The Bayesian sampling in a canonical recurrent circuit with a diversity of inhibitory interneurons: Supplementary Information

Eryn Sale[1,2]
eryn.sale@utsouthwestern.edu

Wen-Hao Zhang[1,2]
wenhao.zhang@utsouthwestern.edu

[1]Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center
[2]O'Donnell Brain Institute, UT Southwestern Medical Center

Contents

1 **Supplementary figures**   **3**

2 **The generative model and Bayesian sampling**   **4**
   2.1 The stimulus likelihood from feedforward inputs 4
   2.2 Langevin sampling 4
   2.3 Hamiltonian sampling 5
   2.4 The mixture of Langevin and Hamiltonian sampling 6
   2.5 Computing the equilibrium distribution via Fokker-Planck approach 6

3 **Theoretical analysis of the nonlinear circuit dynamics**   **7**
   3.1 Verifying the Gaussian ansatz of equilibrium attractor states 7
   3.2 Critical Recurrent Weight 9
   3.3 The circuit dynamics on the stimulus feature manifold 10

4 **Bayesian sampling in the recurrent circuit dynamics**   **11**
   4.1 Bridging the circuit dynamics with mixed sampling 11
   4.2 Conditions of Bayesian sampling in the circuit 12
   4.3 Evaluating sampling performance by eigenvalue analysis 13

5 **Bivariate stimulus posterior sampling in coupled recurrent circuit**   **14**
   5.1 The coupled circuit dynamics on the stimulus manifold 15
   5.2 Identifying the bivariate stimulus prior 16

6 **Simulation details and model parameters**   **16**
   6.1 Network parameters and simulation 16

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

6.2   Read out stimulus samples from the population responses   16

6.3   Power spectrum analysis   17

6.4   Comparing the sampling distributions with posteriors   18

6.5   Reproducing E neurons' tuning curves from modulating interneurons   18

2

1 Supplementary figures
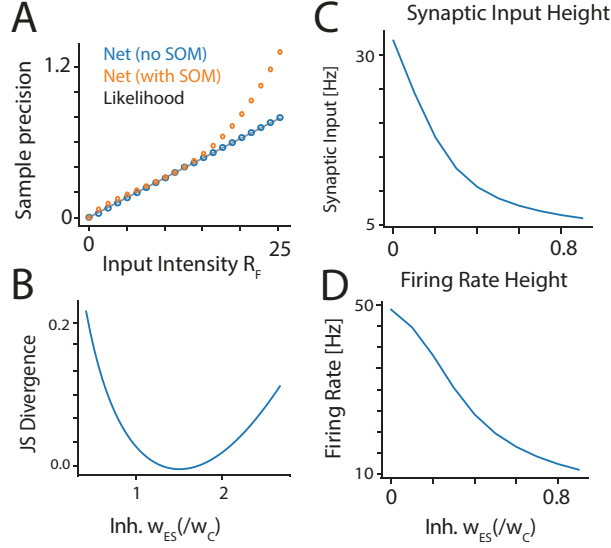
Figure S1: (A) Sample precision for likelihood distribution from generative model and network with or without SOM input. The circuit with SOM has higher sampling precision than without for high feedforward input intensity. (B) With increasing inhibitory to excitatory input, there exists a local minimum where the JS Divergence reaches zero, properly sampling the theoretical posterior distribution. (C-D) Synaptic input and Firing Rate height decrease for excitatory neurons with increasing SOM input.
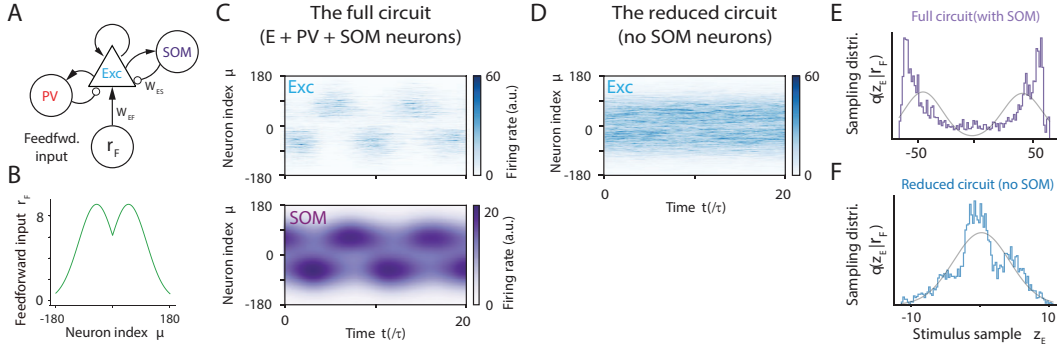
Figure S2: Sampling bimodal posteriors in the proposed circuit with interneurons when simultaneously presenting two stimuli to the circuit model. (A-B) The neural circuit model (A) that receives a bimodal feedforward input (B) with two stimuli located at $-50°$ and $50°$, mimicking the mixture of two orientations. (C) The population responses of E and SOM neurons in the full circuit (shown in A) when receiving the bimodal feedforward inputs shown in (B).(D) Same as (C) but removing SOM neurons in (A).Without SOM neurons, the E neurons are not able to distinguish the two stimuli located at different locations, and only sample a unimodal distribution. (E-F) The sampling distributions read out from excitatory (E) neurons from the full circuit model (E) and the reduced circuit after removing SOM neurons (F). For figure parameters, see Table S3.

3

2 The generative model and Bayesian sampling

2.1 The stimulus likelihood from feedforward inputs

We present the math of deriving the stimulus likelihood function (Eq. (7)) from the feedforward inputs (Eq. (3)). To ease the reading, we copy the definition of stochastic feedforward input here (Eq. (3) in the main text),

$$\mathbf{r}_F(\theta|z) \sim \text{Poisson}[\lambda_F(\theta|z)], \quad \lambda_F(\theta|z) = R_F \exp[-(\theta - z)^2/2a^2],$$

Substituting the feedforward firing rate $\lambda_F(\theta|z)$ into the Poisson distribution, the probability of observing a specific feedforward input $\mathbf{r}_F$ given the stimulus feature z is (the subscript F is suppressed for concision),

$$p(\mathbf{r}|z) = \prod_{j=1}^{N_E} \text{Poisson}\left(\mathbf{r}_j|\boldsymbol{\lambda}_j \Delta t\right) = \prod_{j=1}^{N_E} \frac{(\lambda_j \Delta t)^{\mathbf{r}_j}}{\mathbf{r}_j!} \exp(-\lambda_j \Delta t), \tag{S1}$$

Taking the logarithm,

$$\begin{aligned} \ln p(\mathbf{r}|z) &= \sum_j \left[\mathbf{r}_j \ln(\lambda_j \Delta t) - \ln(\mathbf{r}_j!) - \lambda_j\right], \\ &= \sum_j \mathbf{r}_j \ln(\lambda_j \Delta t) + \text{const.} \end{aligned} \tag{S2}$$

To obtain the last line in the above equation, we assume the sum of population firing rate $\sum_j \lambda_j(z)$ is a constant irrelevant to z, which is true in a homogeneous population with a large number of neurons. Substituting the Gaussian profile of feedforward firing rate $\lambda_F(z)$,

$$\begin{aligned} \ln p(\mathbf{r}|z) &= -\sum_j \mathbf{r}_j \frac{(\theta - z)^2}{2a^2} + \text{const,} \\ &= -\frac{(z - \mu_z)^2}{2\Lambda^{-1}} + \text{const,} \end{aligned} \tag{S3}$$

where

$$\mu_z = \frac{\sum_j \mathbf{r}(\theta_j)\theta_j}{\sum_j \mathbf{r}(\theta_j)}, \quad \Lambda = a^{-2} \sum_j \mathbf{r}(\theta_j). \tag{S4}$$

which is the Eq. (8) in the main text.

In our theoretical calculation below, we approximate the likelihood precision as a function of the peak feedforward firing rate R_F (Eq. 3),

$$\begin{aligned} \Lambda &\approx a^{-2} \sum_j \lambda_F(\theta_j), \\ &\approx a^{-2}\rho \int \lambda_F(\theta)d\theta \\ &= a^{-2}\rho R_F \int e^{-(\theta - z)^2/2a^2} d\theta \\ &= \sqrt{2\pi}\rho a^{-1} R_F, \end{aligned} \tag{S5}$$

where the first approximation is approximating the sum of feedforward population spike counts as the sum of the feedforward input firing rate, which works well in the case of a large number of neurons. The second approximation comes from converting the summation into integral, with $1/\rho$ denoting the distance between neighbor neurons in the stimulus feature manifold.

2.2 Langevin sampling

We present the background of utilizing Langevin dynamics to sample a posterior,

$$\frac{dz_t}{dt} = (\tau_L)^{-1} \frac{d \ln p(z|\mathbf{r}_F)}{dz} + \sqrt{2\tau_L^{-1}} \xi_t, \tag{S6}$$

4

where τ_L is the time constant controlling the sampling speed and ξ_t is a standard Gaussian white noise with zero mean and unit variance. Importantly, to sample the posterior, i.e., where the equilibrium distribution of z generated by the above dynamics is the same as the posterior, the drift and diffusion coefficients in the above equation should govern by the same τ_L. Note that τ_L only changes the speed of sampling, how long the z_t goes into equilibrium, but won't change the equilibrium distribution.

A characteristic of Langevin sampling is the cross correlation function of samples exponentially decaying with time,

$$\rho(\Delta t) = \exp(\Delta t/\tau_L), \tag{S7}$$

which is used in Fig. 2F in the main text.

2.3 Hamiltonian sampling

Hamiltonian Monte Carlo is a Markov Chain Monte Carlo (MCMC) machine learning method that uses gradient information to draw samples from a target distribution[1, 2]. The Hamiltonian function $H(z,y)$ is defined as the sum of potential energy $U(z)$, a function of the state z, and the kinetic energy with momentum $K(y)$, a function of the momentum y,

$$H(z,y) = U(z) + K(y) \tag{S8}$$

To utilize Hamiltonian sampling to sample the distribution $p(z)$ (e.g., posterior), the potential energy $U(z)$ is defined as,

$$U(z) = -\ln p(z). \tag{S9}$$

There is freedom of defining the kinetic energy $K(y)$ and a usual choice is,

$$K(y) = \frac{1}{2}y^T M^{-1} y. \tag{S10}$$

Hamiltonian dynamics (without noise) is

$$
\begin{aligned}
\dot{z} &= \frac{\partial H(z,y)}{\partial y} = \frac{\partial K(y)}{\partial y} = M^{-1}y; \\
\dot{y} &= -\frac{\partial H(z,y)}{\partial z} = -\frac{\partial U(z)}{\partial z} \triangleq -\nabla U(z).
\end{aligned}
\tag{S11}
$$

It can be proved that in the above Hamiltonian dynamics the Hamiltonian function $H(z,y)$ remains unchanged over time. Intuitively, using an analogy from classic mechanics, imagine a ball rolling down a hill. In the conservation of energy principle dictating Hamiltonian mechanics, as the ball rolls down the hill, it loses potential energy, but gains kinetic energy. This kinetic energy in the form of momentum, allows the ball to roll up the next hill.

To utilize the Hamiltonian dynamics to draw random samples from the posterior, a common strategy is injecting noise into the y dynamics,

$$
\begin{aligned}
\dot{z} &= M^{-1}y; \\
\dot{y} &= -\nabla U(z) + \xi_t.
\end{aligned}
\tag{S12}
$$

In the present study, we consider a Hamiltonian sampling with friction, because it can be mapped to the proposed circuit with a diversity of interneurons [2, 3],

$$
\begin{aligned}
\tau_z \dot{z} &= M^{-1}y; \\
\tau_y \dot{y} &= -\nabla U(z) - \beta M^{-1}y + (2\beta\tau_y)^{1/2}\eta_t
\end{aligned}
\tag{S13}
$$

where β represents the friction strength which helps decrease the influence of noise on the system and M is the inertia of the system. The above Hamiltonian dynamics are equivalent to second-order Langevin dynamics. Intuitively, it corresponds to friction forces existing when a ball rolls down a hill, e.g., friction from the grass or dirt under the ball, which prevents it from continuing up the hill. The friction effectively lowers the energy function $H(z,y)$, thereby mitigating the effect of noise [2].

5

2.4 The mixture of Langevin and Hamiltonian sampling

We present the math details linking the proposed circuit model with sampling-based Bayesian inference. Based on the functional form of the bump position dynamics in the proposed circuit model, we hypothesize the proposed circuit model performs a mixed sampling of Langevin sampling and the Hamlitonian sampling. To explore such a possibility, we rewrite the Hamiltonian sampling dynamics below.

$$
\begin{aligned}
\dot{z} &= \tau_z^{-1} y \\
\tau_y \dot{y} &= -\beta y + \nabla \ln p(z|\mathbf{r}_F) + (2\beta\tau_z)^{1/2}\eta_t.
\end{aligned}
\tag{S14}
$$

Meanwhile, the Langevin sampling dynamics of the posterior distribution is,

$$
\dot{z} = \tau_L^{-1} \nabla \ln p(z|\mathbf{r}_F) + (\tau_L/2)^{-1/2}\xi_t,
\tag{S15}
$$

where τ_L is the time constant of Langevin sampling controlling the convergence speed. It can be verified that both Langevin and Hamiltonian sampling dynamics converge to the same posterior distribution. Then we mix both sampling dynamics together, i.e., summing up the z dynamics in the Hamiltonian sampling (Eq. S14) and Langevin sampling (Eq. S15),

$$
\begin{aligned}
\dot{z} &= \tau_z^{-1} y + \tau_L^{-1} \nabla \ln p(z|\mathbf{r}_F) + (\tau_L/2)^{-1/2}\xi_t, \\
\tau_y \dot{y} &= -\beta y + \nabla \ln p(z|\mathbf{r}_F) + (2\beta\tau_z)^{1/2}\eta_t
\end{aligned}
\tag{S16}
$$

Substituting the gradient of the log posterior into the above equation, i.e., $\nabla \ln p(z|\mathbf{r}_F) = \Lambda(\mu_z - z)$ (Eqs. 6 - 8),

$$
\begin{aligned}
\dot{z} &= \tau_z^{-1} y + \tau_L^{-1}\Lambda(\mu_z - z) + (\tau_L/2)^{-1/2}\xi_t, \\
\tau_y \dot{y} &= -\beta y + \Lambda(\mu_z - z) + (2\beta\tau_z)^{1/2}\eta_t
\end{aligned}
\tag{S17}
$$

2.5 Computing the equilibrium distribution via Fokker-Planck approach

Hamiltonian sampling

Here we provide an analysis showing that Hamiltonian sampling can sample the desired distribution $p(z)$ (with $p(z)$ as the equilibrium distribution), which is defined as,

$$
\pi(z,y) = \exp[-H(z,y)] = \exp[-U(z) - K(y)] = \exp[\ln p(z) - K(y)]
\tag{S18}
$$

Reorganizing Eq. (S13),

$$
\begin{aligned}
\dot{z} &= \tau_z^{-1} y \\
\dot{y} &= -\tau_y^{-1}\beta y - \tau_y^{-1}\nabla U(z) + \tau_y^{-1}(2\beta\tau_z)^{1/2}\eta_t,
\end{aligned}
$$

where $-\nabla U(z) = \nabla \ln p(z)$, To convert into matrix notation, we define the vector $Z = (z,y)^\top$.

$$
\frac{d}{dt}\begin{pmatrix} z \\ y \end{pmatrix} = -\begin{pmatrix} 0 & -\tau_z^{-1} \\ \tau_y^{-1} & \tau_y^{-1}\beta \end{pmatrix}\begin{pmatrix} \nabla U(z) \\ y \end{pmatrix} + \sqrt{2}\begin{pmatrix} 0 & 0 \\ 0 & \tau_y^{-1}(\tau_z\beta)^{1/2} \end{pmatrix}\eta_t.
$$

In the Hamiltonian sampling, we have the freedom to define the kinetic energy $K(y)$. To facilitate the calculation of equilibrium distribution, we reorganize the above equation as

$$
\frac{d}{dt}\begin{pmatrix} z \\ y \end{pmatrix} = -\begin{pmatrix} 0 & -\tau_y^{-1} \\ \tau_y^{-1} & \beta\tau_z/\tau_y^2 \end{pmatrix}\begin{pmatrix} \nabla U(z) \\ \tau_y y/\tau_z \end{pmatrix} + \sqrt{2}\begin{pmatrix} 0 & 0 \\ 0 & (\beta\tau_z)^{1/2}/\tau_y \end{pmatrix}\eta_t
\tag{S19}
$$

where the kinetic energy becomes $K(y) = -\tau_y y^2/(2\tau_z)$. The time evolution of the distribution $Z = (z,y)^\top$ in the above dynamics is governed by the following Fokker-Planck equation [4, 2],

$$
\partial_t p(Z,t) = \nabla^\top M[p(Z,t)\nabla H(Z)] + \nabla^\top[D\nabla p(Z,t)]
$$

where $H(Z)$ is the Hamiltonian defined in Eq. (S8), and the matrices are defined as

$$
D = \begin{pmatrix} 0 & 0 \\ 0 & \beta\tau_z/\tau_y^2 \end{pmatrix}, \quad M = \begin{pmatrix} 0 & -\tau_y^{-1} \\ \tau_y^{-1} & \beta\tau_z/\tau_y^2 \end{pmatrix}
\tag{S20}
$$

6

Further, decomposing the matrix M as the sum of the matrix D and an anti-symmetric matrix G,

$$M = \underbrace{\begin{pmatrix} 0 & -\tau_y^{-1} \\ \tau_y^{-1} & 0 \end{pmatrix}}_{G} + \underbrace{\begin{pmatrix} 0 & 0 \\ 0 & \beta\tau_z/\tau_y^2 \end{pmatrix}}_{D}$$

Then, the Fokker-Planck equation can be converted into,

$$\partial_t p_t(Z) = \nabla^\top (D + G)[\nabla H(Z)p_t(Z)] + \nabla^\top [D\nabla\partial p_t(Z)]$$

Since G is anti-symmetric, it can be checked that

$$\nabla^\top [G\nabla p_t(Z)] = \tau_y^{-1}[-\partial_z\partial_y p_t(z, y) + \partial_z\partial_y p_t(z, y)] = 0.$$

Therefore, we have,

$$\nabla^\top [D\nabla\partial_t p(Z)] = \nabla^\top [(D + G)\nabla\partial p_t(Z)].$$

Substituting the above relation into the Fokker-Planck equation,

$$\partial_t p_t(Z) = \nabla^\top (D + G)[\nabla H(Z)p_t(Z) + \nabla\partial p_t(Z)].$$

Therefore, the equilibrium distribution should satisfy

$$\nabla H(Z)p_t(Z) + \nabla\partial p_t(Z) = 0.$$

It can be verified that the $p(Z) \propto \exp[-H(Z)]$ is indeed the equilibrium distribution.

The mixture of Langevin and Hamiltonian sampling

For the mixed sampling of Langevin and Hamiltonian dynamics (Eq. S16), it can be converted into the below matrix form,

$$\frac{d}{dt}\begin{pmatrix} z \\ y \end{pmatrix} = -\begin{pmatrix} \tau_L^{-1} & -\tau_y^{-1} \\ \tau_y^{-1} & \beta\tau_z/\tau_y^2 \end{pmatrix}\begin{pmatrix} \nabla U(z) \\ \frac{\tau_y}{\tau_z}y \end{pmatrix} + \sqrt{2}\begin{pmatrix} \tau_L^{-1/2} & 0 \\ 0 & (\beta\tau_z)^{1/2}/\tau_y \end{pmatrix}\eta_t \quad \text{(S21)}$$

where the drift term matrix can be decomposed as

$$\underbrace{\begin{pmatrix} \tau_L^{-1} & -\tau_y^{-1} \\ \tau_y^{-1} & \beta\tau_z/\tau_y^2 \end{pmatrix}}_{M} = \underbrace{\begin{pmatrix} 0 & -\tau_y^{-1} \\ \tau_y^{-1} & 0 \end{pmatrix}}_{G} + \underbrace{\begin{pmatrix} \tau_L^{-1} & 0 \\ 0 & \beta\tau_z/\tau_y^2 \end{pmatrix}}_{D'}$$

Then its corresponding Fokker-Plank equation is,

$$\partial_t p_t(Z) = \nabla^\top (D' + G)[\nabla H(Z)p_t(Z)] + \nabla^\top [D'\nabla\partial_t p(Z)] \quad \text{(S22)}$$

Redo the similar calculation as above, we can check that $p(Z) \propto \exp[-H(Z)]$ is still the equilibrium distribution. That is, the mixture of Langevin and Hamiltonian sampling dynamics doesn't alter the equilibrium distribution.

3 Theoretical analysis of the nonlinear circuit dynamics

3.1 Verifying the Gaussian ansatz of equilibrium attractor states

We present the math in verifying the Gaussian anstaz of population responses in the proposed network model (Eq. 9). To ease reading, we copy the network dynamics (Eq. 1 and Eq. 5) below.

$$\tau\frac{\partial \mathbf{u}_E(\theta, t)}{\partial t} = -\mathbf{u}_E(\theta, t) + \rho\sum_{X=E,F,S}(\mathbf{W}_{EX} * \mathbf{r}_X)(\theta, t) + \sqrt{\tau\mathsf{F}_E[\mathbf{u}_E(\theta, t)]_+}\xi(\theta, t),$$

$$\tau\frac{\partial \mathbf{u}_S(\theta, t)}{\partial t} = -\mathbf{u}_S(\theta, t) + \rho\sum_{X=E,F}(\mathbf{W}_{SX} * \mathbf{r}_X)(\theta, t); \quad \mathbf{r}_S(\theta, t) = g_S \cdot [\mathbf{u}_S(\theta, t)]_+, \quad \text{(S23)}$$

7

For an equilibrium attractor state, the mean responses in the network satisfy,

$$\langle \mathbf{u}_E(\theta) \rangle = \rho \sum_{X=E,F,S} (\mathbf{W}_{EX} \cdot \langle \mathbf{r}_X \rangle)(\theta),$$

$$\langle \mathbf{u}_S(\theta) \rangle = \rho \sum_{X=E,F} (\mathbf{W}_{SX} \cdot \langle \mathbf{r}_X \rangle)(\theta),$$

(S24)

where $\langle \cdot \rangle$ denotes the average in the equilibrium.

We propose the following Gaussian ansatz for synaptic input of E neurons as in Eq. (9),

$$\langle \mathbf{u}_E(\theta) \rangle = U_E \exp\left[-\frac{(\theta - z_E)^2}{4a_E^2} \right].$$

(S25)

Next we will verify the above Gausisan ansatz works for the proposed recurrent circuit dynamics. The ansatz of the E neurons' firing rates can be obtained by substituting $\langle U_E(\theta) \rangle$ into divisive normalization (Eq. 4),

$$\langle \mathbf{r}_E(\theta) \rangle = \frac{[U_E^2(\theta, t)]^2}{1 + \rho w_{EP} \int [U_E(\theta, t)]^2 d\theta} = \underbrace{\frac{U_E^2}{1 + \rho w_{EP} U_E^2 \sqrt{2\pi}a_E}}_{R_E} \exp\left[-\frac{(\theta - z_E)^2}{2a_E^2} \right].$$

(S26)

Next, we substitute the above Gaussian ansatz into the stationary state of the circuit dynamics (S24). We consider the connection width between E and SOM neuronal population to be different, a_E and a_S, respectively (Table S1). The neural population input from the neuronal population of type Y to the one with type X in the circuit model can generally be calculated as,

$$\langle I_{XY}(\theta) \rangle = \rho \mathbf{W}_{XY} * \langle \mathbf{r}_Y(\theta) \rangle$$

$$= \rho \int \mathbf{w}_{XY}(\theta' - \theta) \langle \mathbf{r}_Y(\theta) \rangle d\theta'$$

$$= \frac{\rho w_{XY} R_Y}{\sqrt{2\pi}a_{XY}} \int \exp\left[-\frac{(\theta' - \theta)^2}{2a_{XY}^2} - \frac{(\theta' - z_Y)^2}{2a_Y^2} \right] d\theta'$$

$$= \frac{\rho w_{XY} R_Y}{\sqrt{2\pi}a_{XY}} \int \exp\left[-\frac{\left(\theta' - \frac{a_Y^2 \theta + a_{XY}^2 z_Y}{a_{XY}^2 + a_Y^2}\right)^2}{2\frac{a_{XY}^2 a_Y^2}{a_{XY}^2 + a_Y^2}} - \frac{(\theta - z_Y)^2}{2(a_{XY}^2 + a_Y^2)} \right] d\theta',$$

$$= \rho w_{XY} R_Y \frac{a_Y}{\sqrt{a_{XY}^2 + a_Y^2}} \exp\left[-\frac{(\theta - z_Y)^2}{2(a_{XY}^2 + a_Y^2)} \right].$$

(S27)

Specifically, plugging in the Gaussian ansatz of E population firing rate (Eq. S26), the recurrent connections leads to the following input,

$$\langle I_{EE}(\theta) \rangle = \rho \mathbf{W}_{EE} * \langle \mathbf{r}_E(\theta) \rangle = \frac{\rho}{\sqrt{2}} w_{EE} R_E \exp\left[-\frac{(\theta - z_E)^2}{4a_E^2} \right].$$

(S28)

Similarly, the feedforward input term is,

$$\langle I_{EF}(\theta) \rangle = \rho \mathbf{W}_{EF} * \langle \mathbf{r}_F(\theta) \rangle = \frac{\rho}{\sqrt{2}} w_{EF} R_F \exp\left[-\frac{(\theta - \mu_E)^2}{4a_E^2} \right].$$

(S29)

Repeat the process for the E to SOM input,

$$\langle I_{SE}(\theta) \rangle = \rho \mathbf{W}_{SE} * \langle \mathbf{r}_E(\theta) \rangle = \rho w_{SE} R_E \frac{a_E}{\sqrt{a_{SE}^2 + a_E^2}} \exp\left[-\frac{(\theta - z_E)^2}{2(a_{SE}^2 + a_E^2)} \right].$$

(S30)

We have analytically calculated the outputs from E neurons. We next calculate the response of SOM neurons. First, we assume SOM neurons receive feedforward inputs, with $\mathbf{u}_S$ being $I_{SE} + I_{SF}$, where we assume the two inputs have the same width. Then, we can derive $\mathbf{r}_S$ from u_S. Substituting the ansatz for the SOM synaptic input,

$$\langle \mathbf{u}_S(\theta) \rangle = I_{SE}(\theta) + I_{SF}(\theta) \triangleq U_S \exp\left[-\frac{(\theta - z_S)^2}{4a_S^2} \right].$$

(S31)

8

where $2a_S^2 = a_{SE}^2 + a_E^2$. We can then derive the SOM firing rate,

$$\langle \mathbf{r}_S(\theta) \rangle = g_S[U_S(\theta, t)] = \underbrace{g_S U_S}_{R_S} \exp\left[-\frac{(\theta - z_S)^2}{4a_S^2}\right]. \tag{S32}$$

Eventually, the inhibitory inputs from SOM to E neurons are calculated as,

$$\langle I_{ES}(\theta) \rangle = \rho \mathbf{W}_{ES} * \langle \mathbf{r}_S(\theta) \rangle = \rho w_{ES} R_S \frac{a_S}{\sqrt{a_{ES}^2 + a_S^2}} \exp\left[-\frac{(\theta - z_S)^2}{2(a_{ES}^2 + a_S^2)}\right], \tag{S33}$$

The feedforward input into SOM is,

$$\langle I_{SF}(\theta) \rangle = \rho \mathbf{W}_{SF} * \langle \mathbf{r}_F(\theta) \rangle = \frac{\rho}{\sqrt{2}} w_{SF} R_F \exp\left[-\frac{(\theta - \mu_S)^2}{4a_S^2}\right]. \tag{S34}$$

To validate the Gaussian ansatz, we need the width of $\langle I_{ES}(\theta) \rangle$ (Eq. S33) is the same as the width of the Gaussian ansatz (Eq. S25), which constrains the connection widths from (S33) and (S31).

$$2a_E^2 = a_{ES}^2 + a_S^2, \tag{S35}$$

Recalling that $2a_S^2 = a_{SE}^2 + a_E^2$, and combining with the above equation,

$$a_E^2 = a_{ES}^2 + a_{SE}^2. \tag{S36}$$

which constrains the connection width between E and SOM neurons.

By appropriately setting the connection width suggested by the above equation, and substituting all above Gaussian functions of synaptic inputs, firing rates into the circuit dynamics (Eqs. 1 and 5),

$$U_E \exp\left[-\frac{(\theta - z_E)^2}{4a_E^2}\right] = \frac{\rho}{\sqrt{2}} \left[w_{EE} R_E e^{-(\theta - z_E)^2/4a_E^2} + w_{EF} R_F e^{-(\theta - \mu_z)^2/4a_E^2}\right]$$

$$+ \rho w_{ES} R_S \frac{a_S}{\sqrt{2}a_E} \exp\left[-\frac{(\theta - z_S)^2}{4a_E^2}\right]. \tag{S37}$$

$$U_S \exp\left[-\frac{(\theta - z_S)^2}{4a_S^2}\right] = \rho w_{SE} R_E \frac{a_E}{\sqrt{a_{SE}^2 + a_E^2}} \exp\left[-\frac{(\theta - z_E)^2}{2(a_{SE}^2 + a_E^2)}\right],$$

Since the above equations are summations of Gaussian functions, it can be checked that when the positions of Gaussian functions are the same, i.e., $z_E = z_S = \mu_z$, the sum of two Gaussian functions will also be a Gaussian function. Therefore we complete our verification about the Gaussian ansatz of equilibrium attractor state.

3.2   Critical Recurrent Weight

We want to scale all the connection strength weights by the smallest recurrent connection strength where the network can hold persistent activity in the absence of feedforward input. Since all exponential functions in the Eq. (S37) can be canceled (the mean of Gaussian functions will be the same in equilibrium), we have

$$U_E = \frac{\rho}{\sqrt{2}}\left[w_{EE} R_E + \frac{a_S}{a_E} w_{ES} R_S\right],$$

$$U_S = \frac{\rho}{\sqrt{2}} \frac{a_E}{a_S} w_{SE} R_E. \tag{S38}$$

From Eq. (S32), we know $R_S = g_S U_S$. Combining this result with Eq. (S38),

$$U_E = \frac{\rho}{\sqrt{2}} R_E \left[w_{EE} + \frac{\rho}{\sqrt{2}} w_{ES} g_S w_{SE}\right] \tag{S39}$$

Recalling the definition of E population firing rate height defined in Eq. (S26)

$$R_E = \frac{U_E^2}{1 + \sqrt{2\pi} k \rho a_E U_E^2}.$$

9

Substituting the above Eq. into Eq. (S39)

$$U_E = \frac{\rho U_E^2}{\sqrt{2} + 2\sqrt{\pi}k\rho a_E U_E^2}\left[w_{EE} + \frac{\rho}{\sqrt{2}}w_{ES}w_{SE}g_S\right].$$

The above equation can be converted into a quadratic equation of U_E,

$$2\sqrt{\pi}k\rho a_E U_E^2 - \rho U_E \underbrace{\left[w_{EE} + \frac{\rho}{\sqrt{2}}w_{ES}w_{SE}g_S\right]}_{w_c} + \sqrt{2} = 0. \tag{S40}$$

with solution

$$U_E = \frac{\rho w_c \pm \sqrt{\rho^2 w_c^2 - 8\sqrt{2\pi}k\rho a_E}}{4\sqrt{\pi}k\rho a_E} \tag{S41}$$

To have non-zero persistent activity without feedforward input, there must exist a value of w_c where the inside of the square root function is positive. Therefore, the smallest w_c for holding the persistent activity is,

$$w_c^2 > \frac{8\sqrt{2\pi}ka_E}{\rho}$$

3.3 The circuit dynamics on the stimulus feature manifold

Previous theoretical studies on the recurrent circuit dynamics have analytically computed the eigenvectors corresponding to the stimulus feature manifold in the neuronal population X [5, 6],

$$\phi_1(\theta|z_X) = \frac{d\langle U_X|\theta\rangle}{dz_X} \propto (\theta - z_X)\exp\left[-\frac{(\theta - z_X)^2}{4a_X^2}\right] \tag{S42}$$

where the above eigenvector has the largest eigenvalue, denoting the circuit dynamics is dominated by the changes along the stimulus feature manifold.

To complete the projection of the high dimensional network dynamics onto the manifold, we first substitute the Gaussian ansatz into the circuit dynamics (Eq. 1).

$$\tau U_E \frac{d}{dt}\exp\left[-\frac{(\theta - z_E)^2}{4a_E^2}\right] = \tau U_E \frac{\theta - z_E}{2a_E^2}\frac{dz_E}{dt}\exp\left[-\frac{(\theta - z_E)^2}{4a_E^2}\right],$$

$$= \left(-U_E + \frac{\rho w_{EE}R_E}{\sqrt{2}}\right)\exp\left[-\frac{(\theta - z_E)^2}{4a_E^2}\right] + \frac{\rho w_{ES}R_S a_S}{\sqrt{2}a_E}\exp\left[-\frac{(\theta - z_S)^2}{4a_E^2}\right]$$

$$+ w_{EF}R_{EF}\exp\left[-\frac{(\theta - \mu_z)^2}{4a_E^2}\right] + \sqrt{\tau F_E U_E}\exp\left[-\frac{(\theta - z_E)^2}{8a_E^2}\right]\xi(\theta, t) \tag{S43}$$

Projecting both sides onto eigenvector $\phi_1(\theta|z_E)$,

$$\tau U_E \frac{dz_E}{dt} = \frac{\rho}{\sqrt{2}}w_{ES}R_S\frac{a_S}{a_E}(z_S - z_E)\exp\left[-\frac{(z_S - z_E)^2}{8a_E^2}\right],$$

$$+ \frac{\rho}{\sqrt{2}}w_{EF}R_F(\mu_z - z_E)\exp\left[-\frac{(\mu_z - z_E)^2}{8a_E^2}\right] + \sqrt{\frac{8a_E F_E}{3\sqrt{3\pi}}}\sqrt{\tau_L U_E}\eta_t \tag{S44}$$

The process is repeated for the SOM neurons network, starting with substituting ansatz into network dynamics and simplifying.

$$\tau U_S \frac{d}{dt}\exp\left[-\frac{(\theta - z_S)^2}{4a_S^2}\right] = \tau U_S \frac{\theta - z_S}{2a_S^2}\frac{dz_S}{dt}\exp\left[-\frac{(\theta - z_S)^2}{4a_S^2}\right],$$

$$= -U_S\exp\left[-\frac{(\theta - z_S)^2}{4a_S^2}\right] + w_{SE}U_E\exp\left[-\frac{(\theta - z_S)^2}{4a_S^2}\right] \tag{S45}$$

$$+ \frac{\rho}{\sqrt{2}}w_{SF}R_F\exp\left[-\frac{(\theta - \mu_z)^2}{4a_S^2}\right].$$

10

Projecting both sides onto the eigenvector $\phi(\theta|z_S) = (\theta - z_S)e^{-(\theta-z_S)^2/4a_S^2}$,

$$\tau U_S \frac{dz_S}{dt} = \frac{\rho}{\sqrt{2}} w_{SE} R_E \frac{a_E}{a_S}(z_E - z_S)\exp\left[-\frac{(z_E - z_S)^2}{8a_S^2}\right] + w_{SF}R_F(\mu_z - z_S)\exp\left[\frac{(\mu_z - z_S)^2}{8a_S^2}\right] \tag{S46}$$

From here, we assume the difference between neuronal populations' positions is small enough compared to the connection width a, i.e., $|z_E - z_S|$ and $|\mu_z - z_E| << 4a_X$. In this case, the projected circuit dynamics can be simplified by ignoring exponential terms in Eqs. (S44) and (S46),

$$\tau U_E \frac{dz_E}{dt} = \frac{\rho}{\sqrt{2}}\left[w_{ES}R_S\frac{a_S}{a_E}(z_S - z_E) + w_{EF}R_F(\mu_z - z_E)\right] + \sqrt{\frac{8F_E a_E}{3\sqrt{3\pi}}}\sqrt{\tau U_E}\eta_t,$$

$$\tau U_S \frac{dz_S}{dt} = \frac{\rho}{\sqrt{2}}\left[\frac{a_E}{a_S}w_{SE}R_E(z_E - z_S) + w_{SF}R_F(\mu_z - z_S)\right] \tag{S47}$$

To simplify notations, we define

$$\tau_X = \tau U_X, \quad g_{XY} = \frac{\rho a_Y}{\sqrt{2}a_X}w_{XY}R_Y, \quad \sigma_E^2 = \frac{8aF_E}{3\sqrt{3\pi}}, \tag{S48}$$

and then the Eq. (S47) is simplified into

$$\tau_E\dot{z}_E = g_{ES}(z_S - z_E) + g_{EF}(\mu_z - z_E) + \sigma_E\sqrt{\tau_E}\eta_t,$$

$$\tau_S\dot{z}_S = g_{SE}(z_E - z_S) + g_{SF}(\mu_z - z_S), \tag{S49}$$

which are the bump position dynamics of Eqs. (11) and (12) in the main text.

4 Bayesian sampling in the recurrent circuit dynamics

In this section, we compare the 1D and 2D dynamics from (S49) and (S83) to the theoretical mixed sampling in Sec. 2.4.

4.1 Bridging the circuit dynamics with mixed sampling

We see the z dynamics in the mixed sampling (Eq. S17) resembles the z_E dynamics in the proposed circuit model (Eq. S49). We next find the auxiliary variable y in the proposed circuit. We firstly decompose the term from feedforward input in the z_E dynamics into two parts,

$$\tau_E\dot{z}_E = [g_{ES}(z_S - z_E) + (1 - \alpha_L)g_{EF}(\mu_z - z_E)] + [\alpha_L g_{EF}(\mu_z - z_E) + \sigma_E\sqrt{\tau_E}\eta_t],$$

$$= \qquad\qquad y_S \qquad\qquad\qquad\quad + [\alpha_L g_{EF}(\mu_z - z_E) + \sigma_E\sqrt{\tau_E}\eta_t], \tag{S50}$$

where $\alpha_L \in [0, 1]$ denotes the proportion of feedforward input coming from the Langevin sampling dynamics. Then the y_S is similar to the auxiliary variable in the mixed sampling dynamics (Eq. S17). To derive the dynamics of y_S, we take the derivative of y_S,

$$\dot{y}_S = g_{ES}(\dot{z}_S - \dot{z}_E) - (1 - \alpha_L)g_{EF}\dot{z}_E. \tag{S51}$$

Based on Eq. (S49), we have

$$\dot{z}_E - \dot{z}_S = \left(-\tau_E^{-1}g_{ES} - \tau_S^{-1}g_{SE}, \tau_E^{-1}g_{EF}, -\tau_S^{-1}g_{SF}\right)\begin{pmatrix} z_E - z_S \\ \mu_z - z_E \\ \mu_z - z_S \end{pmatrix} + \sigma_E\sqrt{\tau_E^{-1}}\eta_t$$

$$\dot{z}_E = \left(-\tau_E^{-1}g_{ES}, \tau_E^{-1}g_{EF}, 0\right)\begin{pmatrix} z_E - z_S \\ \mu_z - z_E \\ \mu_z - z_S \end{pmatrix} + \sigma_E\sqrt{\tau_E^{-1}}\eta_t$$

Substitute the above equations into Eq. (S51), the y_S dynamics can be converted into,

$$\dot{y}_S = \begin{pmatrix} g_{ES}(\tau_E^{-1}g_{ES} + \tau_S^{-1}g_{SE}) + (1 - \alpha_L)g_{EF}\tau_E^{-1}g_{ES} \\ -\tau_E^{-1}g_{EF}[g_{ES} + (1 - \alpha_L)g_{EF}] \\ \tau_S^{-1}g_{SF}g_{ES} \\ \sigma_E(\tau_E)^{-1/2}[g_{ES} - (1 - \alpha_L)g_{EF}] \end{pmatrix}^\top \begin{pmatrix} z_E - z_S \\ \mu_z - z_E \\ \mu_z - z_S \\ \eta_t \end{pmatrix} \tag{S52}$$

11

To derive a y_S dynamics with a decaying term of y_S itself, we utilize that (based on the definition of y_S, Eq. S50),

$$(z_E - z_S) = -g_{ES}^{-1}[y_S - (1-\alpha_L)g_{EF}(\mu_z - z_E)]$$

Substituting it back to Eq. (S52), we arrive at a y_S dynamics similar to the one defined in Hamiltonian sampling (Eq. S17),

$$\dot{y}_S = \begin{pmatrix} \tau_E^{-1}[g_{ES} + (1-\alpha_L)g_{EF}] + \tau_S^{-1}g_{SE} \\ -\alpha_L\tau_E^{-1}g_{EF}[g_{ES} - (1-\alpha_L)g_{EF}] + \tau_S^{-1}g_{SE}(1-\alpha_L)g_{EF} \\ \tau_S^{-1}g_{SF}g_{ES} \\ \sigma_E(\tau_E)^{-1/2}[g_{ES} - (1-\alpha_L)g_{EF}] \end{pmatrix}^{\top} \begin{pmatrix} -y_S \\ \mu_z - z_E \\ \mu_z - z_S \\ \eta_t \end{pmatrix}$$

$$\equiv (\beta_y, \beta_E, \beta_S, \sigma_y) \begin{pmatrix} -y_S \\ \mu_z - z_E \\ \mu_z - z_S \\ \eta_t \end{pmatrix}.$$

(S53)

Combining with the z_E dynamics (Eq. S50) together, the bump position dynamics of E and SOM neurons in the proposed circuit is reorganized into a form similar to the mixed sampling (Eq. S17)

$$\text{E neurons}: \quad \dot{z}_E = \tau_E^{-1}y_S + \alpha_L\tau_E^{-1}g_{EF}(\mu_z - z_E) + \sigma_E\tau_E^{-1/2}\eta_t \tag{S54}$$

$$\text{Auxiliary}: \quad \dot{y}_S = -\beta_y y_S + \beta_E(\mu_z - z_E) + \beta_S(\mu_z - z_S) + \sigma_y\eta_t. \tag{S55}$$

To help readers compare the circuit dynamics on the stimulus feature manifold with the mixed sampling dynamics, we copy the Eq. (S17) in below

$$\text{Sample } z: \quad \dot{z} = \tau_z^{-1}y + \tau_L^{-1}\Lambda(\mu_z - z) + (\tau_L/2)^{-1/2}\xi_t, \tag{S56}$$

$$\text{Auxiliary}: \quad \tau_y\dot{y} = -\beta y + \Lambda(\mu_z - z) + (2\beta\tau_z)^{1/2}\eta_t \tag{S57}$$

4.2 Conditions of Bayesian sampling in the circuit

The circuit weights must be set appropriately to sample from the posterior. Now we investigate the conditions for realizing Bayesian sampling in the proposed circuit model.

Langevin sampling part

To implement the Langevin sampling part in the z_E dynamics (Eq. S54, last two terms), we need

$$\frac{\alpha_L\tau_E^{-1}g_{EF}}{\Lambda} = \frac{\sigma_E^2\tau_E^{-1}}{2} \quad \Rightarrow \quad \sigma_E^2\Lambda = 2\alpha_L g_{EF}. \tag{S58}$$

Substituting the expression of g_{EF}, σ_E^2 (Eq. S48), and Λ (Eq. S5), it yields a constraint on the feedforward weight,

$$w_{EF} = \frac{\sqrt{\pi}\sigma_E^2}{a\alpha_L} = \left(\frac{2}{\sqrt{3}}\right)^3 \frac{\mathsf{F}_E}{\alpha_L}, \tag{S59}$$

which gives rise to the Eq. (19) in the main text.

Hamiltonian sampling part

Comparing Eqs. (S54 - S55) and Eqs. (S56 - S57), we can derive the conditions for realizing Hamiltonian sampling in the circuit,

$$\tau_z = \tau_E; \tag{S60a}$$

$$\beta_S = 0; \tag{S60b}$$

$$\frac{\tau_y}{1} = \frac{\beta}{\beta_y} = \frac{\Lambda}{\beta_E} = \frac{(2\beta\tau_z)^{1/2}}{\sigma_y}, \tag{S60c}$$

12

which is the same as the Eq. (20) in the main text. Combining the above three equations could yield the following condition,

$$\tau_y^2 = \frac{2\beta\tau_z}{\sigma_y^2} = \frac{2\tau_y\beta_y\tau_z}{\sigma_y^2} \quad \Leftrightarrow \quad \tau_y = \frac{2\beta_y\tau_z}{\sigma_y^2} = \frac{\Lambda}{\beta_E} \quad \Leftrightarrow \quad \Lambda\sigma_y^2 = 2\tau_E\beta_y\beta_E. \tag{S61}$$

Substituting the detailed expression of coefficients σ_y, β_y, and β_E (Eq. S53) in the above Eq. (S61),

$$\Lambda\sigma_E^2\tau_E^{-1}[g_{ES} - (1-\alpha_L)g_{EF}]^2 = 2\tau_E\Big[\tau_E^{-1}[g_{ES} + (1-\alpha_L)g_{EF}] + \tau_S^{-1}g_{SE}\Big]$$
$$\times g_{EF}\Big[-\alpha_L\tau_E^{-1}[g_{ES} - (1-\alpha_L)g_{EF}] + \tau_S^{-1}g_{SE}(1-\alpha_L)\Big]$$

Utilizing the relation (Eq. S58) to cancel $\Lambda\sigma_E^2$ with g_{EF} at two sides in the above equation,

$$\alpha_L\tau_E^{-2}[g_{ES} - (1-\alpha_L)g_{EF}]^2 = \Big[\tau_E^{-1}[g_{ES} + (1-\alpha_L)g_{EF}] + \tau_S^{-1}g_{SE}\Big]$$
$$\times \Big[-\alpha_L\tau_E^{-1}[g_{ES} - (1-\alpha_L)g_{EF}] + \tau_S^{-1}g_{SE}(1-\alpha_L)\Big]$$

To simplify notations in the above equation, we define

$$h_E \equiv \tau_E^{-1}[g_{ES} - (1-\alpha_L)g_{EF}]; \quad h_S \equiv \tau_S^{-1}g_{SE}. \tag{S62}$$

which simplified the above equation into,

$$\alpha_L h_E^2 = (h_E + h_S)[-\alpha_L h_E + (1-\alpha_L)h_S]. \tag{S63}$$

Reorganizing the above equation into a quadratic equation of h_E,

$$2\alpha_L \cdot h_E^2 + (2\alpha_L - 1)h_S h_E + (\alpha_L - 1)h_S^2 = 0 \tag{S64}$$

Then the root of h_E is,

$$h_E = h_S \frac{(1 - 2\alpha_L) \pm \sqrt{1 + 4\alpha_L - 4\alpha_L^2}}{4\alpha_L} \equiv G(\alpha_L) \cdot h_S \tag{S65}$$

Substituting the expressions of h_E and h_S (Eq. S62) back into the above equation,

$$\tau_E^{-1}[g_{ES} - (1-\alpha_L)g_{EF}] = G(\alpha_L)\tau_S^{-1}g_{SE}. \tag{S66}$$

Further, substituting the coefficients in Eq. (S48) we eventually have

$$\big(U_E^{-1}R_S\big) \cdot w_{ES} - \big[(1-\alpha_L)U_E^{-1}R_F\big] \cdot w_{EF} = \big[G(\alpha_L)U_S^{-1}R_E\big] \cdot w_{SE} \tag{S67}$$

which is the Eq. (21) in the main text.

4.3 Evaluating sampling performance by eigenvalue analysis

We calculate the eigenvalues of the circuit dynamics on the stimulus feature manifold (Eq. S49), in order to analyze the circuit sampling performance. Since our derivation found the Hamiltonian sampling SOM neurons don't receive feedforward input, i.e., $w_{SF} = 0$, we set $g_{SF} = 0$ in Eq. S49) and rearrange it into the matrix form,

$$\begin{pmatrix} \dot{z}_E \\ \dot{z}_S \end{pmatrix} = -\underbrace{\begin{pmatrix} \tau_E^{-1}(g_{ES} + g_{EF}) & -\tau_E^{-1}g_{ES} \\ -\tau_S^{-1}g_{SE} & \tau_S^{-1}g_{SE} \end{pmatrix}}_{\mathbf{M}},$$
$$+ \underbrace{\begin{pmatrix} \tau_E^{-1}g_{EF} \\ \tau_S^{-1}g_{SF} \end{pmatrix} + \begin{pmatrix} \mu_z + \sigma_E\tau_E^{-1/2}\eta_t \\ 0 \end{pmatrix}}_{\boldsymbol{\mu}} \tag{S68}$$

The eigenvalue with the smallest real-part limits the sampling speed, which is calculated as

$$\lambda_- = \Big[\mathrm{tr}(\mathbf{M}) - \sqrt{\mathrm{tr}(\mathbf{M})^2 - 4\det(\mathbf{M})}\Big]/2 \tag{S69}$$

with the trace $\mathrm{tr}(\mathbf{M})$

$$\mathrm{tr}(\mathbf{M}) = \tau_E^{-1}(g_{ES} + g_{EF}) + \tau_S^{-1}g_{SE}, \tag{S70}$$

and determinant $\det(\mathbf{M})$

$$\det(\mathbf{M}) = \tau_E^{-1}\tau_S^{-1}g_{SE}g_{EF}. \tag{S71}$$

We next analyze how PV and SOM neurons affect the sampling performance respectively.

13

PV neurons

We first consider the reduced circuit model without SOM neurons (setting $g_{ES} = g_{SE} = 0$), and then the dynamics of z_E and z_S are disconnected. The z_E dynamics reduces to

$$
\begin{aligned}
\dot{z}_E &= \tau_E^{-1} g_{EF}(\mu_z - z_E) + \sigma_E \sqrt{\tau_E^{-1}} \xi, \\
&= -\tau_E^{-1} g_{EF} z_E + \left[\tau_E^{-1} g_{EF} \mu_z + \sigma_E \sqrt{\tau_E^{-1}} \xi \right],
\end{aligned}
\tag{S72}
$$

whose eigenvalue is

$$
\lambda = \tau_E^{-1} g_{EF} = \frac{\rho R_F}{\sqrt{2}\tau} \frac{1}{U_E(w_{EP})}.
\tag{S73}
$$

The U_E refers to the peak value of the population synaptic input $\mathbf{u}_E(\theta)$, and is a decreasing function with the inhibitory strength from PV to E neurons w_{EP} (Eq. 4). Therefore, larger inhibition from PV to E neurons will increase the eigenvalue of the z_E sampling dynamics, and hence increase the sampling speed, i.e., the z_E dynamics converges faster (Eq. S72).

SOM neurons

We next investigate how the inhibitory feedback from SOM to E neurons will affect the sampling performance. To obtain theoretical insight, we firstly consider a simple case that the w_{ES} is small enough and its increment doesn't cause a significant change in population response height, i.e., U_X and R_X, otherwise $\tau_X \propto \tau U_X$ in Eq. (S69) will change. In the simplified case, increasing w_{ES} will only change the g_{ES} occurring in the $\mathrm{tr}(\mathbf{M})$ in the slowest eigenvalue λ_-. In this simple case, we find increasing g_{ES} may non-monotonically change the real part of λ_-, depending on whether λ_- has an imaginary part or not.

a). **Pure real λ_-.**
For pure real λ_-, it means the $g_{ES} \propto w_{ES}$ is not large enough and in the weak regime of w_{ES}. In this regime, $\mathrm{Re}(\lambda_-)$ increases with stronger (more negative) w_{ES}.

b). λ_- **has an imaginary part** .
Increasing w_{ES} will lead to the imaginary part of λ_-, and the real part of λ_- is,

$$
\mathrm{Re}(\lambda_-) = \mathrm{tr(M)} = \tau_E^{-1}(g_{ES} + g_{EF}) + \tau_S^{-1} g_{SE},
\tag{S74}
$$

which will decrease with stronger (more negative) w_{ES}.

Figs. 4C plots the real and imaginary parts of λ_- with w_{ES}. We see the $\mathrm{Re}(\lambda_-)$ increases with w_{ES} if λ_- doesn't have an imaginary part. However, when w_{ES} is large enough to induce imaginary λ_-, both the $\mathrm{Re}(\lambda_-)$ and $\mathrm{Im}(\lambda_-)$ decay.

5 Bivariate stimulus posterior sampling in coupled recurrent circuit

We extend our recurrent circuit model above to sample bivariate stimulus posterior. We consider two coupled recurrent circuits, with each the same as Fig. 2C. Each circuit m will receive a feedforward input I_m generated by latent stimulus s_m, and will sample the stimulus s_m. For simplicity, we consider only E neurons across circuits to be coupled together. For coupled networks,

$$
\tau \frac{\partial \mathbf{u}_m^E(\theta, t)}{\partial t} = -\mathbf{u}_m^E(\theta, t) + \rho \sum_n (\mathbf{W}_{mn}^E * \mathbf{r}_n^E)(\theta, t) + \rho \mathbf{W}_m^{ES} * \mathbf{r}_m^S,
$$

$$
+ \rho \mathbf{W}_m^{EF} * \mathbf{r}_m^F + \sqrt{\tau F_m^E [u_m^E(\theta, t)]} \xi_m(\theta, t)
$$

$$
\tau \frac{\partial \mathbf{u}_m^S(\theta, t)}{\partial t} = -\mathbf{u}_m^S(\theta, t) + \rho \sum_{X=E,F} (\mathbf{W}_m^{SX} * \mathbf{r}_X)(\theta, t); \quad \mathbf{r}_S(\theta, t) = g_S \cdot [\mathbf{u}_S(\theta, t)]_+,
$$

where $\mathbf{u}_{Em}(\theta, t)$ and $\mathbf{r}_{Em}(\theta, t)$ represent the synaptic input and firing rate, respectively of each each network receiving independent input, z_{Xm} and has its own population of SOM interneurons. When $m = n$, W_{Emm} is the recurrent connection kernel within the same network; whereas $m \neq n$, W_{Emn} is the connection kernel between neurons from network n to network m. In the case of the coupled network, $m = 1$, and $n = 2$ since there are only two networks.

14

5.1 The coupled circuit dynamics on the stimulus manifold

As described in Sec. 3.3 and the main text, the eigenvector with the largest eigenvalue dominating the circuit dynamics along the stimulus feature manifold of the neuronal population is,

$$\phi(\theta|z_m^X) \propto \frac{d\langle U_m^X|\theta\rangle}{dz_m^X} = (\theta - z_m^X)\exp{-\frac{(\theta - z_m^X)^2}{4(a^X)^2}} \tag{S75}$$

Projecting the circuit dynamics onto the position mode(eigenfunction defined in (S75)), we get the bump position dynamics which has an extra term compared with the Eq. (S47)

$$\tau U_1^E \frac{dz_1^E}{dt} = \frac{\rho}{\sqrt{2}}\Big[w_{12}^{EE}R_2^E(z_2^E - z_1^E) + w_{11}^{ES}R_1^S\frac{a^S}{a^E}(z_1^S - z_1^E) + w^{EF}R_1^F(\mu_1 - z_1^E)\Big],$$
$$+ \sqrt{\frac{8a^E F^E}{3\sqrt{3\pi}}}\sqrt{\tau_L U_1^E}\xi_1^E(\theta, t), \tag{S76}$$

where the 1st term on the right-hand side denotes the effect of the input from another circuit.

Since the SOM populations are not coupled together, the substitution and projection onto the eigenvector remains to same as (S46).

$$\tau_E U_1^E \frac{dz_1^E}{dt} = \frac{\rho}{\sqrt{2}}\Big[w_{12}^{EE}R_2^E(z_2^E - z_1^E) + w_1^{ES}R_1^S\frac{a^S}{a^E}(z_1^S - z_1^E) + w_1^{EF}R_1^F(\mu_1 - z_1^E)\Big]$$
$$+ \sqrt{\frac{8F_1^E a^E}{3\sqrt{3\pi}}}\sqrt{\tau U_1^E}\eta_t \tag{S77}$$
$$\tau_S U_1^S \frac{dz_1^S}{dt} = \frac{\rho}{\sqrt{2}}\Big[w_1^{SE}R_1^E\frac{a^E}{a^S}(z_1^E - z_1^S) + w_1^{SF}R_1^F(\mu_1 - z_1^S)\Big]$$

We define the following notations,

$$\tau_m^X = \tau U_m^X, \quad g_{mn}^{XY} = \frac{\rho a^Y}{\sqrt{2}a^X}w_{mn}^{XY}R_m^Y, \quad \sigma_E^2 = \frac{8a^E F^E}{3\sqrt{3\pi}}, \tag{S78}$$

We simplify (S77) as,

$$\tau_1^E \dot{z}_1^E = g_{12}^{EE}(z_2^E - z_1^E) + g_{11}^{ES}(z_1^S - z_1^E) + g_{11}^{EF}(\mu_1 - z_1^E) + \sigma_E\sqrt{\tau_1^E}\xi_t$$
$$\tau_1^S \dot{z}_1^S = g_{11}^{SE}(z_1^E - z_1^S) + g_{11}^{SF}(\mu_1 - z_1^S). \tag{S79}$$

By changing the subscripts denoting the circuit index, we could obtain the bump position dynamics for circuit 2. Combining the position dynamics in both circuits, we can denote the dynamics by using matrix form,

$$\dot{\mathbf{z}}^E = (\mathbf{D}_\tau^E)^{-1}\Big[-(\mathbf{G}^{EE} + \mathbf{D}^{EF})\mathbf{z}^E + \mathbf{D}^{EF}\boldsymbol{\mu} + \mathbf{D}^{ES}(\mathbf{z}^S - \mathbf{z}^E)\Big] + \sigma_E(\mathbf{D}_\tau^E)^{-1/2}\boldsymbol{\xi}_t$$
$$\dot{\mathbf{z}}^S = (\mathbf{D}_\tau^S)^{-1}\Big[\mathbf{D}^{SE}(\mathbf{z}^E - \mathbf{z}^S) + \mathbf{D}^{SF}(\boldsymbol{\mu} - \mathbf{z}^S)\Big], \tag{S80}$$

where

$$\mathbf{z}^E = (z_1^E, z_2^E), \quad \mathbf{z}^S = (z_1^S, z_2^S), \quad \boldsymbol{\mu} = (\mu_1, \mu_2). \tag{S81}$$

$$\mathbf{G}^{EE} = \begin{pmatrix} g_{12}^{EE} & -g_{12}^{EE} \\ -g_{21}^{EE} & g_{21}^{EE} \end{pmatrix}, \; \mathbf{D}_\tau^X = \begin{pmatrix} \tau_1^X & 0 \\ 0 & \tau_1^X \end{pmatrix}, \; \mathbf{D}^{XY} = \begin{pmatrix} g_{11}^{XY} & 0 \\ 0 & g_{22}^{XY} \end{pmatrix}, \; X, Y \in \{E, S, F\} \tag{S82}$$

15

5.2 Identifying the bivariate stimulus prior

To investigate the stimulus prior stored in the coupled circuits, we consider a simple case that there are no SOM neurons in both circuits. Here we assume the SOM will not change the equilibrium distribution but only speed up sampling, as a conclusion from the sampling in a single recurrent circuit. In this case, the $\mathbf{z}^E$ dynamics reduces to (setting $\mathbf{D}^{ES} = 0$),

$$\dot{\mathbf{z}}^E = (\mathbf{D}^E_\tau)^{-1}\big[-(\mathbf{G}^{EE} + \mathbf{D}^{EF})\mathbf{z}^E + \mathbf{D}^{EF}\boldsymbol{\mu}\big] + \sigma_E(\mathbf{D}^E_\tau)^{-1/2}\boldsymbol{\xi}_t \tag{S83}$$

The drift term can be related to the gradient of posteriors, i.e.,

$$\nabla \ln p(\mathbf{z}|\mathbf{r}_F) = \nabla \ln p(\mathbf{r}_F|\mathbf{z}) + \nabla \ln p(\mathbf{z})$$
$$= \mathbf{D}^{EF}(\boldsymbol{\mu} - \mathbf{z}^E) - \mathbf{G}^{EE}\mathbf{z}^E \tag{S84}$$

Similar to the 1D case, we regard the $\mathbf{D}^{EF}(\boldsymbol{\mu}-\mathbf{z}^E)$ as the gradient of the likelihood, i.e., $\nabla \ln p(\mathbf{r}_F|\mathbf{z})$. Then the $-\mathbf{G}^{EE}\mathbf{z}^E$ will be treated as the gradient of the prior, $\nabla \ln p(\mathbf{z})$. For simplicity, considering the interaction strength of two circuits' samples are symmetric, i.e., $g^{EE}_{12} = g^{EE}_{21} \equiv \Lambda_s$, then the prior is

$$p(z_1, z_2) \propto \exp\left(-\frac{1}{2}\mathbf{z}^\top \boldsymbol{\Lambda}_s \mathbf{z}\right) = \exp[-\Lambda_s(z_1 - z_2)^2/2], \quad \boldsymbol{\Lambda}_s = \Lambda_s \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \tag{S85}$$

We see the coupled circuits store an associative prior of two stimuli which stores the co-occurrence probability of the two (Fig. 5C). Moreover, the marginal prior of each stimulus is still uniform.

6 Simulation details and model parameters

6.1 Network parameters and simulation

Table S1 includes typical parameters used in our network simulation. Each network includes $N_E = 180$ excitatory neurons, $N_S = 180$ SOM interneurons, both of which are uniformly distributed in the stimulus feature space $z \in (-180°, 180°]$. The neuronal density is $\rho = N/w_z$ where $w_z = 360$ is the width for stimulus feature space. All other parameters can be found in Table S1.

In particular, we use the critical E-to-E recurrent weight w_c to scale the weight in the circuit model. w_c is the minimal value of w_{EE} for holding a non-zero population response without feedforward input and the inhibition from SOM, and can be calculated by solving a non-zero U_E when setting $R_F = w_{ES} = 0$ (see SI Sec. 3.2).

$$w_c = 2\sqrt{2}(2\pi)^{1/4}\sqrt{ka\rho} \approx 0.896. \tag{S86}$$

Then, the feedforward input intensity is scaled by the peak value of E population synaptic input, U_c, that is self-sustained by the E neurons with recurrent weight w_c without receiving feedforward input and SOM's inhibition,

$$U_c = \frac{w_c}{2\sqrt{\pi}ka} \tag{S87}$$

Changes to default parameters in Fig.5 are included in the Table S2. For specialized parameters in Fig. S2 refer to Table S3.

The network dynamics were simulated using Eulers method and a time step of 0.01τ. The empirical distribution of stimulus samples is estimated from running the model for 500τ, and the responses and samples in the first 10τ are discarded to exclude the effect from non-equilibrium responses.

The circuit model was simulated on an Asus ROG Zephyrus laptop which has an i7 intel core and 32 RAM. Most simulations take 3.2 seconds. Parameter scans were run on a HPC 512 GB RAM computing cluster using 36 parallel jobs taking approximately six minutes.

6.2 Read out stimulus samples from the population responses

We use a linear decoder called the population vector to read out the instantaneous stimulus sample, z_E, from the E neuron population responses,

$$z_E(t) = \frac{\sum_j \mathbf{r}_E(\theta_j, t)\theta_j}{\sum_j \mathbf{r}_E(\theta_j, t)} \tag{S88}$$

16

Table S1: Default network parameters

Parameter	Variable	Value
Time constant	τ	1
Feedforward weight	w_{EF}	$1.3w_c$
E to SOM weight	w_{SE}	$0.5w_c$
E to E weight	w_{EE}	$0.5w_c$
PV to E weight	w_{PV}	0.0005
SOM to E weight	w_{ES}	$0.6w_c$
Connection width	a_E	40°
Inhibitory gain	g_s	10
Feedforward input intensity	R_F	$0.8U_c$
Feedforward input location	z	0
Fano factor of injected variability	F_E	0.5

Table S2: Fig. 5 network parameters

Parameter	Variable	Value
Feedforward weight	w_{EF}	$1.7w_c$
E to SOM weight	w_{SE}	$0.5w_c$
E to E weight	w_{EE}	$0.5w_c$
SOM to E weight	w_{ES}	$0.7w_c$
E1 to E2 weight	$w_{12} = w_{21}$	$0.2w_c$
Connection width	a_E	40°
SOM connection width	a_S	37.4°
E to SOM connection width	a_{SE}	34.6°
SOM to E connection width	a_{ES}	20°
SOM Time constant	τ_S	$5\tau_E$

Then the sampling distribution generated by the circuit model is defined as the empirical distribution of the stimulus samples

$$p(z) = \sum_t \delta(z - z_E(t)) \tag{S89}$$

The sample z_S is read out similarly from SOM neurons, which contributes to the auxiliary variable y in the Hamiltonian sampling (Eq. 17).

6.3 Power spectrum analysis

For the power spectrum analysis, the local field potential (LFP) signals can be approximated as the sum of the inhibitory or excitatory synaptic currents. Specifically, we use the sum of the synaptic input from both the E and SOM networks,

$$\text{LFP}(t) = \sum_\theta \mathbf{u}_X(\theta, t) \tag{S90}$$

Then the LFP was bandpass filtered from 5 to 50 Hz with a Butterworth filter. The power spectral density was estimated using the periodogram.

Table S3: Fig. S2 network parameters

Parameter	Variable	Value
E Time constant	τ_E	1
Feedforward weight	w_{EF}	$1.3w_c$
SOM Time constant	τ_S	$5\tau_E$
E to SOM weight	w_{SE}	$0.5w_c$
E to E weight	w_{EE}	$1.0w_c$
SOM to E weight	w_{ES}	$1.1w_c$

17

6.4 Comparing the sampling distributions with posteriors

To determine whether the circuit model can sample the posterior, we calculate the KL divergence to measure the discrepancy from the posterior $p(z|\mathbf{r}_F)$ to the circuit's sampling distribution $p(z) = \sum_t \delta(z - z_E(t))$,

$$D_{KL}[p(z|\mathbf{r}_F)||p(z)] = \int p(z|\mathbf{r}_F) \ln \frac{p(z|\mathbf{r}_F)}{p(z)} dz \tag{S91}$$

where the posterior (or the likelihood since the subjective circuit prior is uniform) is directly read out from the feedforward input (Eq. 7). Since the posterior is a Gaussian distribution, we also parameterize the empirical sampling distribution as a Gaussian, i.e., we numerically estimate the mean and covariance of samples and then calculate the KL divergence.

6.5 Reproducing E neurons' tuning curves from modulating interneurons

Mimicking the experiment measuring the E neurons' tuning curves by perturbing interneurons[7], we also perturb each type of interneurons in the circuit model individually and measure how these perturbations change E neurons' tuning curves. The experiments applied a full-field light to the same type of neuron which can be approximately treated as the neurons of the same type receive the same amount of perturbation [7]. Hence in our simulation, when perturbing one type of neurons, we apple the same offset input to all neurons of that type.

Specifically, when we perturbed the SOM neurons,

$$\tau \frac{\partial \mathbf{u}_S(x,t)}{\partial t} = -\mathbf{u}_S(x,t) + \rho \sum_{X=E,F} (\mathbf{W}_{SX} * \mathbf{r}_X)(\theta, t) + I_S; \tag{S92}$$

where I_S is a constant input applied to every SOM neuron in the circuit model.

Similarly, for perturbing PV neurons, we add another offset input into the divisive normalization (Eq. 4),

$$\mathbf{r}_E(\theta, t) = \frac{[\mathbf{u}_E(\theta, t)]_+^2}{1 + \rho w_{EP} \int ([\mathbf{u}_E(\theta, t)]_+^2 + I_P) d\theta'}, \tag{S93}$$

where I_P is the perturbing input.

Then, with the existence of one of these offset inputs, we change the presented feedforward input location z (Eq. 3) and measure the mean firing rate of an example E neuron.

18

References

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

[2] Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pages 1683–1691. PMLR, 2014.

[3] Xingsi Dong, Zilong Ji, Tianhao Chu, Tiejun Huang, Wenhao Zhang, and Si Wu. Adaptation accelerating sampling-based bayesian inference in attractor neural networks. *Advances in Neural Information Processing Systems*, 35:21534–21547, 2022.

[4] Crispin W Gardiner et al. *Handbook of stochastic methods*, volume 3. springer Berlin, 1985.

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

[6] C. C Alan Fung, K. Y. Michael Wong, and Si Wu. A moving bump in a continuous manifold: A comprehensive study of the tracking dynamics of continuous attractor neural networks. *Neural Computation*, 22(3):752–792, 2010.

[7] Nathan R. Wilson, Caroline A. Runyan, Forea L. Wang, and Mriganka Sur. Division and subtraction by distinct cortical inhibitory networks in vivo. *Nature*, 488(7411):343–348, August 2012.

19