Fast Sampling-Based Inference in Balanced Neuronal Networks

Guillaume Hennequin[1]
gjeh2@cam.ac.uk

Laurence Aitchison[2]
laurence@gatsby.ucl.ac.uk

Máté Lengyel[1]
m.lengyel@eng.cam.ac.uk

[1]Computational & Biological Learning Lab, Dept. of Engineering, University of Cambridge, UK
[2]Gatsby Computational Neuroscience Unit, University College London, UK

Abstract

Multiple lines of evidence support the notion that the brain performs probabilistic inference in multiple cognitive domains, including perception and decision making. There is also evidence that probabilistic inference may be implemented in the brain through the (quasi-)stochastic activity of neural circuits, producing samples from the appropriate posterior distributions, effectively implementing a Markov chain Monte Carlo algorithm. However, time becomes a fundamental bottleneck in such sampling-based probabilistic representations: the quality of inferences depends on how fast the neural circuit generates new, uncorrelated samples from its stationary distribution (the posterior). We explore this bottleneck in a simple, linear-Gaussian latent variable model, in which posterior sampling can be achieved by stochastic neural networks with linear dynamics. The well-known Langevin sampling (LS) recipe, so far the only sampling algorithm for continuous variables of which a neural implementation has been suggested, naturally fits into this dynamical framework. However, we first show analytically and through simulations that the symmetry of the synaptic weight matrix implied by LS yields critically slow mixing when the posterior is high-dimensional. Next, using methods from control theory, we construct and inspect networks that are optimally fast, and hence orders of magnitude faster than LS, while being far more biologically plausible. In these networks, strong – but transient – selective amplification of external noise generates the spatially correlated activity fluctuations prescribed by the posterior. Intriguingly, although a detailed balance of excitation and inhibition is dynamically maintained, detailed balance of Markov chain steps in the resulting sampler is violated, consistent with recent findings on how statistical irreversibility can overcome the speed limitation of random walks in other domains.

1   Introduction

The high speed of human sensory perception [1] is puzzling given its inherent computational complexity: sensory inputs are noisy and ambiguous, and therefore do not uniquely determine the state of the environment for the observer, which makes perception akin to a statistical inference problem. Thus, the brain must represent and compute with complex and often high-dimensional probability distributions over relevant environmental variables. Most state-of-the-art machine learning techniques for large scale inference trade inference accuracy for computing speed (e.g. [2]). The brain, on the contrary, seems to enjoy both simultaneously [3].

Some probabilistic computations can be made easier through an appropriate choice of representation for the probability distributions of interest. Sampling-based representations used in Monte Carlo

techniques, for example, make computing moments of the distribution or its marginals straightforward. Indeed, recent behavioural and neurophysiological evidence suggests that the brain uses such sampling-based representations by neural circuit dynamics implementing a Markov chain Monte Carlo (MCMC) algorithm such that their trajectories in state space produce sequential samples from the appropriate posterior distribution [4, 5, 6].

However, for sampling-based representations, speed becomes a key bottleneck: computations involving the posterior distribution become accurate only after enough samples have been collected, and one has no choice but *to wait* for those samples to be delivered by the circuit dynamics. For sampling to be of any practical use, the interval that separates the generation of two independent samples must be short relative to the desired behavioral timescale. Single neurons can integrate their inputs on a timescale $\tau_{\mathrm{m}} \approx 10 - 50$ ms, whereas we must often make decisions in less than a second: this leaves just enough time to use (i.e. read out) a few tens of samples. What kinds of neural circuit dynamics are capable of producing *uncorrelated* samples at $\sim$100 Hz remains unclear.

Here, we introduce a simple yet non-trivial generative model and seek plausible neuronal network dynamics for *fast* sampling from the corresponding posterior distribution. While some standard machine learning techniques such as Langevin or Gibbs sampling do suggest "neural network"-type solutions to sampling, not only are the corresponding architectures implausible in fundamental ways (e.g. they violate Dale's law), but we show here that they lead to unacceptably slow mixing in high dimensions. Although the issue of sampling speed in general is well appreciated in the context of machine learning, there have been no systematic approaches to tackle it owing to a large part to the fact that sampling speed can only be evaluated empirically in most cases. In contrast, the simplicity of our generative model allowed us to draw an analytical picture of the problem which in turn suggested a systematic approach for solving it. Specifically, we used methods from robust control to discover the *fastest* neural-like sampler for our generative model, and to study its structure. We find that it corresponds to greatly non-symmetric synaptic interactions (leading to statistical irreversibility), and mathematically nonnormal[1] circuit dynamics [7, 8] in close analogy with the dynamical regime in which the cortex has been suggested to operate [9].

2   Linear networks perform sampling under a linear Gaussian model

We focus on a linear Gaussian latent variable model which generates observations $\mathbf{h} \in \mathbb{R}^M$ as weighted sums of N features $\mathbf{A} \equiv (\mathbf{a}_1; \dots; \mathbf{a}_N) \in \mathbb{R}^{M \times N}$ with jointly Gaussian coefficients $\mathbf{r} \in \mathbb{R}^N$, plus independent additive noise terms (Fig. 1, left). More formally:

$$p(\mathbf{r}) = \mathcal{N}(\mathbf{r}; 0, \mathbf{C}) \qquad \text{and} \qquad p(\mathbf{h}|\mathbf{r}) = \mathcal{N}\left(\mathbf{h}; \mathbf{Ar}, \sigma_h^2 \mathbf{I}\right) \tag{1}$$

where $\mathbf{I}$ denotes the identity matrix. The posterior distribution is multivariate Gaussian, $p(\mathbf{r}|\mathbf{h}) = \mathcal{N}(\mathbf{r}; \boldsymbol{\mu}(\mathbf{h}), \boldsymbol{\Sigma})$, with

$$\boldsymbol{\Sigma} = \left(\mathbf{C}^{-1} + \mathbf{A}^\top \mathbf{A}/\sigma_h^2\right)^{-1} \qquad \text{and} \qquad \boldsymbol{\mu}(\mathbf{h}) = \boldsymbol{\Sigma}\mathbf{A}^\top \mathbf{h}/\sigma_h^2. \tag{2}$$

where we made explicit the fact that under this simple model, only the mean, $\boldsymbol{\mu}(\mathbf{h})$, but not the covariance of the posterior, $\boldsymbol{\Sigma}$, depends on the input, $\mathbf{h}$.

We are interested in neural circuit dynamics for sampling from $p(\mathbf{r}|\mathbf{h})$, whereby the data (observation) $\mathbf{h}$ is given as a constant feedforward input to a population of recurrently connected neurons, each of which encodes one of the latent variables and also receives inputs from an external, private source of noise $\boldsymbol{\xi}$ (Fig. 1, right). Our goal is to devise a network such that the activity fluctuations $\mathbf{r}(t)$ in the recurrent layer have a stationary distribution that matches the posterior, *for any* $\mathbf{h}$.

Specifically, we consider linear recurrent stochastic dynamics of the form:

$$d\mathbf{r} = \frac{dt}{\tau_{\mathrm{m}}} \left[-\mathbf{r}(t) + \mathbf{Wr}(t) + \mathbf{Fh}\right] + \sigma_\xi \sqrt{\frac{2}{\tau_{\mathrm{m}}}} \, d\boldsymbol{\xi}(t) \tag{3}$$

where $\tau_{\mathrm{m}} = 20$ ms is the single-unit "membrane" time constant, and $d\boldsymbol{\xi}$ is a Wiener process of unit variance, which is scaled by a scalar noise intensity σ_ξ. The activity $r_i(t)$ could represent either the

---

[1]"Nonnormal" should not be confused with "non-Gaussian": a matrix $\mathbf{M}$ is nonnormal iff $\mathbf{MM}^\top \neq \mathbf{M}^\top \mathbf{M}$.
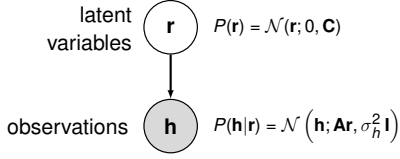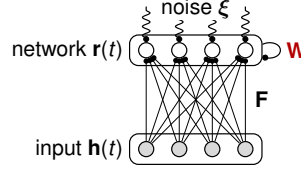
Figure 1: **Sampling under a linear Gaussian latent variable model using neuronal network dynamics.** Left: schematics of the generative model. Right: schematics of the recognition model. See text for details.

membrane potential of neuron i, or the deviation of its momentary firing rate from a baseline. The matrices $\mathbf{F}$ and $\mathbf{W}$ contain the feedforward and recurrent connection weights, respectively.

The stationary distribution of $\mathbf{r}$ is indeed Gaussian with a mean $\boldsymbol{\mu}^{\mathbf{r}}(\mathbf{h}) = (\mathbf{I} - \mathbf{W})^{-1}\mathbf{F}\mathbf{h}$ and a co-variance matrix $\boldsymbol{\Sigma}^{\mathbf{r}} \equiv \left\langle (\mathbf{r}(t) - \boldsymbol{\mu}^{\mathbf{r}})(\mathbf{r}(t) - \boldsymbol{\mu}^{\mathbf{r}})^{\top} \right\rangle_t$. For the following, we will use the dependence of $\boldsymbol{\Sigma}^{\mathbf{r}}$ on $\mathbf{W}$ (and σ_ξ) given implicitly by the following Lyapunov equation [10]:

$$(\mathbf{W} - \mathbf{I})\boldsymbol{\Sigma}^{\mathbf{r}} + \boldsymbol{\Sigma}^{\mathbf{r}}(\mathbf{W} - \mathbf{I})^{\top} = -2\sigma_\xi^2 \mathbf{I} \tag{4}$$

Note that in the absence of recurrent connectivity ($\mathbf{W} = 0$), the variance of every $r_i(t)$ would be exactly σ_ξ^2. Note also that, just as required (see above), only the mean, $\boldsymbol{\mu}^{\mathbf{r}}(\mathbf{h})$, but not the covariance, $\boldsymbol{\Sigma}^{\mathbf{r}}$, depends on the input, $\mathbf{h}$.

In order for the dynamics of Eq. 3 to sample from the correct posteriors, we must choose $\mathbf{F}$, $\mathbf{W}$ and σ_ξ such that $\boldsymbol{\mu}^{\mathbf{r}}(\mathbf{h}) = \boldsymbol{\mu}(\mathbf{h})$ for any $\mathbf{h}$, and $\boldsymbol{\Sigma}^{\mathbf{r}} = \boldsymbol{\Sigma}$. One possible solution (which, importantly, is not unique, as we show later) is

$$\mathbf{F} = (\sigma_\xi/\sigma_h)^2 \, \mathbf{A}^{\top} \qquad \text{and} \qquad \mathbf{W} = \mathbf{W}_{\mathrm{L}} \equiv \mathbf{I} - \sigma_\xi^2 \, \boldsymbol{\Sigma}^{-1} \tag{5}$$

with arbitrary $\sigma_\xi > 0$.

In the following, we will be interested in the likelihood matrix $\mathbf{A}$ only insofar as it affects the posterior covariance matrix $\boldsymbol{\Sigma}$, which turns out to be the main determinant of sampling speed. We will therefore directly choose some covariance matrix $\boldsymbol{\Sigma}$, and set $\mathbf{h} = 0$ without loss of generality.

3 Langevin sampling is very slow

Langevin sampling (LS) is a common sampling technique [2, 11, 12], and in fact the only one that has been proposed to be neurally implemented for continuous variables [6, 13]. According to LS, a stochastic dynamical system performs "noisy gradient ascent of the log posterior":

$$d\mathbf{r} = \frac{1}{2}\frac{\partial}{\partial \mathbf{r}} \log p(\mathbf{r}|\mathbf{h}) \, dt + d\boldsymbol{\xi} \tag{6}$$

where $d\boldsymbol{\xi}$ is a unitary Wiener process. When $\mathbf{r}|\mathbf{h}$ is Gaussian, Eq. 6 reduces to Eq. 3 for $\sigma_\xi = 1$ and the choice of $\mathbf{F}$ and $\mathbf{W}$ given in Eq. 5 – hence the notation $\mathbf{W}_{\mathrm{L}}$ above. Note that $\mathbf{W}_{\mathrm{L}}$ is symmetric.

As we show now, this choice of weight matrix leads to critically slow mixing (i.e. very long auto-correlation time scales in $\mathbf{r}(t)$) when N is large. In a linear network, the average autocorrelation length is dominated by the decay time constant $\tau_{\max}$ of the slowest eigenmode, i.e. the eigenvector of $(\mathbf{W} - \mathbf{I})$ associated with the eigenvalue $\lambda_{\max}^{\mathbf{W}-\mathbf{I}}$ which, of all the eigenvalues of $(\mathbf{W} - \mathbf{I})$, has the largest real part (which must still be negative, to ensure stability). The contribution of the slowest eigenmode to the sample autocorrelation time is $\tau_{\max} = -\tau_{\mathrm{m}}/\mathrm{Re}\left(\lambda_{\max}^{\mathbf{W}-\mathbf{I}}\right)$, so sampling becomes very slow when $\mathrm{Re}\left(\lambda_{\max}^{\mathbf{W}-\mathbf{I}}\right)$ approaches 0. This is, in fact, what happens with LS as $N \to \infty$. Indeed, we could derive the following generic lower bound (details can be found in our Supplementary Information, SI):

$$\lambda_{\max}^{\mathbf{W}_{\mathrm{L}}-\mathbf{I}} \geq \frac{-(\sigma_\xi/\sigma_0)^2}{\sqrt{1 + N\sigma_r^2}} \tag{7}$$

which is shown as dashed lines in Fig. 2. Thus, LS becomes infinitely slow in the large N limit when pairwise correlations do not vanish in that limit (or at least not as fast as $N^{-\frac{1}{2}}$ in their std.).

Slowing becomes even worse when $\boldsymbol{\Sigma}$ is drawn from the inverse Wishart distribution with ν degrees of freedom and scale matrix $\omega^{-2}\mathbf{I}$ (Fig. 2). We choose $\nu = N-1+\lfloor\sigma_r^{-2}\rfloor$ and $\omega^{-2} = \sigma_0^2(\nu-N-1)$
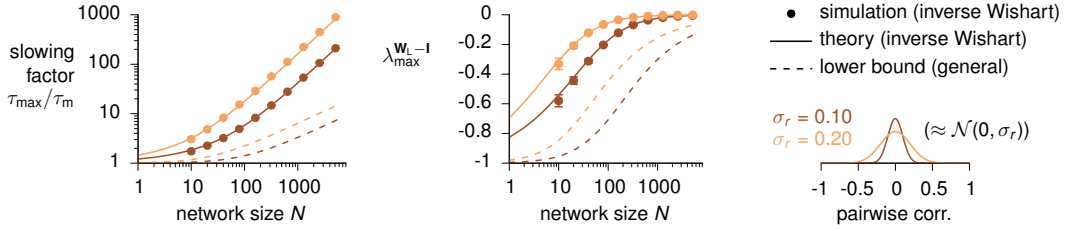
Figure 2: **Langevin sampling (LS) is slow in high-dimension**. Random covariance matrices $\mathbf{\Sigma}$ of size N are drawn from an inverse Wishart distribution with parameters chosen such that the average diagonal element (variance) is $\sigma_0^2 = 1$ and the distribution of pairwise correlations has zero mean and variance σ_r^2 (right). Sampling from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ using a stochastic neural network (cf. Fig. 1) with $\mathbf{W} = \mathbf{W}_\mathrm{L}$ (LS, symmetric solution) becomes increasingly slow as N grows, as indicated by the relative decay time constant $\tau_\mathrm{max}/\tau_\mathrm{m}$ of the slowest eigenmode of $(\mathbf{W}_\mathrm{L} - \mathbf{I})$ (left), which is also the negative inverse of its largest eigenvalue (middle). Dots indicate the numerical evaluation of the corresponding quantities, and errorbars (barely noticeable) denote standard deviation across several random realizations of $\mathbf{\Sigma}$. Dashed lines correspond to the generic bound in Eq. 7. Solid lines are obtained from random matrix theory under the asssumption that $\mathbf{\Sigma}$ is drawn from an inverse Wishart distribution (Eq. 8). Parameters: $\sigma_\xi = \sigma_0 = 1$.

such that the expected value of a diagonal element (variance) in $\mathbf{\Sigma}$ is σ_0^2, and the distribution of pairwise correlations is centered on zero with variance σ_r^2. The asymptotic behavior of the largest eigenvalue of $\mathbf{\Sigma}^{-1}$ (the square of the smallest singular value of a random $\nu \times N$ rectangular matrix) is known from random matrix theory (e.g. [14]), and we have for large N:

$$\lambda_\mathrm{max}^{\mathbf{W}_\mathrm{L}-\mathbf{I}} \approx -\frac{(\sigma_\xi/\sigma_0)^2}{\lfloor \sigma_r^{-2} \rfloor - 2}\left(\sqrt{N - 1 + \lfloor \sigma_r^{-2} \rfloor} - \sqrt{N}\right)^2 \qquad \sim -\mathcal{O}\left(\frac{1}{N}\right) \tag{8}$$

This scaling behavior is shown in Fig. 2 (solid lines). In fact, we can also show (cf. SI) that LS is (locally) the slowest possible choice (see Sec. 4 below for a precise definition of "slowest", and SI for details).

Note that both Eqs. 7-8 are inversely proportional to the ratio (σ_0/σ_ξ), which tells us how much the recurrent interactions must amplify the external noise in order to produce samples from the right stationary activity distribution. The more amplification is required ($\sigma_0 \gg \sigma_\xi$), the slower the dynamics of LS. Conversely, one could potentially make Langevin sampling faster by increasing σ_ξ, but σ_ξ would need to scale as $\sqrt{N}$ to annihilate the critical slowing problem. This – in itself – is unrealistic; moreover, it would also require the resulting connectivity matrix to have a large negative diagonal ($\mathcal{O}(-N)$) – ie. the intrinsic neuronal time constant τ_m to scale as $\mathcal{O}(1/N)$ –, which is perhaps even more unrealistic.[2]

Note also that LS can be sped up by appropriate "preconditioning" (e.g. [15, 16]), for example using the inverse Hessian of the log-posterior. In our case, a simple calculation shows that this corresponds to removing all recurrent connections, and pushing the posterior covariance matrix to the external noise sources, which is only postponing the problem to some other brain network.

Finally, LS is fundamentally implausible as a neuronal implementation: it imposes symmetric synaptic interactions, which is simply not possible in the brain due to the existence of distinct classes of excitatory and inhibitory neurons (Dale's principle). In the following section, we show that networks can be constructed that overcome all the above limitations of LS in a principled way.

4   General solution and quantification of sampling speed

While Langevin dynamics (Eq. 6) provide a general recipe for sampling from any given posterior density, they unduly constrain the recurrent interactions to be symmetric – at least in the Gaussian

---

[2]From a pure machine learning perspective, increasing σ_ξ is not an option either: the increasing stiffness of Eq. 6 would either require the use of a very small integration step, or would lead to arbitrarily small acceptance ratios in the context of Metropolis-Hastings proposals.
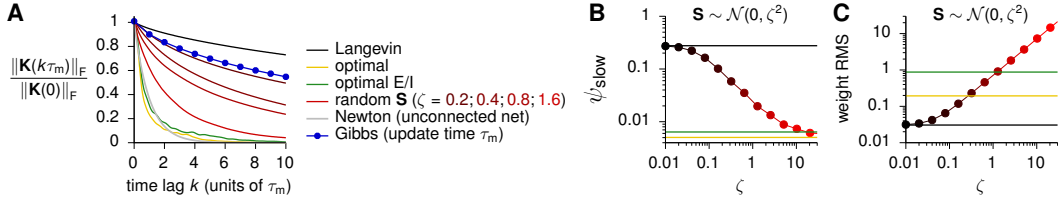
Figure 3: **How fast is the fastest sampler?** (**A**) Scalar measure of the statistical dependency between any two samples collected $k\tau_{\mathrm{m}}$ seconds apart (cf. main text), for Langevin sampling (black), Gibbs sampling (blue, assuming a full update sweep is done every τ_{m}), a series of networks (brown to red) with connectivities given by Eq. 9 where the elements of the skew-symmetric matrix $\mathbf{S}$ were drawn iid. from $\mathcal{N}(0, \zeta^2)$ for different values of ζ (see also panel B), the unconstrained optimized network (yellow), and the optimized E/I network (green). For reference, the dashed gray line shows the behavior of a network in which there are no recurrent interactions, and the posterior covariance is encoded in the covariance of the input noise, which in fact corresponds to Langevin sampling with inverse Hessian ("Newton"-like) preconditioning [16]. (**B**) Total slowing cost $\psi_{\mathrm{slow}}(\mathbf{S})$ when $S_{i<j} \sim \mathcal{N}(0, \zeta^2)$, for increasing values of ζ. The Langevin and the two optimized networks are shown as horizontal lines for comparison. (**C**) Same as in (B), showing the root mean square (RMS) value of the synaptic weights. Parameter values: $N = 200$, $N_I = 100$, $\sigma_\xi = 1$, $\tau_{\mathrm{m}} = 20$ ms.

case. To see why this is a drastic restriction, let us observe that any connectivity matrix of the form

$$\mathbf{W}(\mathbf{S}) = \mathbf{I} + \left(-\sigma_\xi^2 \mathbf{I} + \mathbf{S}\right) \boldsymbol{\Sigma}^{-1} \tag{9}$$

where $\mathbf{S}$ is an arbitrary skew-symmetric matrix ($\mathbf{S}^\top = -\mathbf{S}$), solves Eq. 4, and therefore induces the correct stationary distribution $\mathcal{N}(\cdot, \boldsymbol{\Sigma})$ under the linear stochastic dynamics of Eq. 3. Note that Langevin sampling corresponds to $\mathbf{S} = 0$ (cf. Eq. 5). In general, though, there are $\mathcal{O}(N^2)$ degrees of freedom in the skew-symmetric matrix $\mathbf{S}$, which could perhaps be exploited to increase the mixing rate. In Sec. 5, we will show that indeed a large gain in sampling speed can be obtained through an appropriate choice of $\mathbf{S}$. For now, let us quantify slowness.

Let $\boldsymbol{\Lambda} \equiv \mathrm{diag}\left(\boldsymbol{\Sigma}\right)$ be the diagonal matrix that contains all the posterior variances, and $\mathbf{K}(\mathbf{S}, \tau) \equiv \left\langle (\mathbf{r}(t+\tau) - \mu)(\mathbf{r}(t) - \mu)^\top \right\rangle_t$ be the matrix of lagged covariances among neurons under the stationary distribution of the dynamics (so that $\boldsymbol{\Lambda}^{-\frac{1}{2}} \mathbf{K}(\mathbf{S}, \tau) \boldsymbol{\Lambda}^{-\frac{1}{2}}$ is the autocorrelation matrix of the network). Note that $\mathbf{K}(\mathbf{S}, 0) = \boldsymbol{\Sigma}$ is the posterior covariance matrix, and that for fixed $\boldsymbol{\Sigma}$, σ_ξ^2 and τ_{m}, $\mathbf{K}(\mathbf{S}, \tau)$ depends only on the lag τ and on the matrix of recurrent weights $\mathbf{W}$, which itself depends only on the skew-symmetric matrix $\mathbf{S}$ of free parameters. We then define a "total slowing cost"

$$\psi_{\mathrm{slow}}(\mathbf{S}) = \frac{1}{2\tau_{\mathrm{m}} N^2} \int_0^\infty \left\| \boldsymbol{\Lambda}^{-\frac{1}{2}} \mathbf{K}(\mathbf{S}, \tau) \boldsymbol{\Lambda}^{-\frac{1}{2}} \right\|_{\mathrm{F}}^2 \, \mathrm{d}\tau \tag{10}$$

which penalizes the magnitude of the temporal (normalized) autocorrelations and pairwise cross-correlations in the sequence of samples generated by the circuit dynamics. Here $\|\mathbf{M}\|_F^2 \equiv \mathrm{trace}(\mathbf{M}\mathbf{M}^\top) = \sum_{ij} M_{ij}^2$ is the squared Frobenius norm of $\mathbf{M}$.

Using the above measure of slowness, we revisit the mixing behavior of LS on a toy covariance matrix $\boldsymbol{\Sigma}$ drawn from the same inverse Wishart distribution mentioned above with parameters $N = 200$, $\sigma_0^2 = 2$ and $\sigma_r = 0.2$. We further regularize $\boldsymbol{\Sigma}$ by adding the identity matrix to it, which does not change anything in terms of the scaling law of Eq. 8 but ensures that the diagonal of $\mathbf{W}_{\mathrm{L}}$ remains bounded as N grows large. We will use the same $\boldsymbol{\Sigma}$ in the rest of the paper. Figure 3A shows $\left\| \boldsymbol{\Lambda}^{-1/2} \mathbf{K}(\mathbf{S}, \tau) \boldsymbol{\Lambda}^{-1/2} \right\|_{\mathrm{F}}$ as a function of the time lag τ: as predicted in Sec. 3, mixing is indeed an order of magnitude slower for LS ($\mathbf{S} = \mathbf{0}$, solid black line) than the single-neuron time constant τ_{m} (grey dashed line). Note that ψ_{slow} (Eq. 10, Fig. 3B) is proportional to the area under the squared curve shown in Fig. 3A. Sample activity traces for this network, implementing LS, can be found in Fig. 4B (top).

Using the same measure of slowness, we also inspected the speed of Gibbs sampling, another widely used sampling technique (e.g. [17]) inspiring neural network dynamics for sampling from distributions over binary variables [18, 19, 20]. Gibbs sampling defines a Markov chain that operates in

discrete time, and also uses a symmetric weight matrix. In order to compare its mixing speed with that of our continuous stochastic dynamics, we assume that a full update step (in which all neurons have been updated once) takes time τ_{m}. We estimated the integrand of the slowing cost (Eq. 10) numerically using 30'000 samples generated by the Gibbs chain (Fig. 3A, blue). Gibbs sampling is comparable to LS here: samples are still correlated on a timescale of order $\sim 50\, \tau_{\mathrm{m}}$.

Finally, one may wonder how a random choice of $\mathbf{S}$ would perform in terms of decorrelation speed. We drew random skew-symmetric $\mathbf{S}$ matrices from the Gaussian ensemble, $S_{i<j} \sim \mathcal{N}(0, \zeta^2)$, and computed the slowing cost (Fig. 3, red). As the magnitude ζ of $\mathbf{S}$ increases, sampling becomes faster and faster until the dynamics is about as fast as the single-neuron time constant τ_{m}. However, the synaptic weights also grow with ζ (Fig. 3C), and we show in Sec. 5 that an even faster sampler exists that has comparatively weaker synapses. It is also interesting to note that the slope of ψ_{slow} at $\zeta = 0$ is zero, suggesting that LS is in fact maximally slow (we prove this formally in the SI).

5   What is the fastest sampler?

We now show that the skew-symmetric matrix $\mathbf{S}$ can be optimized for sampling speed, by directly minimizing the slowing cost $\psi_{\mathrm{slow}}(\mathbf{S})$ (Eq. 10), subject to an L_2-norm penalty. We thus seek to minimize:

$$\mathcal{L}(\mathbf{S}) \quad \equiv \quad \psi_{\mathrm{slow}}(\mathbf{S}) + \frac{\lambda_{L_2}}{2N^2} \left\| \mathbf{W}(\mathbf{S}) \right\|_F^2 . \tag{11}$$

The key to performing this minimization is to use classical Ornstein-Uhlenbeck theory (e.g. [10]) to bring our slowness cost under a form mathematically analogous to a different optimization problem that has arisen recently in the field of robust control [21]. We can then use analytical results obtained there concerning the gradient of ψ_{slow}, and obtain the overall gradient:

$$\frac{\partial \mathcal{L}(\mathbf{S})}{\partial \mathbf{S}} \quad = \quad \frac{1}{N^2} \left[(\mathbf{\Sigma}^{-1}\mathbf{P}\mathbf{Q})^\top - (\mathbf{\Sigma}^{-1}\mathbf{P}\mathbf{Q}) \right] + \frac{\lambda_{L_2}}{N^2} \left[\mathbf{S}\mathbf{\Sigma}^{-2} + \mathbf{\Sigma}^{-2}\mathbf{S} \right] \tag{12}$$

where matrices $\mathbf{P}$ and $\mathbf{Q}$ are obtained by solving two dual Lyapunov equations. All details can be found in our SI.

We initialized $\mathbf{S}$ with random, weak and uncorrelated elements (cf. the end of Sec. 4, with $\zeta = 0.01$), and ran the L-BFGS optimization algorithm using the gradient of Eq. 12 to minimize $\mathcal{L}(\mathbf{S})$ (with $\lambda_{L_2} = 0.1$). The resulting, optimal sampler is an order of magnitude faster than either Langevin or Gibbs sampling: samples are decorrelated on a timescale that is even faster than the single-neuron time constant τ_{m} (Fig. 3A, orange). We also found that fast solutions (with correlation length $\sim \tau_m$) can be found irrespective of the size N of the state space (not shown), meaning that the relative speed-up between the optimal solution and LS grows with N (cf. Fig. 2).

The optimal $\mathbf{S}_{\mathrm{opt}}$ induces a weight matrix $\mathbf{W}_{\mathrm{opt}}$ given by Eq. 9 and shown in Fig. 4A (middle). Notably, $\mathbf{W}_{\mathrm{opt}}$ is no longer symmetric, and its elements are much larger than in the Langevin symmetric solution $\mathbf{W}_{\mathrm{L}}$ with the same stationary covariance, albeit orders of magnitude smaller than in random networks of comparable decorrelation speed (Fig. 3C).

It is illuminating to visualize activity trajectories in the plane defined by the topmost and bottommost eigenvectors of $\mathbf{\Sigma}$, i.e. the first and last principal components (PCs) of the network activity (Fig. 4C). The distribution of interest is broad along some dimensions, and narrow along others. In order to sample efficiently, large steps ought to be taken along directions in which the distribution is broad, and small steps along directions in which the distribution is narrow. This is exactly what our optimal sampler does, whereas LS takes small steps along both broad and narrow directions (Fig. 4C).

6   Balanced E/I networks for fast sampling

We can further constrain our network to obey Dale's law, i.e. the separation of neurons into separate excitatory (E) and inhibitory (I) groups. The main difficulty in building such networks is that picking an arbitrary skew-symmetric matrix $\mathbf{S}$ in Eq. 9 will not yield the column sign structure of an E/I network in general. Therefore, we no longer have a parametric form for the solution matrix manifold on which to find the fastest network. However, by extending the methods of Sec. 5, described in
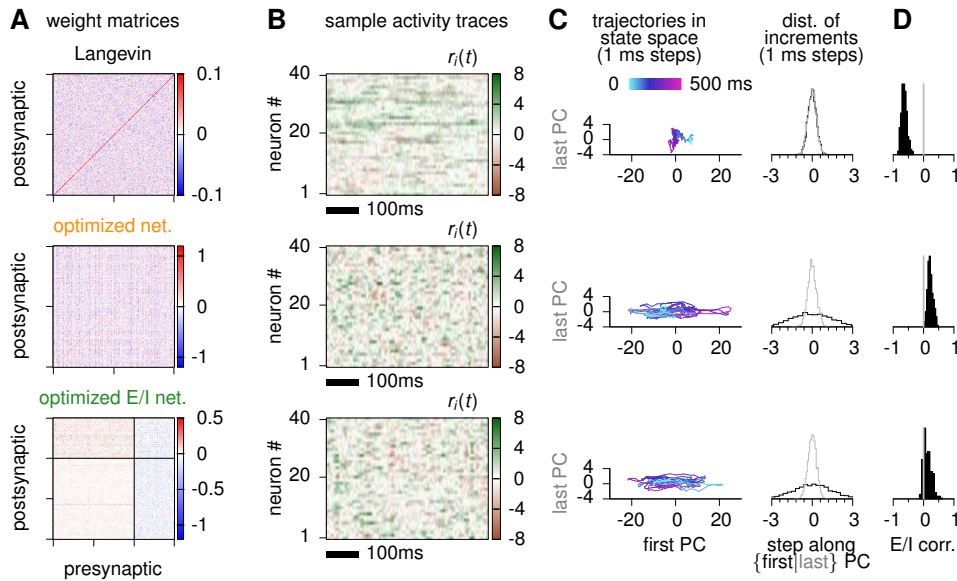
Figure 4: **Fast sampling with optimized networks.** (A) Synaptic weight matrices for the Langevin network (top), the fastest sampler (middle) and the fastest sampler that obeys Dale's law (bottom). Note that the synaptic weights in both optimized networks are an order of magnitude larger than in the symmetric Langevin solution. The first two networks are of size $N = 200$, while the optimized E/I network has size $N + N_I = 300$. (B) 500 ms of spontaneous network activity ($\mathbf{h} = 0$) in each of the three networks, for all of which the stationary distribution of $\mathbf{r}$ (restricted here to the first 40 neurons) is the same multivariate Gaussian. (C) Left: activity trajectories (the same 500 ms as shown in (B)) in the plane defined by the topmost and bottommost eigenvectors of the posterior covariance matrix $\boldsymbol{\Sigma}$ (corresponding to the first and last principal components of the activity fluctuations $\mathbf{r}(t)$). For the E/I network, the projection is restricted to the excitatory neurons. Right: distribution of increments along both axes, measured in 1 ms time steps. Langevin sampling takes steps of comparable size along all directions, while the optimized networks take much larger steps along the directions of large variance prescribed by the posterior. (D) Distributions of correlations between the time courses of total excitatory and inhibitory input in individual neurons.

detail in our SI, we can still formulate the problem as one of unconstrained optimization, and obtain the fastest, balanced E/I sampler.

We consider the posterior to be encoded in the activity of the $N = 200$ excitatory neurons, and add $N_I = 100$ inhibitory neurons which we regard as auxiliary variables, in the spirit of Hamiltonian Monte Carlo methods [11]. Consequently, the E-I and I-I covariances are free parameters, while the E-E covariance is given by the target posterior. For additional biological realism, we also forbid self-connections as they can be interpreted as a modification of the intrinsic membrane time constant of the single neurons, which in principle cannot be arbitrarily learned.

The speed optimization yields the connectivity matrix shown in Fig. 4A (bottom). Results for this network are presented in a similar format as before, in the same figures. Sampling is almost as fast as in the best (regularized) unconstrained network (compare yellow and green in Fig. 3), indicating that Dale's law – unlike the symmetry constraint implicitly present in Langevin sampling – is not fundamentally detrimental to mixing speed. Moreover, the network operates in a regime of excitation/inhibition balance, whereby the total E and I input time courses are correlated in single cells (Fig. 4D, bottom). This is true also in the unconstrained optimal sampler. In contrast, E and I inputs are strongly anti-correlated in LS.

7

7 Discussion

We have studied sampling for Bayesian inference in neural circuits, and observed that a linear stochastic network is able to sample from the posterior under a linear Gaussian latent variable model. Hidden variables are directly encoded in the activity of single neurons, and their joint activity undergoes moment-to-moment fluctuations that visit each portion of the state space at a frequency given by the target posterior density. To achieve this, external noise sources fed into the network are amplified by the recurrent circuitry, but preferentially amplified along the state-space directions of large posterior variance. Although, for the very simple linear Gaussian model we considered here, a purely feed-forward architecture would also trivially be able to provide independent samples (ie. provide samples that are decorrelated at the time scale of τ_{m}), the network required to achieve this is deeply biologically implausible (see SI).

We have shown that the choice of a *symmetric weight matrix* – equivalent to LS, a popular machine learning technique [2, 11, 12] that has been suggested to underlie neuronal network dynamics sampling continuous variables [6, 13] – is most unfortunate. We presented an analytical argument predicting dramatic slowing in high-dimensional latent spaces, supported by numerical simulations. Even in moderately large networks, samples were correlated on timescales much longer than the single-neuron decay time constant.

We have also shown that when the above symmetry constraint is relaxed, a family of other solutions opens up that can potentially lead to much faster sampling. We chose to explore this possibility from a normative viewpoint, optimizing the network connectivity directly for sampling speed. The fastest sampler turned out to be highly asymmetric and typically an order of magnitude faster than Langevin sampling. Notably, we also found that constraining each neuron to be either excitatory or inhibitory does not impair performance while giving a far more biologically plausible sampler. Dale's law could even provide a natural safeguard against reaching slow symmetric solutions such as Langevin sampling, which we saw was the worst-case scenario (cf. also SI).

It is worth noting that $\mathbf{W}_{\mathrm{opt}}$ is strongly nonnormal.[3] Deviation from normality has important consequences for the dynamics of our networks: it makes the network sensitive to perturbations along some directions in state space. Such perturbations are rapidly amplified into large, transient excursions along other, relevant directions. This phenomenon has been shown to explain some key features of spontaneous activity in primary visual cortex [9] and primary motor cortex [22].

Several aspects would need to be addressed before our proposal can crystalize into a more thorough understanding of the neural implementation of the sampling hypothesis. First, can local synaptic plasticity rules perform the optimization that we have approached from an algorithmic viewpoint? Second, what is the origin of the noise that we have hypothesized to come from external sources? Third, what kind of nonlinearity must be added in order to allow sampling from non-Gaussian distributions, whose shapes may have non-trivial dependencies on the observations? Also, does the main insight reached here – namely that fast samplers are to be found among nonsymmetric, nonnormal networks – carry over to the nonlinear case? As a proof of principle, in preliminary simulations, we have shown that speed optimization in a linearized version of a nonlinear network (with a tanh gain function) does yield fast sampling in the nonlinear regime, even when fluctuations are strong enough to trigger the nonlinearity and make the resulting sampled distribution non-Gaussian (details in SI).

Finally, we have also shown (see SI) that the Langevin solution is the only network that satisfies the detailed balance condition [23] in our model class; reversibility is violated in all other stochastic networks we have presented here (random, optimal, optimal E/I). The fact that these networks are faster samplers is in line with recent machine learning studies on how non-reversible Markov chains can mix faster than their reversible counterparts [24]. The construction of such Monte-Carlo algorithms has proven challenging [25, 26, 27], suggesting that the brain – if it does indeed use sampling-based representations – might have something yet to teach us about machine learning.

---

[3]Indeed, the sum of the squared moduli of its eigenvalues accounts for only 25% of $\|\mathbf{W}_{\mathrm{opt}}\|_{\mathrm{F}}^2$ [7]. For a normal matrix $\mathbf{W}$ (such as the Langevin solution, $\mathbf{W}_{\mathrm{L}}$), $\sum_i |\lambda_i|^2 = \|\mathbf{W}\|_{\mathrm{F}}^2$, i.e. this ratio is 100%.

References

[1] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381:520—522, 1996.

[2] D. MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003.

[3] D. Knill and A. Pouget. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27:712–719, 2004.

[4] J. Fiser, P. Berkes, G. Orbán, and M. Lengyel. Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, 14:119–130, 2010.

[5] P. Berkes, G. Orbán, M. Lengyel, and J. Fiser. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331:83–87, 2011.

[6] R. Moreno-Bote, D. C. Knill, and A. Pouget. Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, 108:12491–12496, 2011.

[7] L. N. Trefethen and M. Embree. *Spectra and pseudospectra: the behavior of nonnormal matrices and operators*. Princeton University Press, 2005.

[8] G. Hennequin, T. P. Vogels, and W. Gerstner. Non-normal amplification in random balanced neuronal networks. *Physical Review E*, 86:011909, 2012.

[9] B. K. Murphy and K. D. Miller. Balanced amplification: A new mechanism of selective amplification of neural activity patterns. *Neuron*, 61:635–648, 2009.

[10] C. W. Gardiner. *Handbook of stochastic methods: for physics, chemistry, and the natural sciences*. Berlin: Springer, 1985.

[11] R. Neal. MCMC using Hamiltonian dynamics. *Handbook of MCMC*, pages 113–162, 2011.

[12] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the International Conference on Machine Learning*, 2011.

[13] A. Grabska-Barwinska, J. Beck, A. Pouget, and P. Latham. Demixing odors - fast inference in olfaction. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1968–1976. Curran Associates, Inc., 2013.

[14] Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62:1707–1739, 2009.

[15] M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:123–214, 2011.

[16] J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34:A1460–A1487, 2012.

[17] M. Mezard and A. Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[18] G. E. Hinton and T. J. Sejnowski. Learning and relearning in Boltzmann machines. In D E Rumelhart, J L McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1: Foundations, chapter 7, pages 282–317. MIT Press, Cambridge, MA, 1986.

[19] G E Hinton, P Dayan, B J Frey, and R M Neal. The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161, 1995.

[20] L. Buesing, J. Bill, B. Nessler, and W. Maass. Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology*, 7:1–22, 2011.

[21] J. Vanbiervliet, B. Vandereycken, W. Michiels, S. Vandewalle, and M. Diehl. The smoothed spectral abscissa for robust stability optimization. *SIAM Journal on Optimization*, 20:156–171, 2009.

[22] G. Hennequin, T. P. Vogels, and W. Gerstner. Optimal control of transient dynamics in balanced networks supports generation of complex movements. *Neuron*, 82, 2014.

[23] W. Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[24] Akihisa Ichiki and Masayuki Ohzeki. Violation of detailed balance accelerates relaxation. *Physical Review E*, 88:020101, 2013.

[25] Y. Sun, J. Schmidhuber, and F. J. Gomez. Improving the asymptotic performance of Markov Chain Monte-Carlo by inserting vortices. In J.D. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2235–2243. 2010.

[26] K. S. Turitsyn, M. Chertkov, and M. Vucelja. Irreversible Monte Carlo algorithms for efficient sampling. *Physica D: Nonlinear Phenomena*, 240:410–414, 2011.

[27] Joris Bierkens. Non-reversible Metropolis-Hastings. *arXiv:1401.8087 [math]*, 2014.

Fast Sampling-Based Inference in Balanced Neuronal Networks
— Supplementary Information —

Guillaume Hennequin[1]
gjeh2@cam.ac.uk

Laurence Aitchison[2]
laurence@gatsby.ucl.ac.uk

Máté Lengyel[1]
m.lengyel@eng.cam.ac.uk

[1]Computational & Biological Learning Lab, Dept. of Engineering, University of Cambridge, UK
[2]Gatsby Computational Neuroscience Unit, University College London, UK

1 Lower bound on the maximum eigenvalue of the Langevin connectivity matrix

Here we give an informal derivation of the lower bound on $\mathrm{Re}\left(\lambda_{\max}^{\mathbf{W_L}-\mathbf{I}}\right)$ found in Eq. 7 of the main text. Let us recall that $(\mathbf{W_L}-\mathbf{I})$ is real and symmetric, so its eigenvalues are all real, and since $\mathbf{W_L}-\mathbf{I} = -\sigma_\xi^2\boldsymbol{\Sigma}^{-1}$ we can write[1]

$$\lambda_{\max}^{\mathbf{W_L}-\mathbf{I}} \quad = \quad -\sigma_\xi^2\lambda_{\min}^{\boldsymbol{\Sigma}^{-1}} \quad = \quad -\frac{\sigma_\xi^2}{\lambda_{\max}^{\boldsymbol{\Sigma}}} \tag{1}$$

Now, again because of its symmetry, $\boldsymbol{\Sigma}$ is a normal matrix, and so it is similar to (i.e. equal to the unitary transformation of) a diagonal matrix that contains its eigenvalues. Since unitary transformations preserve the Frobenius norm, we can write $\sum_{i,j}\Sigma_{ij}^2 = \sum_i\left(\lambda_i^{\boldsymbol{\Sigma}}\right)^2$ and since all the eigenvalues of $\boldsymbol{\Sigma}$ are positive, we have $N\left(\lambda_{\max}^{\boldsymbol{\Sigma}}\right)^2 \geq \sum_i\left(\lambda_i^{\boldsymbol{\Sigma}}\right)^2$. Plugging this into Eq. 1, we arrive at a bound that relates the maximum eigenvalue of $(\mathbf{W_L}-\mathbf{I})$ to a basic summary statistics, the sum of all (co)variances, of the posterior covariance matrix $\boldsymbol{\Sigma}$:

$$\lambda_{\max}^{\mathbf{W_L}-\mathbf{I}} \quad \geq \quad -\sigma_\xi^2\sqrt{\frac{N}{\sum_{ij}\Sigma_{ij}^2}} \tag{2}$$

In the $N \to \infty$ limit, assuming that pairwise correlations do not vanish, the denominator is $\mathcal{O}(N^2)$, meaning that $0 > \lambda_{\max}^{\mathbf{W_L}-\mathbf{I}} \geq -\mathcal{O}(1/\sqrt{N})$: the slowest eigenmode of $\mathbf{W_L}$ becomes critically slow. To make this bound more concrete, let us assume that $\Sigma_{ii} \simeq \sigma_0^2$ (all posterior variances are roughly equal) and that the distribution of pairwise posterior correlations has zero mean and standard deviation σ_r. We can then rewrite Eq. 2 as

$$\lambda_{\max}^{\mathbf{W_L}-\mathbf{I}} \quad \geq \quad \frac{-(\sigma_\xi/\sigma_0)^2}{\sqrt{1+N\sigma_r^2}} \tag{3}$$

which is Eq. 7 of our main text.

---

[1]For a non-singular matrix $\mathbf{M}$, the eigenvalues of $\mathbf{M}^{-1}$ are the inverses of those of $\mathbf{M}$; and since $\boldsymbol{\Sigma}$ is a positive definite covariance matrix, all its eigenvalues are positive, which yields Eq. 1.

2 Minimization of the slowing cost ψ_{slow}

Let us recall the definition of the slowing cost for convenience:

$$\psi_{\text{slow}}(\mathbf{S}) = \frac{1}{2\tau_{\text{m}}N^2} \int_0^\infty \left\| \mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{K}(\mathbf{S}, \tau) \mathbf{\Lambda}^{-\frac{1}{2}} \right\|_{\text{F}}^2 \, d\tau \tag{4}$$

where $\mathbf{K}(\mathbf{S}, \tau) \equiv \left\langle \delta\mathbf{r}(t + \tau)\, \delta\mathbf{r}(t)^\top \right\rangle_t$.

From Ornstein-Uhlenbeck theory [1], we know that $\mathbf{K}(\mathbf{S}, \tau)$ obeys the following differential equation:

$$\tau_{\text{m}} \frac{d\mathbf{K}(\mathbf{S}, \tau)}{d\tau} = [\mathbf{W}(\mathbf{S}) - \mathbf{I}]\, \mathbf{K}(\mathbf{S}, \tau) \tag{5}$$

such that for $\tau \geq 0$, we have $\mathbf{K}(\mathbf{S}, \tau) = e^{[\mathbf{W}(\mathbf{S}) - \mathbf{I}]\, \tau/\tau_{\text{m}}} \mathbf{\Sigma}$. We may thus rewrite $\psi_{\text{slow}}(\mathbf{S})$ as

$$\psi_{\text{slow}}(\mathbf{S}) = \frac{1}{2N^2} \text{trace} \left[\mathbf{V} \left(\int_0^\infty e^{\tau[\mathbf{W}(\mathbf{S}) - \mathbf{I}]} \mathbf{U}\mathbf{U}^\top e^{\tau[\mathbf{W}(\mathbf{S})^\top - \mathbf{I}]} d\tau \right) \mathbf{V}^\top \right] \tag{6}$$

with the shorthand notation $\mathbf{U} \equiv \mathbf{\Sigma}\mathbf{\Lambda}^{-\frac{1}{2}}$ and $\mathbf{V} \equiv \mathbf{\Lambda}^{-\frac{1}{2}}$. Equation 6 is the canonical form used in linear quadratic control theory [2] and affords the following gradient:

$$\frac{\partial \psi_{\text{slow}}(\mathbf{S})}{\partial \mathbf{S}} = \frac{1}{N^2} \left[(\mathbf{\Sigma}^{-1}\mathbf{P}\mathbf{Q})^\top - (\mathbf{\Sigma}^{-1}\mathbf{P}\mathbf{Q}) \right] \tag{7}$$

Here $\mathbf{P}$ and $\mathbf{Q}$ are the unique solutions of a pair of Lyapunov equations,

$$(\mathbf{W} - \mathbf{I})\mathbf{P} + \mathbf{P}(\mathbf{W} - \mathbf{I})^\top = -\mathbf{\Sigma}\mathbf{\Lambda}^{-1}\mathbf{\Sigma} \tag{8}$$

$$(\mathbf{W} - \mathbf{I})^\top \mathbf{Q} + \mathbf{Q}(\mathbf{W} - \mathbf{I}) = -\mathbf{\Lambda}^{-1}, \tag{9}$$

which can be solved efficiently [3], e.g. using the Matlab function `lyap`. Note also that $\psi_{\text{slow}}(\mathbf{S}) = \text{trace}(\mathbf{\Lambda}^{-1/2}\mathbf{P}\mathbf{\Lambda}^{-1/2})/2N^2$ [2].

The L_2-penalty term in the overal cost function (Eq. 13 of the main text) is more easily differentiated, yielding the gradient

$$\frac{\partial \mathcal{L}(\mathbf{S})}{\partial \mathbf{S}} = \frac{1}{N^2} \left[(\mathbf{\Sigma}^{-1}\mathbf{P}\mathbf{Q})^\top - (\mathbf{\Sigma}^{-1}\mathbf{P}\mathbf{Q}) \right] + \frac{\lambda_{L_2}}{N^2} \left[\mathbf{S}\mathbf{\Sigma}^{-2} + \mathbf{\Sigma}^{-2}\mathbf{S} \right] \tag{10}$$

which is skew-symmetric, as it should.

3 *Only* Langevin sampling (LS) satisfies detailed balance (in our model class)

Consider a Markov chain $\{\mathbf{x}_t\}$ with stationary distribution $p(\mathbf{x}_t)$ and a probability of transitioning from state $\mathbf{x}_t$ into state $\mathbf{x}_{t+1}$ given by $p(\mathbf{x}_{t+1}|\mathbf{x}_t)$. Detailed balance is satisfied if, and only if for any pair of states $(\mathbf{x}_t, \mathbf{x}_{t+1})$, we have

$$p(\mathbf{x}_{t+1}|\mathbf{x}_t)\, p(\mathbf{x}_t) = p(\mathbf{x}_t|\mathbf{x}_{t+1})\, p(\mathbf{x}_{t+1}) \tag{11}$$

Equation 11 states that any state sequence $\mathbf{x}_t \to \mathbf{x}_{t+1}$ should be visited as often as the reverse sequence $\mathbf{x}_{t+1} \to \mathbf{x}_t$, that is, time is reversible. Taking the logarithm on both sides, we rewrite the detailed balance condition as

$$\log p(\mathbf{x}_{t+1}|\mathbf{x}_t) + \log p(\mathbf{x}_t) = \log p(\mathbf{x}_t|\mathbf{x}_{t+1}) + \log p(\mathbf{x}_{t+1}) \tag{12}$$

To see whether or not detailed balance holds in our samplers, we write the network dynamics (Eq. 3 in the main text) in discrete time, i.e. for $\epsilon \to 0$ we have

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \epsilon\mathbf{A}\mathbf{x}_t + \sqrt{2\epsilon}\eta_t \tag{13}$$

where $\mathbf{A} \equiv \mathbf{W} - \mathbf{I} = (-\mathbf{I} + \mathbf{S})\mathbf{\Sigma}^{-1}$ is the effective connectivity (it includes the leak term), $\eta_t \sim \mathcal{N}(0, \mathbf{I})$, and both τ_{m} and σ_ξ have been set to unity without loss of generality.

Thus

$$\log p(\mathbf{x}_{t+1}|\mathbf{x}_t) = -\frac{N}{2}\log(2\pi\epsilon) - \frac{\|\mathbf{x}_{t+1} - (\mathbf{I} + \epsilon\mathbf{A})\mathbf{x}_t\|^2}{4\epsilon} \tag{14}$$

2

and, given that our samplers have the right stationary distribution $\mathcal{N}(0, \boldsymbol{\Sigma})$ (detailed balance is not required to show this, see [1]),

$$\log p(\mathbf{x}_t) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\mathbf{x}_t^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}_t \qquad (15)$$

Therefore, detailed balance is satisfied if, and only if for any state pair $(\mathbf{x}_t, \mathbf{x}_{t+1})$ we have:

$$\|\mathbf{x}_{t+1} - (\mathbf{I} + \epsilon\mathbf{A})\mathbf{x}_t\|^2 + 2\epsilon\mathbf{x}_t^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}_t = \|\mathbf{x}_t - (\mathbf{I} + \epsilon\mathbf{A})\mathbf{x}_{t+1}\|^2 + 2\epsilon\mathbf{x}_{t+1}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}_{t+1} \qquad (16)$$

Keeping only first-order terms in ϵ, and inserting the parameterization $\mathbf{A} = (-\mathbf{I} + \mathbf{S})\boldsymbol{\Sigma}^{-1}$, we obtain the following necessary and sufficient condition for reversibility:

$$2(\mathbf{x}_{t+1} - \mathbf{x}_t)^\top \mathbf{S}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_{t+1} - \mathbf{x}_t) = 0 \qquad (17)$$

Clearly, if $\mathbf{S} = 0$, the condition is satisfied, therefore detailed balance holds. Conversely, if detailed balance holds, then Eq. 17 must hold for *any* pair $(\mathbf{x}_t, \mathbf{x}_{t+1})$, from which it is easy to see that $\mathbf{S}\boldsymbol{\Sigma}^{-1}$ must be zero, and therefore $\mathbf{S} = 0$ too. Therefore, only the Langevin solution (which corresponds to $\mathbf{S} = 0$) satisfies time reversibility.

4   LS is at the pessimum of the slowness cost function

Here we prove that LS corresponds to a pessimum of the slowness cost function ψ_{slow} used throughout the paper to optimize for sampling speed. To do this, we show that the gradient $\partial\psi_{\text{slow}}/\partial\mathbf{S}$ is zero at $\mathbf{S} = 0$.

Let us assume that $\mathbf{S} = 0$. Then,

$$\mathbf{A} = -\sigma_\xi^2 \boldsymbol{\Sigma}^{-1} \qquad (18)$$

such that the two Lyapunov equations (Eqs. 8 and 9) of Sec. 2 become:

$$\sigma_\xi^2 \left(\boldsymbol{\Sigma}^{-1}\mathbf{P} + \mathbf{P}\boldsymbol{\Sigma}^{-1} \right) = \boldsymbol{\Sigma}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma} \qquad (19)$$

$$\sigma_\xi^2 \left(\boldsymbol{\Sigma}^{-1}\mathbf{Q} + \mathbf{Q}\boldsymbol{\Sigma}^{-1} \right) = \boldsymbol{\Lambda}^{-1} \qquad (20)$$

Now,

$$\sigma_\xi^2 \left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}\mathbf{Q}\boldsymbol{\Sigma}) + (\boldsymbol{\Sigma}\mathbf{Q}\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1} \right) = \sigma_\xi^2 \left(\mathbf{Q}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\mathbf{Q} \right) \qquad (21)$$

$$= \sigma_\xi^2 \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma}^{-1}\mathbf{Q} + \mathbf{Q}\boldsymbol{\Sigma}^{-1} \right) \boldsymbol{\Sigma} \qquad (22)$$

$$= \boldsymbol{\Sigma}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma} \qquad (23)$$

(the last equality uses Eq. 20). Thus, $\boldsymbol{\Sigma}\mathbf{Q}\boldsymbol{\Sigma}$ is a solution to Eq. 19, and since the solution is unique because $\mathbf{P}$ is positive definite [2], we conclude that

$$\mathbf{P} = \boldsymbol{\Sigma}\mathbf{Q}\boldsymbol{\Sigma} \qquad (24)$$

i.e

$$\boldsymbol{\Sigma}^{-1}\mathbf{P} = \mathbf{Q}\boldsymbol{\Sigma} \qquad (25)$$

Using this result in Eq. 7, together with the fact that $\mathbf{P}$, $\mathbf{Q}$ and $\boldsymbol{\Sigma}$ are symmetric, we compute:

$$\frac{\partial\psi_{\text{slow}}(\mathbf{S})}{\partial\mathbf{S}} \propto \left(\boldsymbol{\Sigma}^{-1}\mathbf{P}\mathbf{Q} \right)^\top - \boldsymbol{\Sigma}^{-1}\mathbf{P}\mathbf{Q} \qquad (26)$$

$$= \mathbf{Q}\mathbf{P}\boldsymbol{\Sigma}^{-1} - \mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q} \qquad (27)$$

$$= \mathbf{Q}(\boldsymbol{\Sigma}^{-1}\mathbf{P})^\top - \mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q} \qquad (28)$$

$$= \mathbf{Q}(\mathbf{Q}\boldsymbol{\Sigma})^\top - \mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q} \qquad (29)$$

$$= \mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q} - \mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q} \qquad (30)$$

$$= 0 \qquad (31)$$

At this stage, we have shown that $\mathbf{S} = 0$ corresponds to a critical point of ψ_{slow}. The fact that small, random (unstructured) perturbations of $\mathbf{S}$ around 0 only decrease ψ_{slow} (Fig. 3 of the main text) suggest that LS is in fact (locally, but perhaps also globally) the slowest possible sampler for our problem.

5   Details of the balanced E/I network optimization

To build optimized networks that obey Dale's law, we assume that there are $N_{\text{exc.}} = N$ excitatory neurons, where N is the dimension of the distribution we want to sample from, and $N_{\text{inh.}}$ inhibitory neurons whose activity distribution is irrelevant (i.e. we regard inhibitory neurons as auxiliary sampling variables, in the spirit of Hamiltonian Monte Carlo methods [4]). In the main paper, $N = 200$ and $N_I = 100$. Let $M = N_{\text{exc.}} + N_{\text{inh.}}$ denote the total network size. The dynamics do not change, i.e. we still have

$$d\mathbf{r} = \frac{dt}{\tau_{\text{m}}} \left[-\mathbf{r}(t) + \mathbf{W}\mathbf{r}(t) + \mathbf{F}\mathbf{h} \right] + \sigma_\xi \sqrt{\frac{2}{\tau_{\text{m}}}} \, d\boldsymbol{\xi}(t) \tag{32}$$

The connectivity matrix $\mathbf{W}$ is now made of N positive columns followed by N_I negative columns. This makes it difficult to apply the approach of Sec. 5 of the main text, as picking an arbitrary skew-symmetric matrix $\mathbf{S}$ in Eq. 11 (main text) will not yield the column sign structure of an E/I network in general. Therefore, we no longer have a parametric form for the solution matrix manifold on which to search for the fastest network. However, with a few simple variations, we can still formulate the problem as one of unconstrained optimization, as explained now.

The first step is to enforce Dale's law through the following re-parameterization of $\mathbf{W}$:

$$W_{ij} = (1 - \delta_{ij}) \, s_j \, \exp \beta_{ij} \tag{33}$$

where s_j is a fixed sign that depends only on presynaptic neuron j ($s_j = +1$ for $j \leq N$, -1 otherwise), and the β_{ij}'s are unconstrained free parameters. Note that we do not allow for autapses, hence the $(1 - \delta_{ij})$ term in Eq. 33). Second, since the target posterior distribution specifies only the $N \times N$ upper-left quadrant $\boldsymbol{\Sigma}$ of the overall covariance matrix which we denote by $\boldsymbol{\Sigma}_{\text{tot}}$, we are free to optimize over the other quadrants. We parameterize $\boldsymbol{\Sigma}_{\text{tot}}$ by its Cholesky factor:

$$\boldsymbol{\Sigma}_{\text{tot}} = \mathbf{L}\mathbf{L}^\top, \qquad \mathbf{L} \equiv \begin{pmatrix} \mathbf{L}_{11} & 0 \\ \mathbf{L}_{12} & \mathbf{L}_{22} \end{pmatrix} \tag{34}$$

where $\mathbf{L}_{11}$ is the Cholesky factor of the posterior covariance matrix $\boldsymbol{\Sigma}$ (i.e. $\boldsymbol{\Sigma} = \mathbf{L}_{11}\mathbf{L}_{11}^\top$), and the two matrices $\mathbf{L}_{12}$ and $\mathbf{L}_{22}$ are free parameters. Note that $\mathbf{L}_{12}$ is a full rectangular matrix of size $N_I \times N$, while $\mathbf{L}_{22}$ is lower-triangular with dimensions $N_I \times N_I$. Third, in order to force the network to sample from the right multivariate Gaussian distribution, we incorporate the Lyapunov equation (cf. Eq. 4 in the main text) as an additional constraint in our loss function. This additional term reads:

$$\psi_{\text{sol.}} \equiv \frac{1}{2M^2} \left\| (\mathbf{W} - \mathbf{I})\boldsymbol{\Sigma}_{\text{tot}} + \boldsymbol{\Sigma}_{\text{tot}}(\mathbf{W} - \mathbf{I})^\top + 2\sigma_\xi^2 \mathbf{I} \right\|_{\text{F}}^2 \tag{35}$$

When $\psi_{\text{sol.}}$ is zero, the Lyapunov equation (Eq. 4 in the main text) is satisfied, and therefore the stationary covariance matrix of the network dynamics matches $\boldsymbol{\Sigma}_{\text{tot}}$. In particular, their upper-left quadrant would then be equal, meaning that the excitatory sub-network would be sampling from the right posterior.

Note that $\psi_{\text{sol.}}$ depends on both the β_{ij}'s (through $\mathbf{W}$) and the free covariance parameters $\mathbf{L}_{12}$ and $\mathbf{L}_{22}$ (through $\boldsymbol{\Sigma}_{\text{tot}}$). The corresponding gradients can be obtained after a bit of algebra, and read:

$$\frac{\partial \psi_{\text{sol.}}}{\partial \mathbf{L}} = \frac{2}{M^2} \left[(\mathbf{G}\mathbf{A}) + (\mathbf{G}\mathbf{A})^\top \right] \mathbf{L} \tag{36}$$

$$\frac{\partial \psi_{\text{sol.}}}{\partial \mathbf{W}} = \frac{2}{M^2} \mathbf{G}\boldsymbol{\Sigma}_{\text{tot}} \tag{37}$$

where

$$\mathbf{G} \equiv (\mathbf{W} - \mathbf{I})\boldsymbol{\Sigma}_{\text{tot}} + \boldsymbol{\Sigma}_{\text{tot}}(\mathbf{W} - \mathbf{I})^\top + 2\sigma_\xi^2 \mathbf{I} \tag{38}$$

Note that we are interested only in the lower-triangular part of Eq. 36. The application of the chain rule to go from $\mathbf{W}$ to the β_{ij}'s in Eq. 37 is straightforward (it can be performed element-wise).

The total cost function we minimize is

$$\mathcal{L} \equiv \psi_{\text{sol.}} + \lambda_{\text{slow}} \psi_{\text{slow}} + \frac{\lambda_{L_2}}{2M^2} \|\mathbf{W}\|_{\text{F}}^2 \tag{39}$$

where ψ_{slow} penalizes the magnitude of lagged auto- and cross-correlations over an infinite time horizon, in the excitatory sub-network only. To give a formal definition of ψ_{slow}, let us use the

notation $\tilde{\mathbf{A}}$ to denote the zeroing of all elements but those in the upper-left $N \times N$ quadrant of any $M \times M$ matrix $\mathbf{A}$. The modified slowness loss is then written as

$$\psi_{\text{slow}}(\mathbf{S}) \equiv \frac{1}{2\tau_{\text{m}} N^2} \int_0^\infty \left\| \mathbf{\Lambda}^{-\frac{1}{2}} \tilde{\mathbf{K}}(\mathbf{S}, \tau) \mathbf{\Lambda}^{-\frac{1}{2}} \right\|_{\text{F}}^2 \, \mathrm{d}\tau \tag{40}$$

where $\mathbf{\Lambda}$ is a diagonal matrix that contains the diagonal of $\mathbf{\Sigma}$ in its upper-left quadrant and zeros everywhere else. The kernel $\mathbf{K}$ is defined as in the main text, i.e.

$$\mathbf{K}(\mathbf{S}, \tau) \equiv \langle \mathbf{r}(t) \mathbf{r}(t+\tau)^\top \rangle_t \tag{41}$$
$$= e^{(\mathbf{W} - \mathbf{I})\tau/\tau_{\text{m}}} \mathbf{\Sigma}_{\text{tot}} \tag{42}$$

(the second equality can be found in e.g. [1]). Observing that $\tilde{\mathbf{K}}(\mathbf{S}, \tau) = \tilde{\mathbf{I}} \mathbf{K}(\mathbf{S}, \tau) \tilde{\mathbf{I}}$, and making the change of variable $\tau/\tau_{\text{m}} \to \tau$, we can rewrite the slowness cost as

$$\psi_{\text{slow}}(\mathbf{S}) = \frac{1}{2N^2} \text{trace} \left[\mathbf{V} \left(\int_0^\infty e^{\tau(\mathbf{W}-\mathbf{I})} \mathbf{U} \mathbf{U}^\top e^{\tau(\mathbf{W}-\mathbf{I})^\top} \mathrm{d}\tau \right) \mathbf{V}^\top \right] \tag{43}$$

with

$$\mathbf{V} \equiv \mathbf{\Lambda}^{-\frac{1}{2}} \tilde{\mathbf{I}} \qquad\qquad = \tilde{\mathbf{\Lambda}}^{-\frac{1}{2}} \tag{44}$$
$$\mathbf{U} \equiv \mathbf{\Sigma} \tilde{\mathbf{I}} \mathbf{\Lambda}^{-\frac{1}{2}} \qquad\qquad = \mathbf{\Sigma} \tilde{\mathbf{\Lambda}}^{-\frac{1}{2}} \tag{45}$$

As in the main text we have to solve two dual Lyapunov equations for matrices $\mathbf{P}$ and $\mathbf{Q}$:

$$\mathbf{A} \mathbf{P} + \mathbf{P} \mathbf{A}^\top = -\mathbf{\Sigma}_{\text{tot}} \tilde{\mathbf{\Lambda}}^{-1} \mathbf{\Sigma}_{\text{tot}} \tag{46}$$
$$\mathbf{A}^\top \mathbf{Q} + \mathbf{Q} \mathbf{A} = -\tilde{\mathbf{\Lambda}}^{-1} \tag{47}$$

Thus [2],

$$\psi_{\text{slow}} = \frac{1}{2N^2} \text{trace} \left(\tilde{\mathbf{\Lambda}}^{-\frac{1}{2}} \mathbf{P} \tilde{\mathbf{\Lambda}}^{-\frac{1}{2}} \right) \tag{48}$$

and the gradient w.r.t. the synaptic weights is again given by

$$\frac{\partial \psi_{\text{slow}}}{\partial \mathbf{W}} = \frac{\mathbf{Q} \mathbf{P}}{N^2} \tag{49}$$

The gradient w.r.t. the Cholesky factor $\mathbf{L}$ requires a bit more algebra, and reads

$$\frac{\partial \psi_{\text{slow}}}{\partial \mathbf{L}} = 2 \left[\left(\tilde{\mathbf{\Lambda}}^{-1} \mathbf{\Sigma}_{\text{tot}} \mathbf{Q} \right) + \left(\tilde{\mathbf{\Lambda}}^{-1} \mathbf{\Sigma}_{\text{tot}} \mathbf{Q} \right) \right] \mathbf{L} \tag{50}$$

We used again the L-BFGS algorithm (from the NLopt library) to optimize $\mathcal{L}$ (Eq. 39), with parameters $\lambda_{L_2} = \lambda_{\text{slow}} = 0.1$.

6 Connection to Newton preconditioning

Given an arbitrary target distribution $p(\mathbf{r}) \propto \exp[-V(\mathbf{r})]$, LS can be more generally written as [5]

$$\mathrm{d}\mathbf{r} = -\mathbf{B} \nabla_{\mathbf{r}} V(\mathbf{r}(t)) \mathrm{d}t + \sqrt{2\mathbf{B}} \mathrm{d}\boldsymbol{\xi}(t) \tag{51}$$

where $\boldsymbol{\xi}$ is a unitary Wiener process and $\mathbf{B}$ is an arbitrary positive definite matrix (a "preconditioning" matrix). Note that the noise must also be preconditioned.

Inspired by classical Newton methods for gradient-based optimization, which take into account the curvature of the objective function, Martin et al. [5] advocate the use of the inverse Hessian as a preconditioner, shown to yield substantial speed improvements in many cases. In our Gaussian case, picking the inverse Hessian $\mathbf{H}^{-1} = -\mathbf{\Sigma}$ of the log-posterior as a preconditioner gives exactly:

$$\mathrm{d}\mathbf{r}(t) = -\mathbf{r}(t) + \sqrt{2\mathbf{\Sigma}} \mathrm{d}\boldsymbol{\xi}(t) \tag{52}$$

i.e. it pushes the target covariance in the noise term and removes all recurrent interactions (note that we have assumed zero posterior mean without loss of generality).

Our optimized samplers are thus quite different from smartly preconditioned Langevin sampling, in that they can still make use of *independent* noise sources.

7   Why feed-forward networks are insufficient

In the main text, we have shown how recurrent neural networks can sample fast from multivariate Gaussian distributions. Algorithmically, it is straightforward to draw independent samples from such distributions, and indeed, the standard Cholesky sampling algorithm can be seen as a feed-forward neural network in which the input layer is made of independent noisy neurons, and the output layer (consisting of the neurons that represent the posterior) mixes those inputs, as well as the external stimulus, linearly. However, this solution assumes that some neurons are stochastic and uncorrelated, and their only role is generating noise for the rest of the brain, while others (those representing the posterior in the output layer) respond deterministically to their inputs. We find this dichotomy physiologically highly unrealistic. Moreover, a feedforward layout is inconsistent with the ubiquitously recurrent nature of cortical connectivity, of which, in contrast, our networks make optimal use to support the computation.

8   Application to a nonlinear, non-Gaussian system

A fundamental motivation for focusing on recurrent networks is the intuition that only such complex architectures will be able to sample from non-Gaussian posterior distributions in which not only the mean, but also higher-order moments might have non-trivial dependencies on the input. Although the analytical results of the main text, in particular the gradient of the slowness cost function, are limited to linear networks, we have made preliminary explorations of nonlinear, non-Gaussian systems through simulations.

Our starting point is a nonlinear, random network with stochastic dynamics of the form:

$$\mathrm{d}\mathbf{u} = \frac{\mathrm{d}t}{\tau_{\mathrm{m}}}\left[-\mathbf{u}(t) + \mathbf{W}^{(0)}\mathbf{r}(t)\right] + \sigma_\xi\sqrt{\frac{2}{\tau_{\mathrm{m}}}}\mathrm{d}\xi(t) \qquad \text{with} \quad r_i(t) = \tanh[u_i(t)] \qquad (53)$$

where $W_{ij}^{(0)} \sim \mathcal{N}(0, R^2/N)$ such that $\mathbf{W}^{(0)}$ has a circular eigenvalue spectrum of radius $R = 5$. We set the network size to $N = 200$. A sample activity trace from this nonlinear system, $r_i(t)$, is shown in Fig. 1A (blue). Neurons tend to spend prolonged periods of time at either lower or upper saturation of their nonlinear gain function ($\tanh$), yielding an autocorrelation length about 3-4 times greater than the membrane time constant (Fig. 1B, blue). There are strong negative and positive correlations in the joint activity of neuron pairs (Fig. 1C, x-axis).

We then asked if we could build a nonlinear network that would sample from (approximately) the same distribution, but faster. To apply our linear framework, we first estimated the covariance matrix $\mathbf{\Sigma_u} \equiv \langle \mathbf{u}(t)\mathbf{u}(t)^\top \rangle_t$ on the basis of a 100 second-long simulation of the nonlinear stochastic dynamics of Eq. 53. We then built an optimal linear network to sample from a normal distribution with covariance $\mathbf{\Sigma_u}$ as described in the main text, and finally used the resulting connectivity matrix $\mathbf{W}$ in place of $\mathbf{W}^{(0)}$ in the dynamics of Eq. 53.

This procedure yielded a nonlinear network that turned out to sample from approximately the same non-Gaussian multivariate distribution $p(\mathbf{r})$ as the original nonlinear network. Indeed, individual pairwise correlations $\langle r_i(t)r_j(t) \rangle_t$ approximated those in the original network to a good degree (Fig. 1C), as did individual marginals of $p(r_i)$ (not shown). Importantly, sampling was several times faster, as can be inferred from the sample activity trace of Fig. 1A (red) and as summarized in the average autocorrelation of $r_i(t)$ shown in Fig. 1B (red). Note, that since the original nonlinear network often operates close to saturation, and thus makes heavy use of its nonlinearities, it is not at all trivial that our speed optimization based on a linear approximation work so well. In fact, if, instead of the speed optimized nonnormal weight matrix, we use the corresponding Langevin solution from the linearized dynamics, then the nonlinear version of the dynamics does not only slow down but even fails to match the correct stationary distribution. This is because the Langevin solution encodes the posterior distribution in the principal eigenvectors of the weight matrix, and those are the directions along which saturation occurs the most.
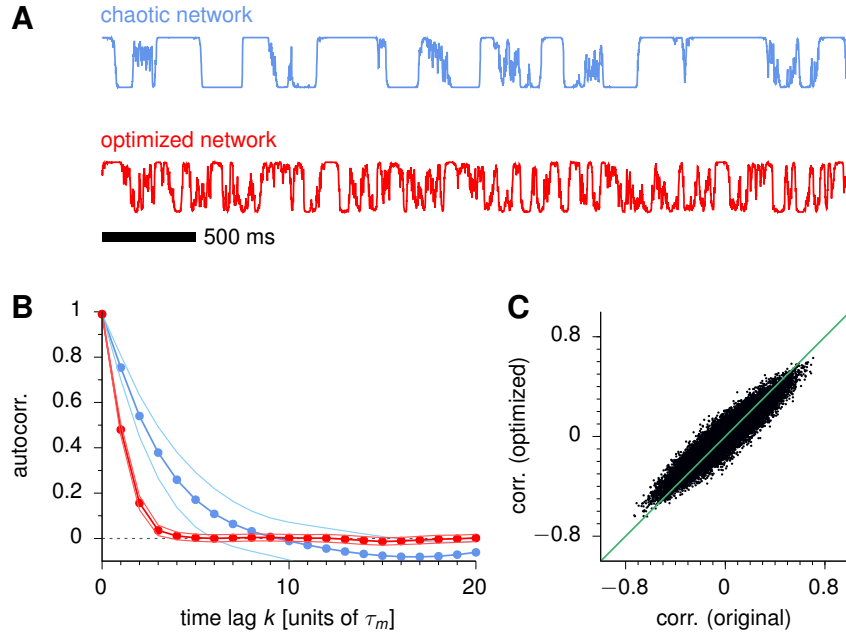
Figure 1: Speeding up sampling for a non-Gaussian distribution in a nonlinear system. **(A)** Sample activity traces for a single unit in the nonlinear, chaotic network (blue), and in the optimized network (red). **(B)** Firing rate autocorrelation $\langle r_i(t)r_i(t + k\tau_{\mathrm{m}})\rangle_t$ (with r_i transformed to z-score) averaged across neurons (flanking lines denote $\pm$ one std.), for the two networks. **(C)** Pairwise activity correlations $\langle r_i(t)r_j(t)\rangle_t$ (with the r_k's transformed to z-scores) in the nonlinear, chaotic network (x-axis) vs. those in the optimized network (y-axis).

References

[1] C. W. Gardiner. *Handbook of stochastic methods: for physics, chemistry, and the natural sciences*. Berlin: Springer, 1985.

[2] J. Vanbiervliet, B. Vandereycken, W. Michiels, S. Vandewalle, and M. Diehl. The smoothed spectral abscissa for robust stability optimization. *SIAM Journal on Optimization*, 20:156–171, 2009.

[3] R. H. Bartels and G. W. Stewart. Solution of the matrix equation AX+XB=C. *Communications of the ACM*, 15:820–826, 1972.

[4] R. Neal. MCMC using Hamiltonian dynamics. *Handbook of MCMC*, pages 113–162, 2011.

[5] J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34:A1460–A1487, 2012.