First, our model requires that the feedforward inputs and the recurrent inputs have the same tuning profile (Eq. 1 and 3), in order to compute the conditional distribution of latent stimulus feature via summing feedforward and recurrent inputs (Eq. 4). The assumption that feedforward and recurrent

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

inputs have the same tuning profile is supported by a recent circuitry experiments in layers 4 and 2/3 in mouse V1 [9]. Second, our model predicts that the magnitude of internal differential correlations is determined by the learned correlation between stimulus and context (Fig. 6b-c). This could be tested with a perceptual learning task in which a subject is repeatedly presented with the same set of stimuli[43]. We predict that increasing the correlation between global context and the local stimulus in the stimulus set, would lead to an increase in differential correlations during and after learning (See Fig. 5c-d). Indirect evidence from previous studies supports the possibility that neuronal correlations and spatiotemporal firing patterns are shaped by external stimuli: It has been shown that neurons become less correlated with perceptual learning with a simple stimulus set[44], and spontaneous population response patterns, including correlation structure, become more similar to evoked response patterns during development[45]. Lastly, to verify the predicted impact of internal differential correlations on information representation, we could test whether there exists a non-monotonic relation between the magnitude of internally generated differential correlations and the mutual information (the information between feedforward inputs and latent features sampled by the network) estimated from population responses (Fig. 7d). For example, suppose an animal is fully adapted to a *reference* stimulus set with a prescribed correlation between the context and local stimulus. Training the animal with a more or less correlated stimulus set (compared to the reference set) would then increase or decrease the amplitude of internal differential correlations respectively. Regardless of whether the differential correlations increase or decrease, our network model predicts that mutual information (with the true generative model, $p(\Theta)$ in Eq. 17, defined as the reference stimulus set) in the population responses will decrease (Fig. 7d).

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)[16;34], and features of sampling-based codes (SBCs) [17;19;37;38;40]. In our network, the conditional distribution of the stimulus feature is represented by instantaneous firing rates using a parametric code similar to a PPC (Eqs. 3 and 4), i.e., the conditional distribution belongs to a linear exponential family of distributions with sufficient statistics that are linear projections of instantaneous firing rates. Further, our network produces samples in the stimulus feature subspace with variability generated internally, similar to previously proposed SBCs (Fig. 3a).

Despite these similarities, there are some fundamental differences between our network and PPCs and conventional SBCs. PPCs are generally implemented in deterministic networks (no internal variability) 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 that drives sampling in the joint stimulus and context feature subspace, as well as variability in feedforward inputs to the network which conveys the likelihood function[34] (Fig. 3a). Second, in PPCs the posterior of stimulus feature is represented

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

parametrically in neuronal responses, while in our 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 latent stimulus feature[17;19;20;37;38;40], 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). Hence, in our proposed network the sampling space has dimension two, while in conventional SBCs the dimension of sampling space is the number of neurons in the network. A benefit of sampling in the feature subspace is that sampling is much more efficient in spaces of lower dimension.

On the network level, the recurrent connections are translation-invariant in the stimulus feature subspace, an assumption widely used in continuous attractor network models (CAN)[26;46–48]. 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[16]. Here we have shown that a network with CAN-like structure and internally spiking 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[4], real cortical circuits may use it to sample stimulus features.

To keep the exposition transparent we have presented a model of minimal complexity. The proposed network mechanism of sampling-based inference can be generalized to more complex generative models, and hence the assumption of Gaussianity (Eqs. 8 and 9) and analytical expression in Eq. (14) are not essential to the proposed mechanism. For example, a similar network can generate samples from other types of distributions belonging to the linear exponential family[34;36]. The type of distribution from which the network samples is determined by the basis function (logarithm of tuning) $\mathbf{h}(s)$ (Eq. 4). These basis functions depend on the tuning of feedforward inputs, along with the recurrent connection profiles, and do not have to be Gaussian. The network model could also infer the latent features with a non-uniform prior, when the preferred stimuli of neurons in the population are not distributed uniformly in the stimulus feature subspace. In addition, 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 Poissonian spiking variability, with each subnetwork inferring the marginal posterior of one 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[49].

In conclusion, we have shown that recurrent connections in a cortical circuit can store information about the learned structure of the world, and that neural noise can drive sampling from the

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 representation of the structure of the world stored in recurrent circuits.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

The probabilistic generative model

We assume that the statistics of stimuli in the world, and the feedforward inputs received by neurons in the recurrent network, 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 can be generalized to explain information processing in other brain areas. We denote by z the one dimensional context feature (e.g., orientation) of an extended 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 bimodal, we assume that z follows a uniform distribution for simplicity (Fig. 2c, marginal plots),

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

where $\mathcal{U}(a, b)$ denotes a uniform distribution over $[a, b]$. The assumption that the prior distribution over the context variable is uniform simplifies our model significantly, as it implies the spatial homogeneity of the network model as given by Eqs. (10-11). However, it doesn't substantially change our main results. We denote by s the one dimensional feature (e.g., orientation) of the local stimulus falling into the central receptive field (cRF) of a single V1 hypercolumn. Due to the differences between the local (stimulus) and global (context) aspects of the sensory scene, the stimulus feature s is not identical to the contextual feature z, but we assume that the two are related as,

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

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 model the feedforward inputs $\mathbf{u}^{\mathsf{f}}$ to the N_E excitatory neurons in V1 as conditionally independent Poissonan variables, with Gaussian intensity tuning over the input,

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

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 the peak firing rate and width of the tuning function, respectively. This model of the input has been widely used in previous neural coding studies, such as those of probabilistic

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

population codes (PPC)[16;34;35].

Under the assumption that the preferred stimulus feature values, $\{\theta_j\}_{j=1}^{N_E}$, of all feedforward inputs are uniformly distributed in the space of the stimulus feature s, we can use Eq. (10) to show that the likelihood of s given a single observation of the input, $\mathbf{u}^{\mathsf{f}}$, belongs to the exponential family of distributions with linear sufficient statistics[16;34;35], and has the form

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

The second line Eq. (11) denotes the standard form for a distribution belonging to the linear exponential family, where $\phi(\mathbf{u}^{\mathsf{f}})$ is a function which only depends on the input, $\mathbf{u}^{\mathsf{f}}$. The *basis function*, $\mathbf{h}(s)$[16], determines the profile of the 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 $\mathbf{u}^{\mathsf{f}}$ can be normalized to obtain a Gaussian distribution (Eq. 11), with mean μ_{f} and precision Λ_{f} which are both linear functions of the inputs, $\{\mathbf{u}_j^{\mathsf{f}}\}_{j=1}^{N_E}$,

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

The mean, μ_{f}, represents the position of $\mathbf{u}^{\mathsf{f}}$ in stimulus feature subspace, and the precision, Λ_{f}, is proportional to the sum of total feedforward spike counts, n_{f} (Fig. 2d-e). In sum, the whole generative model is determined by the joint distribution,

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

Approximate Bayesian inference via Gibbs sampling

We assume that the neural circuit needs to compute information about both the local feature, s, and context feature, z, from the feedforward input $\mathbf{u}^{\mathsf{f}}$. The circuit approximates the joint posterior distribution over the features, rather than making a point estimate[12;32]. The simplicity of the generative model (Eq. 13), allows us to obtain an analytical expression for the posterior

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

distribution,

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

Our goal is to develop a biologically plausible sampling algorithm to approximate this posterior distribution, and use Eq. (14) to verify the output of the algorithm.

We consider using Gibbs sampling[36] (a type of Monte Carlo method) to approximate the posterior. To describe the iterative Gibbs algorithm, we assume that a sample, $\tilde{z}_t$, of the context feature, z, is provided at time t. We can use this sample, $\tilde{z}_t$, with the feedforward input, $\mathbf{u}^{\mathsf{f}}$, to update the conditional distribution of stimulus feature s (step 1 in Fig. 2f),

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

Here $\hat{s}_t$ and Λ denote the mean and precision (inverse of variance) of the conditional distribution $p(s|\tilde{z}_t, \mathbf{u}^{\mathsf{f}})$. The next step in the algorithm is to draw a sample, $\tilde{s}_t$, from the conditional distribution $p(s|\tilde{z}_t, \mathbf{u}^{\mathsf{f}})$ (step 2 in Fig. 2f),

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

Next, the conditional distribution over the context feature, z, is updated given this new sample, $\tilde{s}_t$, and a new sample, $\tilde{z}_{t+\Delta t}$, is drawn (step 3 in Fig. 2f),

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

These three steps in the Gibbs sampling algorithm (Eqs. 15-16) are performed iteratively until the distribution of samples, $\tilde{s}_t$ and $\tilde{z}_t$, approximate the true posterior distribution with sufficient accuracy (Fig. 2g; compare the red dots with the blue contour map).

Inference from an information-theoretic point of view

The goal of the sampling algorithm is to approximate the posterior distribution over the features, $\Theta = \{s, z\}$, given a feedforward input, $\mathbf{u}^{\mathsf{f}}$. When the sampling algorithm uses a model which does not match the structure of the generative model for the stimuli, the samples will converge to distribution $q(\Theta|\mathbf{u}^{\mathsf{f}})$ which differs from the true posterior, $p(\Theta|\mathbf{u}^{\mathsf{f}})$ (Eq. 14). In this case the mutual information between the asymptotic samples of latent features, Θ, and feedforward input will be

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

smaller than in the case when samples come from the true posterior, $p(\Theta|\mathbf{u}^{\mathsf{f}})$,

$$\begin{aligned} I(\Theta, \mathbf{u}^{\mathsf{f}}) &= -\mathbb{E}_{p(\Theta)}[\log p(\Theta)] + \mathbb{E}_{p(\Theta,\mathbf{u}^{\mathsf{f}})}[\log p(\Theta|\mathbf{u}^{\mathsf{f}})] \\ &\geq -\mathbb{E}_{p(\Theta)}[\log p(\Theta)] + \mathbb{E}_{p(\Theta,\mathbf{u}^{\mathsf{f}})}[\log q(\Theta|\mathbf{u}^{\mathsf{f}})] \equiv I_q(\Theta; \mathbf{u}^{\mathsf{f}}), \end{aligned} \tag{17}$$

It is straightforward to show that the difference between $I(\Theta, \mathbf{u}^{\mathsf{f}})$ and $I_q(\Theta, \mathbf{u}^{\mathsf{f}})$ is the Kullback-Leibler (KL) divergence between p and q, i.e., $D_{KL}[p||q] = I(\Theta, \mathbf{u}^{\mathsf{f}}) - I_q(\Theta, \mathbf{u}^{\mathsf{f}}) = \mathbb{E}_p(\ln p - \ln q) \geq 0$. Equality in Eq. (17) holds only if the distribution q matches the true posterior p exactly.

The mutual information $I_q(\Theta; \mathbf{u}^{\mathsf{f}})$ can be computed analytically when the approximating distribution $q(\Theta|x) = \mathcal{N}\left(\Theta|\boldsymbol{\mu}_q, \mathbf{K}_q^{-1}\right)$ is a bivariate normal (substituting Eqs. 13 and 14 into Eq. 17),

$$I_q(\Theta; \mathbf{u}^{\mathsf{f}}) = \log L + \frac{1}{2}[1 - \log(2\pi)] + \frac{1}{2}\log\frac{|\mathbf{K}_q|}{\Lambda_s} - \frac{1}{2}\left[\mathrm{tr}(\mathbf{K}_p^{-1}\mathbf{K}_q) + (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^\top \mathbf{K}_q(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)\right], \tag{18}$$

where $L = 360°$ 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 posterior distribution (Eq. 14). When q matches the posterior distribution, p, we have, $I(\Theta; \mathbf{u}^{\mathsf{f}}) = \log L - \frac{1}{2}[1 + \log(2\pi)] + \frac{1}{2}\log\Lambda_{\mathsf{f}}$. As expected mutual information increases with the reliability of sensory transmission, Λ_{f} (Fig. 7b).

A stochastic recurrent network sampling latent features

We consider a generic recurrent network consisting of N_E excitatory Poisson neurons and implement sampling from the posterior over the features using network dynamics. For simplicity, we describe neuronal activity using a Hawkes process (a type of inhomogeneous Poisson process[50]). The instantaneous firing rates of the neurons in the network at time t, $\boldsymbol{\lambda}_t$, obey the following

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

recurrent equations:

$$\boldsymbol{\lambda}_t = \mathbf{u}^{\mathsf{f}} + \mathbf{u}^{\mathsf{r}}_t = \mathbf{u}^{\mathsf{f}} + (w_E \mathbf{r}_{t-\Delta t} + \sigma_r \boldsymbol{\xi}_t)\,, \tag{19}$$

$$\mathbf{r}_t \sim \prod_{j=1}^{N_E} \text{Poisson}(\boldsymbol{\lambda}_{tj}), \tag{20}$$

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 denote the excitatory recurrent weight and recurrent variability, respectively. Below we present the conditions for the network to implement Gibbs sampling.

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 activity, $\mathbf{r}$ represents the stimulus feature, s. Specifically, we assume that with fixed 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 $\tilde{s}_{t-\Delta t}$,

$$\begin{aligned} p(\mathbf{u}^{\mathsf{r}}_t | \tilde{s}_{t-\Delta t}) &= \textstyle\prod_{j=1}^{N_E} \mathcal{N}[\mathbf{u}^{\mathsf{r}}_{tj} | \langle \mathbf{u}^{\mathsf{r}}_{tj} | \tilde{s}_{t-\Delta t} \rangle, \langle \mathbf{u}^{\mathsf{r}}_{tj} | \tilde{s}_{t-\Delta t} \rangle], \\ \langle \mathbf{u}^{\mathsf{r}}_{tj} | \tilde{s}_{t-\Delta t} \rangle &= U^{\mathsf{r}} \exp[\mathbf{h}_j(\tilde{s}_{t-\Delta t})] = U^{\mathsf{r}} \exp[-(\theta_j - \tilde{s}_{t-\Delta t})^2 / 2a^2]. \end{aligned} \tag{21}$$

In Eq. (21) we also assume that the recurrent input, $\mathbf{u}^{\mathsf{r}}$, 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. The assumption of independence of the components of $\mathbf{u}^{\mathsf{r}}$ given a stimulus feature sample, $\tilde{s}_{t-\Delta t}$, does not mean that the components of $\mathbf{u}^{\mathsf{r}}$ conditioned on an external stimulus feature s are independent. Due to the nature of the 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 we consider a network that transforms its inputs linearly (Eq. 19), if we only consider the linear Fisher information of s contained in $\mathbf{u}^{\mathsf{r}}$, then the likelihood of s given a recurrent input $\mathbf{u}^{\mathsf{r}}$ can be normalized to obtain a Gaussian distribution as follows,

$$\begin{aligned} p(s|\mathbf{u}^{\mathsf{r}}) &\propto \exp\left[\mathbf{h}(s)^\top \mathbf{u}^{\mathsf{r}}\right] = \exp\left[-\textstyle\sum_j \mathbf{u}^{\mathsf{r}}_j \frac{(\theta_j - s)^2}{2a^2}\right], \\ &\propto \mathcal{N}(s|z, \Lambda_s^{-1}). \end{aligned} \tag{22}$$

Comparing Eqs. (22) and (9) we see that the recurrent input $\mathbf{u}^{\mathsf{r}}$ can represent a conditional distribution over the stimulus feature s, and that the mean and precision of the prior can be linearly

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

read out from $\mathbf{u}^{\mathsf{r}}$,

$$\Lambda_s = a^{-2} n_r = a^{-2} \textstyle\sum_j \mathbf{u}^{\mathsf{r}}_j, \quad z = n_r^{-1} \textstyle\sum_j \mathbf{u}^{\mathsf{r}}_j \theta_j. \tag{23}$$

Because the recurrent and feedforward inputs represent the distributions in the same way by design (see Eqs. 11 and 22), the sum of feedforward and recurrent inputs gives the instantaneous firing rate $\boldsymbol{\lambda}_t$ (Eq. 19),

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

Therefore 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 determined by linear functions of the rates, $\Lambda = a^{-2} \sum_j \boldsymbol{\lambda}_{tj}$ 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. The average (tuning) of $\boldsymbol{\lambda}$ can be computed as,

$$\begin{aligned} \langle \boldsymbol{\lambda}_j(s) \rangle &= \langle \mathbf{u}^{\mathsf{f}}_j(s) \rangle + \langle \mathbf{u}^{\mathsf{r}}_j(s) \rangle = (U^{\mathsf{f}} + U^{\mathsf{r}}) \exp[\mathbf{h}_j(s)] \\ &= R \exp[-(s-\theta_j)^2/2a^2]. \end{aligned} \tag{24}$$

Poisson spike generation samples stimulus feature subspace

We demonstrate next that the internally generated Poisson spike variability generates random samples from the conditional distribution $p(s|\tilde{z}_t, \mathbf{u}^{\mathsf{f}})$ of the stimulus feature, s, 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 show that the empirical distribution of the 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 $\boldsymbol{\lambda}_t$. To simplify the derivation, suppose the time window or neural population are sufficiently large so that the instantaneous firing rate, $\boldsymbol{\lambda}_t$, is close to its trial average, i.e., a smooth Gaussian profile (Eq. 24), 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. 20), the likelihood of an emitted spike $\mathbf{r}_t$ given $\hat{s}_t$ is proportional to a Gaussian distribution,

$$p(\mathbf{r}_t|\hat{s}_t) = \textstyle\prod_{j=1}^{N_E} \text{Poisson}\left[\mathbf{r}_{tj}|\boldsymbol{\lambda}_{tj}(\hat{s}_t)\right] \propto \mathcal{N}\left(\tilde{s}_t|\hat{s}_t, \Lambda_{\mathbf{r}}^{-1}\right), \tag{25}$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with precision $\Lambda_r = a^{-2} n_r = a^{-2} \sum_j \mathbf{r}_{tj}$ and $\tilde{s}_t = n_r^{-1} \sum_j \mathbf{r}_{tj} \theta_j$. Comparing the empirical distribution of stimulus feature sample $\tilde{s}_t$ collected over trials (Eq. 25), 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 (Λ_r versus Λ). However, the precision of the empirical distribution, Λ_r, equals the precision of conditional distribution, Λ, in the limit of a large neural population, or a long time window (see Supplementary Information for a rigorous derivation),

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

This shows the Poisson spike generation leads the neural dynamics to asymptotically sample stimulus features from the true conditional in the limit of a large neural population, or over long time windows. We confirm this conclusion in simulations (Fig. 3h).

Recurrent connections store the association between stimulus and context

The representation of context feature samples in recurrent input, $\mathbf{u}^{\mathsf{r}}$, (Eqs. 21 and 22) puts constraints on the recurrent weight, w_E, and recurrent variability, σ_r. Because the stimulus tuning of the recurrent input, $\mathbf{u}^{\mathsf{r}}$, and responses, $\mathbf{r}$, have the same width a (Eqs. 21 and 24), every neuron in the population can only be connected to itself in the simple network consisting of only excitatory neurons, since otherwise the tuning of $\mathbf{u}^{\mathsf{r}}$ would be wider than that of $\mathbf{r}$. Combining Eqs. (21 and 24) and using Table 2, we can show that every neuron is self-connected with the same weight w_E, with the optimal weight satisfying the requirement of recurrent inputs (Eq. 21),

$$w_E^* = \frac{\langle \mathbf{u}_j^{\mathsf{r}} \rangle}{\langle \mathbf{r}_j \rangle} = \frac{U^{\mathsf{r}} e^{-(\theta_j - s)^2/2a^2}}{R e^{-(\theta_j - s)^2/2a^2}} = \frac{\Lambda_s}{\Lambda_{\mathsf{f}} + \Lambda_s}, \tag{26}$$

which yields Eq. (6) in the main text. The last equality in Eq. (26) follows from the observation that the sum of the vector components $\mathbf{r}$ 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 self-connections are a result of the simplifying assumption that the network consists solely of excitatory neurons (Eq. 19). This assumption simplified the analysis, and is relaxed below (see Eq. 33 and Fig. 1).

Another assumption on the recurrent connections is that the components of $\mathbf{u}^{\mathsf{r}}$ exhibit independent, Poisson-like variability (the variance is the same as mean), given the network's activity at the previous time step, $\mathbf{r}_{t-\Delta t}$. This constrains the recurrent variability σ_r (Eq. 19). Hence,

$$\mathbf{u}_t^{\mathsf{r}} = \hat{\mathbf{u}}_{rt} + \sqrt{\hat{\mathbf{u}}_{rt}} \boldsymbol{\xi}_t, \quad \hat{\mathbf{u}}_{rt} = w_E^* \mathbf{r}_{t-\Delta t}, \tag{27}$$

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. The Poisson-like variability in

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

recurrent interactions allow for the sampling of the context feature (Eq. 2c), via the same mechanism that allowed for sampling of the stimulus feature using Poisson spike generation (Eq. 5).

Analysis of the conditional distribution of the neuronal response

We compute the distribution of neuronal responses, $\mathbf{r}$, given an external stimulus feature, s, which can be compared with experimental observations[22]. 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 internally generated variability described by $p(\mathbf{r}|\mathbf{u}^{\mathsf{f}})$ (Fig. 3a). Together, the distribution of $\mathbf{r}$ in response to an external stimulus s therefore has the form

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

For simplicity, we only compute the covariability of $p(\mathbf{r}|\mathbf{u}^{\mathsf{f}})$ along the stimulus feature space, since we are mainly interested in how stimuli are represented in this subspace. By approximating the Poissonian spiking variability $p(\mathbf{r}|\boldsymbol{\lambda})$ by a multivariate normal distribution (Eq. 20), and considering the limit of weak fluctuations in $\boldsymbol{\lambda}$ along the stimulus feature subspace, $p(\mathbf{r}|\mathbf{u}^{\mathsf{f}})$ can be computed approximately as (see math details in Supplementary Information),

$$\begin{aligned} p(\mathbf{r}|\mathbf{u}^{\mathsf{f}}) &= \int p(\mathbf{r}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}|\mathbf{u}^{\mathsf{f}})d\boldsymbol{\lambda}, \\ &\approx \mathcal{N}\big[\mathbf{r}\big|\boldsymbol{\lambda}(s), \mathrm{diag}(\boldsymbol{\lambda}(s)) + V(\hat{s}|\mu_{\mathsf{f}})\boldsymbol{\lambda}'_s\boldsymbol{\lambda}'^{\top}_s\big], \quad \text{where } s = \mu_{\mathsf{f}}. \end{aligned} \tag{28}$$

The covariance structure of the neuronal response includes two terms: $\mathrm{diag}(\boldsymbol{\lambda}(s))$, a diagonal matrix whose entries equal that of the vector $\boldsymbol{\lambda}$ denoting the (independent) Poisson spiking variability (Eq. 13), and $V(\hat{s}|\mu_{\mathsf{f}})\boldsymbol{\lambda}'_s\boldsymbol{\lambda}'^{\top}_s$, a term that 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 $\boldsymbol{\lambda}'_s\boldsymbol{\lambda}'^{\top}_s$ is often termed differential correlations [2;21]. With the Gaussian tuning of $\boldsymbol{\lambda}_s$ (Eq. 24), $\boldsymbol{\lambda}'_s\boldsymbol{\lambda}'^{\top}_s$ exhibits anti-symmetric structure (Fig. 6b), as we noted previously[21;26;51–53].

In Eq. (28), $V(\hat{s}|\mu_{\mathsf{f}})$ characterizes the amplitude of internal differential correlations, and $\hat{s}$ and μ_{f} denote the position of $\boldsymbol{\lambda}$ and $\mathbf{u}^{\mathsf{f}}$ on the stimulus feature subspace respectively (Table 2). The dynamics of Gibbs sampling (Eq. S5 in Supplementary Information) and the network structure (Eq. 6) imply that

$$V(\hat{s}|\mu_{\mathsf{f}}) = \frac{\Lambda_s}{\Lambda_{\mathsf{f}}(\Lambda_{\mathsf{f}} + \Lambda_s)} = a^2 n_{\mathsf{f}}^{-1} w_E^*. \tag{29}$$

Note that $V(\hat{s}|\mu_{\mathsf{f}})$ is constrained by network connections, in that it is internally generated by the network. Consequently, any internally generated differential correlations can be regarded as a neural signature of the network performing sampling in stimulus feature subspace.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

An expression for $p(\mathbf{r}|s)$ can be derived similarly, and includes an additional term contributing to differential correlations compared with $p(\mathbf{r}|\mathbf{u}^{\mathsf{f}})$ (Eq. 28) due to fluctuation in the feedforward inputs,

$$\begin{aligned} p(\mathbf{r}|s) &\approx \mathcal{N}\big[\mathbf{r}\big|\boldsymbol{\lambda}(s), \mathrm{diag}(\boldsymbol{\lambda}(s)) + V(\hat{s}|s)\boldsymbol{\lambda}_s'\boldsymbol{\lambda}_s'^{\top}\big], \\ V(\hat{s}|s) &= V(\hat{s}|\mu_{\mathsf{f}}) + V(\mu_{\mathsf{f}}|s) = \frac{\Lambda_s}{\Lambda_{\mathsf{f}}(\Lambda_{\mathsf{f}}+\Lambda_s)} + \frac{1}{\Lambda_{\mathsf{f}}} = a^2 n_{\mathsf{f}}^{-1}(w_E^* + 1). \end{aligned} \tag{30}$$

Here the variance $V(\hat{s}|s)$ in the stimulus feature subspace is a mixture of internal variability, $V(\hat{s}|\mu_{\mathsf{f}})$, and sensory noise, $V(\mu_{\mathsf{f}}|s)$ (Eq. 13).

A spiking network model with excitatory and inhibitory Poisson neurons

To test the validity of the proposed mechanism we developed a recurrent network model consistent with many previous mechanistic models of cortical dynamics[25;26]. The network consists of N_E excitatory (E) and N_I inhibitory (I) spiking neurons, with the activity of each neuron modeled as a Hawkes process[50]. At time t, we represent the response of neuron j in population $a = \{E, I\}$, $\mathbf{r}_j^a(t)$, as a spike count drawn from a Poisson distribution with instantaneous firing rate, $\boldsymbol{\lambda}_j^a(t)$,

$$\mathbf{r}_j^a(t) \sim \mathrm{Poisson}[\boldsymbol{\lambda}_j^a(t)]. \tag{31}$$

The firing rate $\boldsymbol{\lambda}_j^a(t)$ is the sum of feedforward input $\mathbf{u}_j^{a\mathsf{f}}(t)$ and recurrent input $\mathbf{u}_j^{a\mathsf{r}}(t)$, so that $\boldsymbol{\lambda}_j^a(t) = \mathbf{u}_j^{a\mathsf{f}}(t) + \mathbf{u}_j^{a\mathsf{r}}(t)$. The feedforward inputs are filtered spikes from upstream neurons, $\mathbf{u}_j^{a\mathsf{f}}(t) = \sum_n \eta\big(t - t_{jn}^{\mathsf{f}}\big)$, 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 $\eta(t) = \exp(-t/\tau_d)/\tau_d$, $(t > 0)$. Throughout, we set the synaptic time constant $\tau_d = 2$ms. The recurrent input received by neuron j in population a is defined by

$$\begin{aligned} \mathbf{u}_j^{a\mathsf{r}}(t) &= \bar{\mathbf{u}}_j^{a\mathsf{r}}(t) + \sqrt{[\bar{\mathbf{u}}_j^{a\mathsf{r}}(t)]_+}\,\xi(t), \\ \bar{\mathbf{u}}_j^{a\mathsf{r}}(t) &= \sum_{b=\{E,I\}} \sum_{k=1}^{N_b} \frac{J_{jk}^{ab}}{\sqrt{N}} \sum_n \eta(t - t_{kn}^b), \end{aligned} \tag{32}$$

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. The recurrent input in the network is corrupted by noise whose variance equals the mean of the recurrent input. In a physiological network, recurrent noise may be generated by the chaotic state in network dynamics[54] or synaptic noise[41;42]. In Eq. (32) the function $[\cdot]_+$ rectifies the negative input, and $\xi(t)$ is a random varialbe following a standard Gaussian distribution. The coefficient J_{ij}^{ab} is the synaptic weight from neuron j in population b to neuron i in population a. The time t_{kn}^b is the time of the nth spike fired by neuron k in population b. The parameter $N = N_E + N_I$

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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[54]. Finally, the synaptic input profile of the recurrent input, $\eta(t)$, is the same as the one we chose for the feedforward input for convenience.

To model the coding of a circular stimulus feature such as orientation, the excitatory neurons are arranged on a ring[26;51]. The preferred stimuli, θ_j, of the excitatory neurons are equally spaced on the interval $[-180°, 180°]$, consistent with the range of latent features (Eq. 8). Inhibitory neurons are not tuned to stimulus orientation, and their role is to stabilize network responses. The recurrent connections between E neurons are modeled using a Gaussian function decaying with the distance between the stimuli preferred by the two cells,

$$J_{jk}^{EE} = \frac{w_{EE}L}{\sqrt{2\pi}a} \exp\left[-\frac{(\theta_j - \theta_k)^2}{2a^2}\right], \tag{33}$$

We imposed periodic boundaries on the Gaussian function to avoid boundary effect in simulations. Although in the generative model we assumed non-periodic feature variables (Eq. 1), as long as the variance of the associated distributions are smaller than the width of the feature space, the network model with periodic boundaries on the recurrent connection (Eq. 33) provides a good approximation of the non-periodic Gaussian posterior (Eq. 14). The weight w_{EE} denotes the average connection strength of all E to E connections. The parameter $a = 40°$ defines the footprint of connectivity in feature space (i.e the ring), and $L = 360°$ is the length of the ring manifold (Eq. 8); Multiplication by L in Eq. (33) sets the sum of all E to E connection strengths equal to $N_E w_{EE}$. Moreover, the excitatory and inhibitory neurons are all-to-all connected with each other (similar for I to I connections). For simplicity, we consider the E to I, I to I and I to E connections all to be unstructured (in feature space) and assume that connections of the same type have equal weight, i.e., $J_{jk}^{EI} = w_{EI}$, $J_{jk}^{IE} = w_{EE}$ and $J_{jk}^{II} = w_{II}$. To simplify the network further, we consider the connections from the same population of neurons to have the same average weight, i.e., $w_{EE} = w_{IE} \equiv w_E$ and $w_{II} = w_{EI} \equiv w_I$.

The feedforward inputs applied to E neurons consist of independent Poisson spike counts as described by Eq. (10), with rate $\langle \mathbf{u}_j^{E\mathsf{f}}(s)\rangle = U^{\mathsf{f}} e^{-(s-\theta_j)^2/(4a^2)}$. The inhibitory neurons also receive feedforward indpendent Poissonian inputs. The firing rate of the input received by every I neuorn is proportional to the overall feedforward rate of input to E neurons, in order to keep the excitatory and inhibitory balance of neuronal activities in the network,

$$\langle \mathbf{u}_j^{I\mathsf{f}} \rangle = \frac{w_{I\mathsf{f}}}{N_I} \sum_{j=1}^{N_E} \langle \mathbf{u}_j^{E\mathsf{f}}(s)\rangle. \tag{34}$$

In the simulations shown in Figs. 4-7, we started with a network of $N_E = 180$ excitatory and $N_I = 45$ inhibitory neurons, and increased the number of neurons by a fixed factor in Fig. 1e.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The ratio between the average connection from I neurons and the one from E neurons was kept fixed with $w_I/w_E = 5$. We set the feedforward weight of input to I neurons to $w_{I\mathrm{f}} = 0.8$. We simulated the dynamics of the model network using an Euler method with a time step of 0.5ms. Further details about the simulations and numerical estimates of mutual information and linear Fisher Information from simulated network responses can be found in Supplementary Information.

Acknowledgements

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

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

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

[3] 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.

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

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

[6] 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.

[7] 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.

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

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

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

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

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

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

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

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

[25] 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.

[26] 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.

[27] 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[28] Chengcheng Huang, Alexandre Pouget, and Brent David Doiron. Internally generated population activity in cortical networks hinders information transmission. *bioRxiv*, 2020.

[29] Maoz Shamir and Haim Sompolinsky. Nonlinear population codes. *Neural computation*, 16(6):1105–1136, 2004.

[30] 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.

[31] 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.

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

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

[34] 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.

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

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

[37] 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.

[38] 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.

[39] 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.

[40] 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.

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.18.389197; this version posted January 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[42] 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.

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

[44] 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.

[45] 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.

[46] 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.

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

[48] 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.

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

[50] 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.

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

[52] 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.

[53] 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.

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