Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference

Rodrigo Echeveste[1], Laurence Aitchison[1], Guillaume Hennequin[1,*], and Máté Lengyel[1,2*]

[1] *Computational and Biological Learning Lab, Dept. of Engineering, University of Cambridge, Cambridge, UK*
[2] *Department of Cognitive Science, Central European University, 7 Oktober 6. utca, Budapest H-1051, Hungary*

[*] *equal contribution*

Sensory cortices display a suite of ubiquitous dynamical features, such as ongoing noise variability, transient overshoots, and oscillations, that have so far escaped a common, principled theoretical account. We developed a unifying model for these phenomena by training a recurrent excitatory–inhibitory neural circuit model of a visual cortical hypercolumn to perform sampling-based probabilistic inference. The optimized network displayed several key biological properties, including divisive normalization, as well as stimulus-modulated noise variability, inhibition-dominated transients at stimulus onset, and strong gamma oscillations. These dynamical features had distinct functional roles in speeding up inferences and made predictions that we confirmed in novel analyses of awake monkey recordings. Our results suggest that the basic motifs of cortical dynamics emerge as a consequence of the efficient implementation of the same computational function—fast sampling-based inference—and predict further properties of these motifs that can be tested in future experiments.

The dynamics of sensory cortices exhibit a set of features that appear ubiquitously across species and experimental conditions. Responses vary over time and across trials even when the same static stimulus is presented[1], and these intrinsic variations have both systematic and seemingly random components (so-called noise variability). The most prominent systematic patterns of neural activity are strong, inhibition-dominated transients at stimulus onset[2] (or, equivalently, strong adaptation following stimulus onset), and stimulus-dependent population oscillations in the gamma band (20–80 Hz)[3,4]. The extent and pattern of noise variability is also stimulus-dependent: variability is quenched at stimulus onset[1], decreasing gradually with stimulus contrast in the primary visual cortex (V1)[5,6], and is further modulated by the content of the stimulus, e.g. the orientation or direction of drifting gratings for cells in V1 or in the middle temporal visual area (MT)[7,8].

While the mechanisms giving rise to these dynamical phenomena are increasingly well understood[8–10], their functional significance remains largely unknown and controversial, with several candidate functional roles having been proposed for each of them. For example, cortical gamma oscillations have been suggested to be a substrate for binding different sources of information about a feature (known as binding by synchrony)[11,12], to mediate information routing (communication by synchrony)[13], or to enable a temporal code of spikes relative to the oscillation phase[14]. Additionally, transient overshoots have been proposed to carry novelty or prediction error signals[15]. Noise variability, when considered to have any function at all rather than being mere nuisance[16], has been argued to bear signatures of specific probabilistic computations in the cortex[6,17,18]. However, it is unclear whether these explanations can be reconciled, as each of them only accounts for select aspects of the data, and has been challenged by alternative accounts[3,19–21].

Here, we present a unifying model in which all of these dynamical phenomena emerge as a consequence of the efficient implementation of the same computational function: probabilistic inference. Probabilistic inference provides a principled solution for the fundamental requirement of perception to continually fuse partial and noisy information from multiple sources (including multiple sensory cues, modalities, and forms of memory)[22,23]. Formally, the result of this fusion is a posterior probability distribution, i.e. the probability that "hidden" quantities that are not directly accessible to the brain may take any particular configuration given information that is directly available to our senses. Behavioral evidence in several domains, including near-optimal performance in multi-sensory integration, decision making, motor control, and learning suggests that the

brain represents posterior distributions (at least approximately)[24]. There have also been several proposals for how the neural responses of sensory cortical populations may implement these probabilistic representations[6,17,25]. While these models successfully explained important aspects of stationary response distributions (e.g. tuning curves, Fano factors, noise correlations), they have so far fallen short of accounting for the rich intrinsic dynamics of sensory cortical areas.

To bring together dynamics (cortical-like activity patterns) and function (computing posterior distributions) in a principled manner, we optimized a recurrent neural network obeying a set of fundamental biological constraints (separation of excitatory and inhibitory cells, and non-saturating firing rates in the physiological regime) for performing probabilistic inference. The network received visual stimuli as inputs and was required to represent in its responses the posterior distributions that would be inferred for the same stimuli by a Bayesian ideal observer. Specifically, network dynamics had to produce activities which represented statistical samples from the posterior distribution. This required the network to modulate not only the mean but also the variability of its responses in a stimulus-dependent manner. Such a sampling-based probabilistic representation of uncertainty has been shown to account for stimulus- and task-dependent aspects of *stationary* variability in V1[6,26,27], thus offering a promising computational target for a network used to study the temporal *dynamics* of V1 responses.

The optimized neural circuit exhibited a number of appealing computational and dynamical features. Computationally, after training on a reduced stimulus set, the network exhibited strong forms of generalization by producing near-optimal response distributions to novel inputs which required qualitatively different responses. Furthermore, the network discovered out-of-equilibrium dynamics, a strategy currently employed by modern machine learning algorithms to produce samples that become statistically independent on short timescales[28]. Biologically, the circuit achieved divisive normalization of its outputs and displayed marked transients at stimulus onset, as well as strong gamma oscillations, such that both the magnitude of transients and the frequency of gamma oscillations scaled with stimulus contrast. Crucially, these dynamical phenomena did not emerge in a control network trained to match posterior mean responses only, without a need to modulate the variability of its responses. Indeed, further analyses of transients and oscillations in the optimized network revealed distinct functional roles for them. Our results allowed us to predict novel properties for cortical dynamics, such as the stimulus-tuning of onset transients, which we confirmed by performing novel analyses of published V1 recordings in the awake monkey[29]. In addition, our model also made further specific predictions about the stimulus tuning of excitatory–inhibitory lags and the distribution of gamma power across the different modes of network dynamics, which can be readily tested in future experiments. In summary, we constructed the first biologically constrained recurrent neural network performing sampling-based probabilistic inference that explained a plethora of electrophysiological observations in sensory cortices. Our model thus provides a unifying theoretical account of the basic motifs of sensory cortical dynamics.

Results

Optimizing a recurrent neural circuit for performing inference

To study neural circuit dynamics implementing probabilistic inference, we used a novel combination of two well-established, though hitherto unrelated, principled computational approaches. First, we used a probabilistic ideal observer model to specify the computational goal of perceptual inference in a simplified visual task. Performing inference requires an internal model that encapsulates one's assumptions about how the inputs to be processed have been generated by the environment. For this, we adopted the Gaussian scale mixture (GSM) model (Fig. 1a, Online Methods), a generative model that has been shown to capture the statistics of natural image patches[30]. Conversely, inference under the GSM model has been shown to account for behavioral and neural data (for stationary responses) in visual perception[6,31,32]. The GSM model assumes that an image patch is generated as a linear combination of oriented Gabor filter-like visual features ("projective fields" that only differed in their orientation, Fig. S1), each present with a different intensity, further scaled by a single global "contrast" variable. The ideal observer was obtained by a Bayesian inversion of this model (Online Methods). Thus, for every image patch taken as sensory input, the ideal observer yielded a high-dimensional posterior distribution quantifying the probability that any particular joint combination of intensities for the Gabor-like projective fields may have generated the input (Fig. 1b).

Second, to model cortical circuit dynamics, we used a canonical, rate-based stochastic recurrent neural network model, the stochastic variant of the stabilized supralinear network (SSN)[8,33,34]

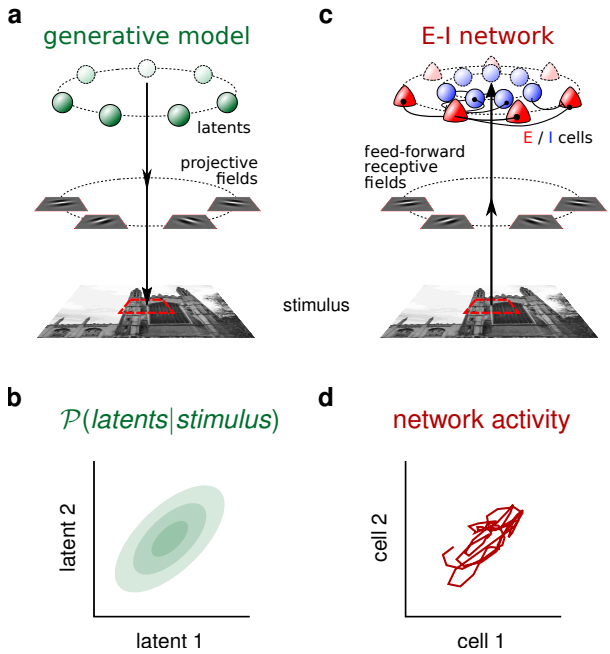

Fig. 1. The statistical generative model, and the corresponding neural circuit implementing sampling-based probabilistic inference. (a) Sketch of the Gaussian scale mixture (GSM) generative model. An image patch is constructed as a linear combination of a fixed set of localised, oriented, Gabor filter-like features (projective fields, differing only in their orientations, uniformly spread between $-90°$ and $90°$), with stimulus-specific feature intensities (latent variables) drawn from a multivariate Gaussian distribution. The resulting image is scaled by a global contrast variable and corrupted by noise (not shown). **(b)** 2-dimensional projection of the posterior distribution over latent variables given a visual stimulus, computed by the Bayesian ideal observer under the generative model. **(c)** An excitatory–inhibitory (E–I) neural network receiving an image as an input filtered by feed-forward receptive fields identical to the projective fields of the generative model in **a**. The activity of each E cell represents the value of one latent variable in the generative model. **(d)** 2-dimensional projection of the neural state space, with the responses of the two E cells corresponding to the latent variables shown in **b**. Neural response trajectory samples from the corresponding posterior distribution over time given the same stimulus.

(Online Methods). The network was constrained to exhibit some basic biological features that have been shown to have fundamental consequences for cortical dynamics: the presence of separate but inter-connected excitatory (E) and inhibitory (I) populations of neurons (Fig. 1c), supralinear (expansive) input/output functions[33,35], and some finite and stimulus-independent process noise (incorporating intrinsic and extrinsic forms of neural variability)[8,36].

We trained this network to perform sampling-based inference under the GSM. For this, the activity of each excitatory cell at any point in time was taken to represent a possible level of intensity of the corresponding projective field of the GSM model. The activities of inhibitory neurons were treated as auxiliary variables which were not explicitly constrained by the computational objective. The network was optimized to produce stationary distributions of (excitatory) neural activities that matched the posteriors computed by the GSM-based ideal observer (up to second-order statistics) for every image in a small training set (Online Methods). In other words, for each stimulus, the network was required to perform inference by using its stochastic dynamics to sample different parts of its state space over time with a frequency that was determined by the posterior distribution corresponding to the same stimulus (Fig. 1d). Critically, as process noise in the network was stimulus-independent, the network had to use its recurrent dynamics to shape this variability appropriately for matching the target posteriors for each input. Moreover, besides requiring a match between GSM posterior statistics and the network's stationary response distribution, the training objective also included terms encouraging fast circuit dynamics (Supplementary Material). Thus, the network had to generate fast fluctuations with the correct stimulus-dependent patterns of trial-by-trial mean and covariance.

Inference and generalization in the optimized network

In line with with neural recordings, activity in the optimized network was highly variable across time and trials, both in response to low-contrast images (Fig. 2a, top), and for higher-contrast stimuli (Fig. 2a, bottom). Critically, the distributions of neural responses at the five training stimuli (the same image at five different contrast levels; Fig. 2b left) closely matched the corresponding GSM posteriors (Fig. 2b–d, compare red to

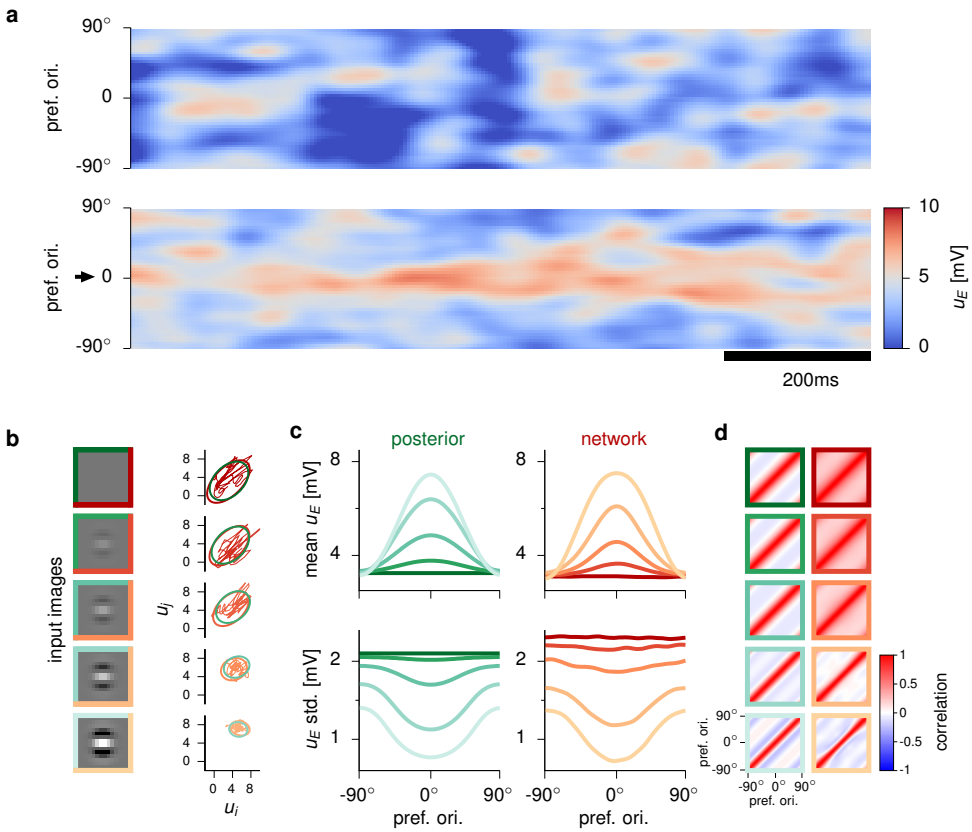

Fig. 2. Inference and responses in the optimized network. (a) Sample population activity of excitatory (E) cell membrane potentials u_E (y-axis, ordered by their preferred orientation) at zero (top) and high (bottom) contrast. The high contrast image has a dominant orientation at $0°$ (arrow). **(b)** Left: input images in the training set (frame color indicates contrast level). Right: covariance ellipses (2 standard deviations) of the ideal observer's posterior distributions (green) and of the network's corresponding response distributions (red). Red trajectories show sample 500 ms-sequences of activities in the network. Projections for two representative latent variables / E cells are shown, with preferred orientations $42°$ (u_i) and $16°$ (u_j). **(c)** Mean (top) and standard deviation (bottom) of latent variables under the ideal observer's posterior distribution (left, green) and of E cell membrane potentials u_E in the network (right, red), ordered by their preferred orientation, for each image in the training set. **(d)** Correlation matrices of the ideal observer's posterior distributions (left, green) and the network's response distributions (right, red). Line colors in **c** and frame colors in **d** correspond to different contrast levels, same colors as image frames in **b**.

green). Specifically, the mean activity of neurons increased while the variability of their responses decreased with contrast as well as with the match between stimulus orientation and their preferred orientation. This was consistent with the behaviour of the moments of the GSM posterior (Fig. 2c, and circles in Fig. 3a). Thus, the network had been trained successfully to perform sampling-based inference on these input images.

We also tested the capacity of the network to compute the appropriate posterior distribution for stimuli outside the training set. First, we employed the same image that was used to construct the training set but presented it at novel, intermediate contrast levels. The mean and variability of network responses smoothly interpolated between the corresponding target moments, closely following the behaviour of the GSM posterior (Fig. 3a, solid curves between circles).

Next, we presented 500 entirely novel, randomly generated images from the GSM to the network (Online Methods). Overall, we found a similar degree of match between network responses and GSM posteriors (Fig. 3b, orange), both in their means (Fig. 3b, top) and covariances (Fig. 3b, bottom), as for the training set (Fig. 3b, purple). Critically, while the inputs of the training set included a single dominant orientation, many test images had a more complex structure, with more than one dominant orientation (Fig. 3c, first column). Consequently, the corresponding GSM posteriors that the network was required to match became qualitatively different, such that both the mean activity profiles across the population (Fig. 3c, 2nd column) and the principal components (PCs) of the noise covariances (Fig. 3c, remaining columns) became multimodal and highly dependent on the stimulus (Fig. 3c, green; compare across rows). The network was able to match the required GSM posteriors with high accuracy even in these challenging cases (Fig. 3c, red). Thus, the optimized network performed approximate Bayesian inference over a wide array

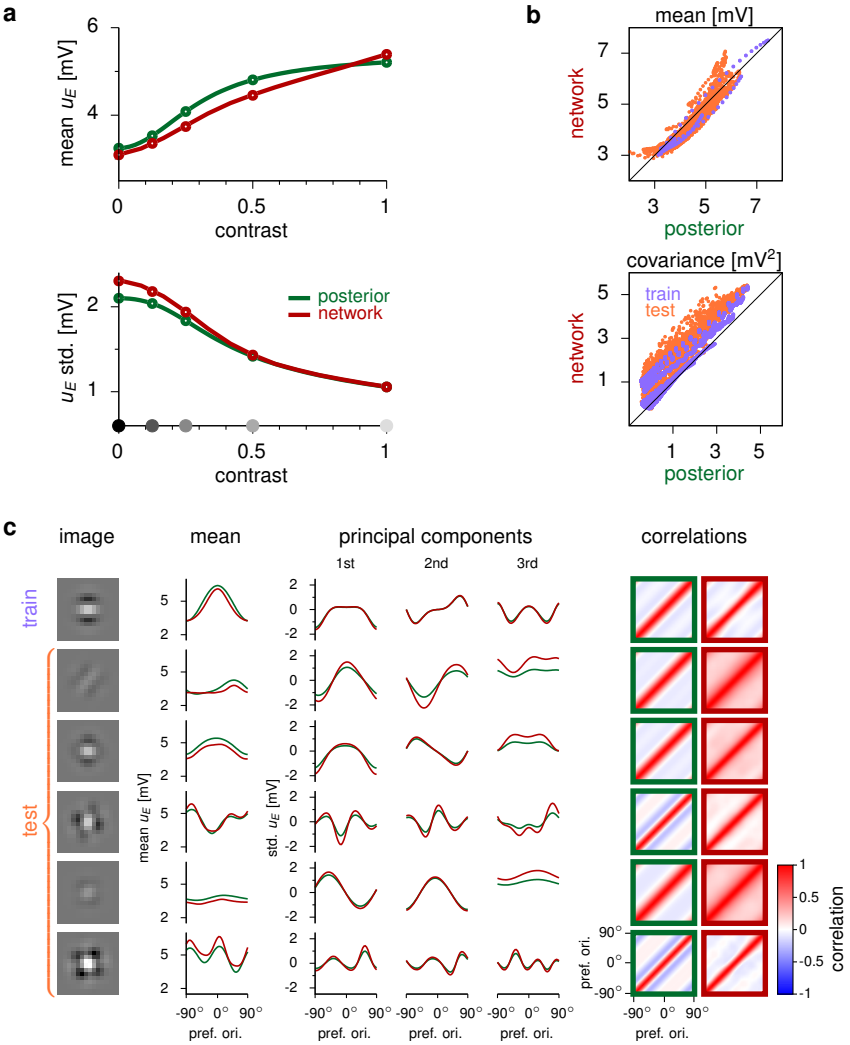

Fig. 3. Generalization in the optimized network. (a) Mean and standard deviation of latent variables (green) and network responses (red) averaged over the population, as a function of contrast. Circles, and gray dots on x-axis indicate training contrast levels. The network correctly generalizes over untrained contrast levels (segments between circles). **(b)** Mean (top) and covariance (bottom) during network activity (y-axis) versus under the posterior (x-axis). Each dot corresponds to the response of an individual cell (top) or cell-pair (bottom) to one of the trained stimuli (purple) or one of the novel, untrained stimuli in the test set (orange). **(c)** Examples of generalization in the network. Each row corresponds to a different input image, and the corresponding statistical moments of latent variables under the GSM posterior (green) and responses in the network (red). As a reference, the top row shows one of the training images. The bottom five rows show generalization to novel test images. Left: example input images. Middle: GSM (green) and network means (red), and the first three principal components of the GSM covariance, scaled by the square root of the variance they explain of the GSM posterior (green) and of the network covariance (red). Right: Correlation matrices of the ideal observer's posterior distributions (left, green frames) and the network's response distributions (right, red frames).

of images by always sampling (approximately) from the appropriate, stimulus-dependent high-dimensional posterior distribution of the GSM-based ideal observer.

The optimized network performs fast sampling

Under sampling-based inference, the time it takes to accurately represent the posterior distribution by collecting successive samples is directly proportional to the timescale over which these samples are correlated[37]. For example, if neural responses were correlated on a 100 ms timescale, it would take on the order of a second to obtain 10 independent samples. Thus, achieving a minimal overall correlation of neural responses on short timescales is desirable for the brain to form percepts and make decisions on typical behavioral timescales. In our optimized network, noise variability generated new, independent samples every few tens of milliseconds across all contrast levels, as evidenced by fast-decaying membrane potential autocorrelations (Fig. 4a, colored curves). These timescales were similar to the timescales of activity fluctuations observed in sensory cortical areas[18,38]. In fact, they were faster than what would have been expected in a

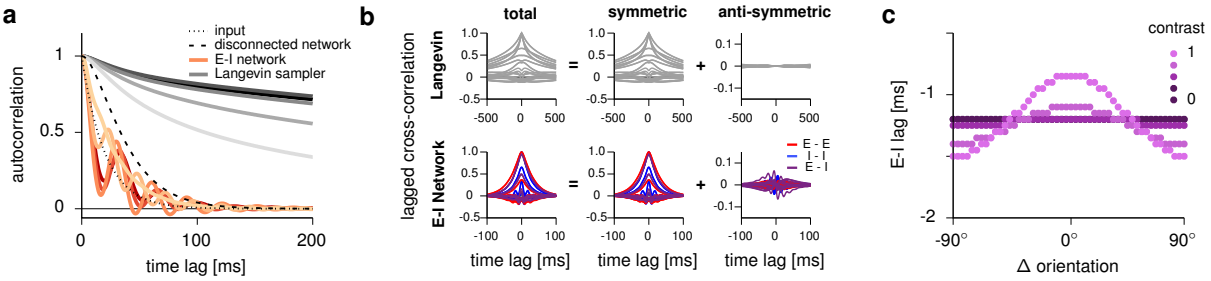

Fig. 4. Temporal correlations. (a) Membrane potential auto-correlations (population average) in the network for increasing levels of stimulus contrast (from dark to pale red; same colors as in Fig. 2b–d). The auto-correlation of a purely feed-forward network (with the same process noise) is shown for comparison (dashed black line), together with those of the process noise (dotted black line), and a collection of networks implementing Langevin sampling at each contrast level (from dark to light gray). **(b)** Lagged cross-correlation (left) in the Langevin sampler (top) and in the optimized E–I network (bottom), decomposed into temporally symmetric (middle) and anti-symmetric components (right). Each line corresponds to a different cell pair, color encodes identity of participating cells (E or I, note that there is no separation of E and I cells in the Langevin networks). Substantial anti-symmetric components between all types of cell pairs indicate out-of-equilibrium dynamics, i.e. violation of (statistical) detailed balance in the E–I network (top vs. bottom, right). **(c)** Lag between total E and I inputs to each E cell, as a function of stimulus orientation (relative to preferred orientation) at different contrast levels (colors).

disconnected network with the same membrane and input time constants (Fig. 4a, dashed curve), and even close to the theoretical limit of a network of infinitely fast neurons in which sampling speed is solely limited by the input time constant (Fig. 4a, dotted curve).

To understand the algorithmic strategy employed by the optimized network to perform fast inference, we compared its dynamics to a popular machine learning algorithm known as Langevin sampling (Online Methods), which has previously been suggested as being implemented by recurrent neural networks (without a separation of E and I cells)[39,40]. For each input, fluctuations in the optimized network were consistently an order of magnitude faster than in the corresponding Langevin network (Fig. 4a, gray curves). While Langevin sampling relies on time-reversible dynamics—i.e. any time series of responses is as probable as its time-reversed counterpart, and thus cross correlations in the network are purely temporally symmetric (Fig. 4b, top)—some of the best-performing, modern machine learning algorithms achieve faster sampling by breaking time-reversibility through out-of-equilibrium dynamics[28]. Remarkably, our optimized network also displayed a marked departure from time-reversibility, as evidenced by a strong asymmetric component in its pairwise cross-correlograms (Fig. 4b, bottom).

The irreversibility of network dynamics implied sequentiality in the activation of particular pairs of neurons. In particular, we found that I cells typically lagged behind E cells. Moreover, for any cell, its total inhibitory input tended to also lag behind its overall excitatory input (Fig. 4c), consistent with known electrophysiology[41]. Interestingly, this lag was smaller for cells that were most strongly driven by the stimulus, and this modulation became stronger with increasing contrast. These form testable predictions of our model.

Cortical-like dynamics in the optimized circuit

Having established that our network fulfilled its function of representing posterior distributions via sampling, we compared the dynamics it used to achieve this with known physiological properties of V1. First, we confirmed that, as expected for a network whose stationary membrane potential response distributions represent GSM posteriors[6], overall spike count statistics in the network (computed from firing rates assuming a doubly stochastic spike emission process, see Online Methods) behaved realistically. Firing rates in the model had a physiologically realistic dynamic range and were tuned to stimulus orientation in a similar way to neurons in macaque V1 (Fig. 5a, left-middle; Ref. 8, analysis of data recorded by Ref. 29). Furthermore, the quenching of variability with increasing contrast that we noted earlier (Fig. 2d, bottom), gave rise to the quenching of spike count variability, as quantified by the Fano factor (Supplementary Material), weaker than but in qualitative agreement with experimental data (cf. Fig. 5b, left-middle). Fano factor suppression was also stronger at the cell's preferred orientation, as in awake monkey V1 (cf. Fig. 5b, left-middle). Moreover, stationary responses in the network exhibited clear signatures of divisive normalization (Fig. S4), a canonical operation of cortical circuits[42].

Next, we focused on the dynamical properties of the optimized network. Recall that the optimization procedure only constrained the stationary response distributions and encouraged general temporal decorrelation on fast timescales, but did

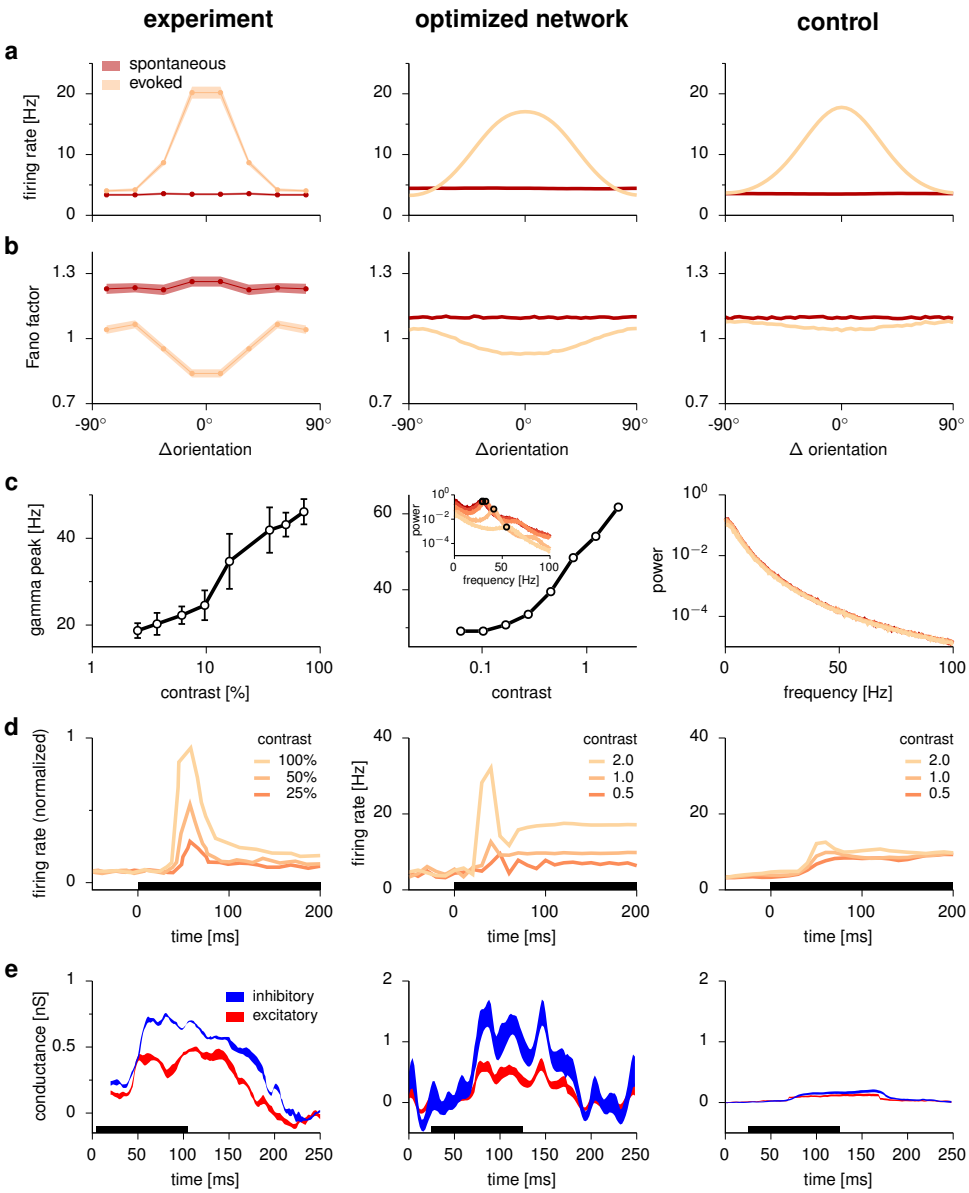

Fig. 5. Cortical-like dynamics in the optimized circuit. Left: experimental data; middle: optimized network; right: control network trained to modulate its mean responses but not its variability. **(a)–(b)** Mean firing rate (**a**) and Fano factor (**b**) of neurons as a function of stimulus orientation (relative to preferred orientation) during spontaneous (dark red) and evoked activity (light orange). Experimental results show mean ± s.e.m. **(c)** Peak gamma frequency in the local field potential (LFP) power spectrum as a function of contrast. Inset for the optimized network and main panel for the control network show LFP power spectra at different contrast levels (colors as in Figs. 2 and 4). Note that no dependence of gamma frequency is shown for the control network as there are no discernible gamma peaks in the power spectra. **(d)** Average rate response around stimulus onset at different contrast levels (colors). **(e)** Excitatory and inhibitory conductance (mean ± s.e.m., relative to baseline, see Supplementary Material for details) during a transient stimulus response. Black bars in **d**–**e** show stimulus period. Panels **a**–**b** reproduce analyses from Ref. 8 of data from Ref. 29 (awake macaque V1). Experimental data in **c** was reproduced from Ref. 4 (awake macaque V1), in **d** reproduced from Ref. 3 (awake macaque V1), and in **e** reproduced from Ref. 2 (awake mouse V1).

not otherwise prescribe any specific dynamics. Nevertheless, we found that the network exhibited a number of experimentally observable features of cortical dynamics. Specifically, the optimized circuit displayed strong gamma oscillations, with a peak frequency increasing with contrast, consistent with V1 recordings in the awake monkey[3,4] (Fig. 5c, left-middle). Moreover, oscillations disappeared entirely when we voltage-clamped each cell in the E population to its mean voltage corresponding to the input (Fig. S8), while analogous voltage-clamp of the inhibitory population led to unstable dynamics (not shown). Thus, in the optimized network, gamma oscillations arose from interactions between E and I cells (i.e. the so-called "PING" mechanism[21]).

The network also showed strong transient responses such that average population rates had a marked contrast-dependent overshoot at stimulus

onset, consistent with recordings in V1[3] (Fig. 5d, left-middle). Finally, we found that inhibition transiently dominated over excitation during stimulus presentation, as in the V1 of the awake mouse[2] (Fig. 5g, left-middle).

To discern how much of these dynamical properties were unavoidable consequences of the E–I architecture and general stimulus tuning of our network, we optimized a 'control' network in which single cell parameters (time constants and firing rate nonlinearities), overall network architecture, and the target mean activities were the same as those used for the original network (Online Methods). Critically, the control network was trained to match the mean of the posterior distributions, but unlike the original network, was not required to modulate its variability. Despite clear stimulus-dependent modulations in mean responses (as required by training; Fig. 5a, right), the control network exhibited only minimal modulations of both membrane potential variability Fig. S5, and Fano factors (Fig. 5b, right). This indicated that modulations of response variability seen in the original network, which are a hallmark of a sampling-based probabilistic inference strategy[6], were not just a generic by-product of non-linear E–I dynamics[34]. Critically, neither gamma oscillations, nor marked inhibition dominated transients emerged in the control network (Fig. 5c-e, right). Oscillations were also absent in a second control network specifically optimized to modulate its mean firing rates while keeping its Fano factors constant (Supplementary Material, Figs. S6 and S7), as would be required by other, non sampling-based probabilistic representations[17]. These results showed that the dynamical features observed in the original network emerged as a consequence of the specific computation for which it was optimized.

Oscillations improve mixing time

To isolate the potential functional benefits of oscillations, we studied a simplified variant of these dynamics with a one-dimensional response (analogous to the time-varying response of a single neuron) fully characterized by its mean and variance (determined by the target distribution), and by its autocorrelogram which we could manipulate directly (Online Methods; Fig. 6a–b). We noted that the envelope of the autocorrelogram in the original network would be ultimately constrained by the time constant of the process noise (Fig. 6a, inset, black dotted; see also Fig. 4). We therefore constrained the autocorrelogram of our one-dimensional surrogate system to have the same envelope as the full network, and only varied its degree of "oscillatoriness" (Fig. 6a, blue, orange, red). These oscillations were able to substantially reduce the area under the autocorrelogram, which we could show analytically directly controlled sampling performance (Online Methods). This could also be seen by the speed with which the distribution of responses measured over a finite time window converged to the true stationary distribution (Fig. 6b). Interestingly, when the oscillations in the response were sufficiently strong to induce negative-going lobes in the autocorrelogram (Fig. 6a, red), sampling accuracy remained better than in the non-oscillatory system even for long sampling times (Fig. 6b, red vs. blue; Supplementary Material). In fact, although the strongly oscillating system exhibited longer correlation decay times than the theoretical minimum (Fig. 6a, inset; envelope of red vs. black dotted line), it achieved even better asymptotic performance than a non-oscillating system operating at that theoretical minimum (Fig. 6b, red vs. black dotted; for sampling times greater than approximately 20 ms).

Importantly, oscillations will only decrease the area under the autocorrelogram, and thus be useful for sampling speed (as long as at least one oscillation cycle fits under the envelope). This requires their period to be sufficiently shorter than the time constant of their envelope (approximately 35 ms), i.e. their frequency to be 30 Hz or higher. Indeed, the lowest frequency we observed in the network was about 30 Hz (Fig. 5c, middle). We next studied the organization of gamma oscillations in the multidimensional responses of the full network. Mathematical analyses have suggested that fast convergence to a high-dimensional target distribution requires temporally irreversible dynamics (such as those exhibited during oscillations)[40]. Specifically, we were able to show that maximal sampling speed is achieved when larger response variance is associated with higher oscillation frequency (Supplementary Material). As we showed above, an inherent property of our network (and of the cortex) was that increasing contrast quenched variability (Fig. 2b-c, 3a, 5b). Therefore, the combination of these two results explained why in our network (as in experimental data), increasing contrast also led to an increase in the frequency of gamma oscillations (Fig. 5c).

Given process noise with a finite time constant, our analyses also predicted that oscillations in an efficiently sampling network should be predominantly expressed along the top PCs of the stationary covariance for each stimulus, i.e. the (stimulus-dependent) network-wide activity patterns that were responsible for most of the overall response variability (Supplementary Material).

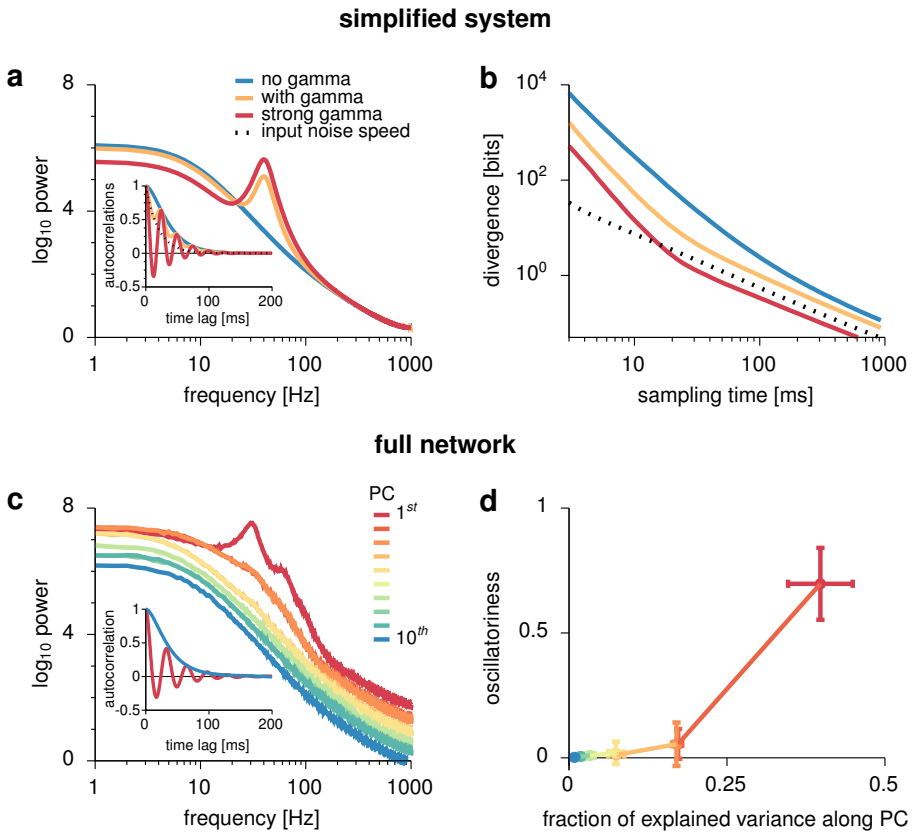

Fig. 6. Oscillations improve mixing time. (a)–(b) Analysis of a family of simplified one-dimensional systems with time-varying scalar responses. **(a)** Power spectra of three simplified systems (colors) with identical response mean and variance but different degrees of oscillatoriness. Black dotted line represents the autocorrelation of the process noise. Inset: Autocorrelation functions. **(b)** Divergence between the distribution estimated from a finite sampling time (x-axis) and the true stationary distribution for the three systems (colors as in **a**). **(c)–(d)** Sampling in the full E–I network. **(c)** Power spectra of the network's neural activity along the directions of the principal components (PCs) of its stationary response distribution, ordered by PC rank (colors). Inset: autocorrelation of neural activity along the directions of the 1st and 10th PCs (colors as in main plot). **(d)** Oscillatoriness of the autocorrelogram along each principal component (colors as in **c**) as a function of the fraction of the total variance of responses they capture. Note that the measure of oscillatoriness is defined to be invariant to the overall magnitude of fluctuations (Online Methods).

This was indeed apparent in the associated power spectra (Fig. 6c), and autocorrelograms that even showed negative-going lobes (Fig. 6c, inset). Specifically, there was a positive relationship between a measure of oscillatoriness (invariant w.r.t. the overall magnitude of fluctuations) along successive PCs and the fraction of variance explained by them (Fig. 6d; Online Methods). As the response patterns corresponding to the top PCs strongly depended on the input to the network (Fig. 3c), this also meant that the network adaptively expressed temporal oscillations along different, stimulus-dependent spatial patterns. In sum, the network used non-trivial temporal dynamics, in the form of contrast-dependent, pattern-selective gamma oscillations, to ensure that even short segments of its activity were maximally representative of the posterior distribution it represented for each stimulus.

Transients support continual inference

While oscillations increased the effective speed of sampling once network responses were representative of the target distribution (also known as "mixing speed"[37]), we found that transients in our network mitigated the other main temporal constraint of sampling: the so-called "burn-in" time it takes for responses to become representative[37]. We observed that, in line with experimental data, during stimulus onset, neural responses tended to overshoot the corresponding stationary response levels (Fig. 5d, 7a). One might naively expect such transients to be detrimental for representing a distribution as they clearly deviate from the target (represented by the steady-state responses). However, in a realistic setting with a changing environment, distributions need to be represented continually, without waiting for the system to achieve steady state. Thus, we considered how a moving decoder of neural responses

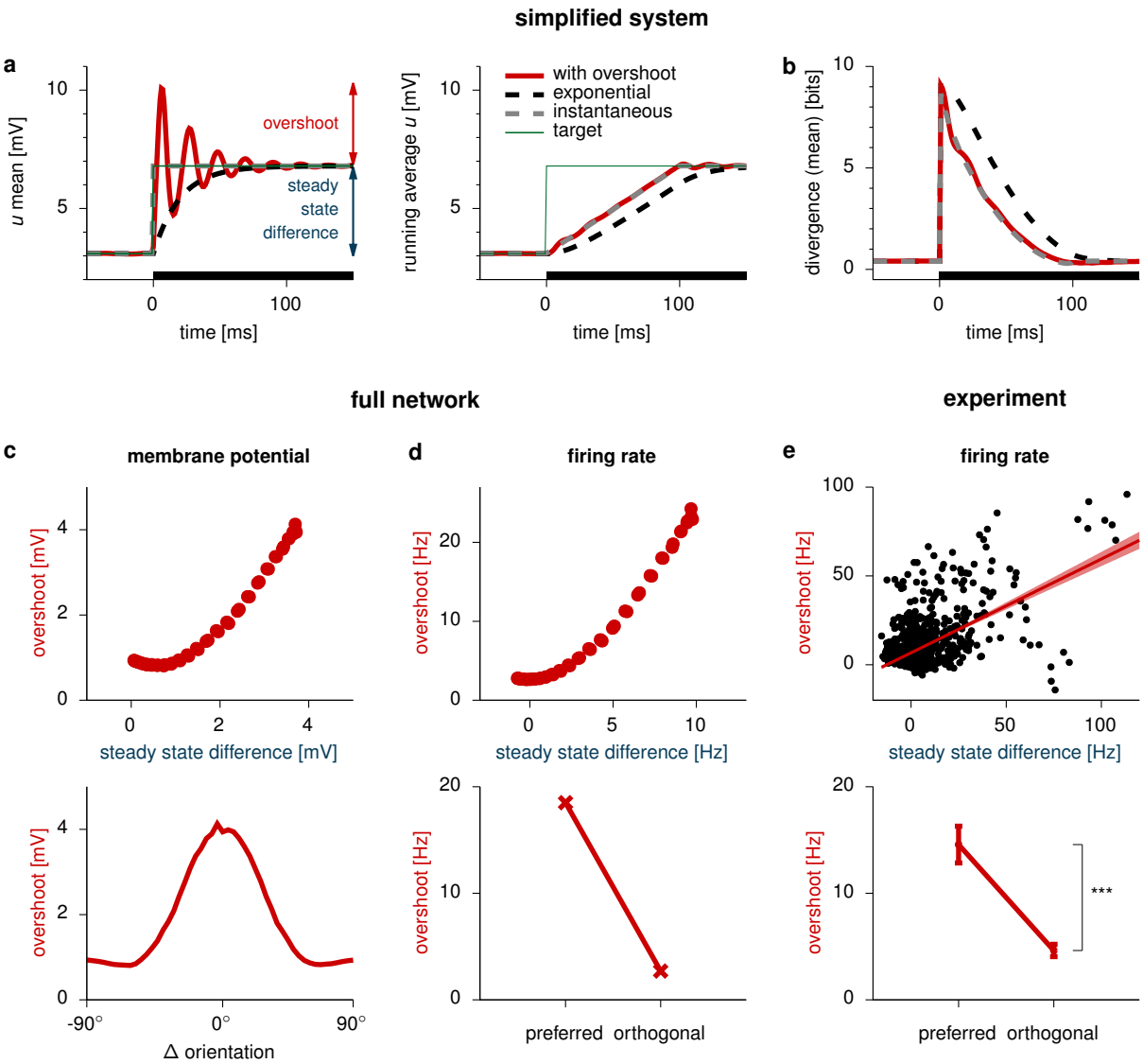

Fig. 7. Transients support continual inference. (a)–(b) Analysis of a family of simplified one-dimensional systems with time-varying scalar responses. **(a)** Temporal evolution of the mean (left) membrane potential (u_E), and its running average (right), in the three simplified systems that had identical response autocorrelations (matched to neural autocorrelations in the full network, Fig. S9, cf. Fig. 6) but different time dependent means (shown here) and variances (Fig. S9). Thin green line shows the time-varying target mean. Three systems are compared: a cell displaying a marked transient overshoot (as seen in the network, red), a cell whose mean and standard deviation converged exponentially with a timescale given by the membrane time constant (dashed black), and a hypothetical instantaneous system that exhibited a discrete jump in its mean (and variance, dashed grey). **(b)** Divergence between the target distribution at a given point in time and the distribution implied by the neural activity sampled in the preceding 100 ms, for each of the three systems (colors as in **a**). The mean-dependent term of the divergence is shown here, which depends on the difference between the target mean and the running average of samples (shown in **a**, right) (see Fig. S9 for the full divergence). Black bars in **a**–**b** show stimulus period. **(c)** Top: overshoot magnitude versus steady state difference in membrane potentials (see **a** for legend). Each dot corresponds to the response of one cell to one particular stimulus. Bottom: overshoot magnitude as a function of stimulus orientation (relative to preferred orientation). **(d)** Top: same as **c**, top, for firing rates. Bottom: average rate overshoot across cells whose preferred orientation is aligned with the stimulus ($0 \pm 30°$), or near-orthogonal to it ($90 \pm 30°$). **(e)** Re-analyzed experimental recordings from awake macaque V1[29]. Top: overshoot magnitude versus steady state difference, as in **d**, top. Red line shows linear regression ($\pm 95\%$ confidence bands), ***: $p < 0.001$ ($n = 1280$ cell-stimulus pairs). Bottom: average (± 1 s.e.m) overshoot magnitude for preferred and orthogonal stimuli, as in **d**, bottom, ***: $p < 0.001$ ($n = 864$ cell-stimulus pairs).

over a finite trailing time window approximated the target.

To isolate the potential functional benefits of transients, beyond those of oscillations we analyzed before, we once again studied the same simplified, 1-dimensional variant of network dynamics (Online Methods; Supplementary Material). In this simplified system, we fixed the autocorrelogram as well as the before- and after-stimulus onset steady-state means and variances to that of the full network, and studied the effects of different ways in which a network could transition be-

tween the two steady states (Fig. 7a). We considered three possibilities: 1. as an upper bound on performance, which is clearly unrealistic, a system that instantaneously switched between the two steady-states (Fig. 7a, gray dashed); 2. a system that exponentially approached the new steady state with the characteristic time constant of the cells in the network, thus lacking overshoots (Fig. 7a, black dashed); and 3. a system that underwent the same overshoots as the original network (Fig. 7a, red). We found that the system with overshoots performed similarly to the instantaneous system and generated samples that allowed a substantially more accurate estimate of the target mean than the system that decayed exponentially (without overshoots) to the new steady-state (Fig. 7a–b). (These results extended qualitatively to the case when the match in the full distributions was considered, Supplementary Material, Fig. S9.) This was because without overshoots at stimulus onset, responses were still sampling from the distribution corresponding to the baseline input, and so including them in the estimation of the new stimulus-related mean inevitably biased the estimate to be too low. This bias was largely offset by the overshoot. Indeed, we were able to show analytically that the optimal way to compensate for this bias was to express transient overshoots at stimulus onset (followed by damped oscillations around the new steady-state value, Supplementary Material). The intuition for this was that the running averaging of responses formally corresponded to a temporal convolution, and so the optimal response was the *de*convolution of the target with the averaging (box-car) kernel. The deconvolution of a step function under basic smoothness constraints yielded similar transients to those that we observed in the network (Fig. S9c).

The hypothesis of increased sampling accuracy by transient compensation made a distinct prediction (which is also supported our mathematical analysis, Supplementary Material): transient overshoots should scale with the change in steady state responses. Indeed, our network exhibited this effect (Fig. 7c, top), which also resulted in transients being orientation tuned, reflecting the tuning of stationary responses (Fig. 7c, bottom). These results also held for firing rates (Fig. 7d). While stimulus-onset transients have been widely observed[3,4], previous reports did not analyze their stimulus tuning. Therefore, we conducted our own analyses of a previously published dataset of V1 responses in the awake monkey[29]. We found that, in line with the predictions of the model, the size of overshoots scaled with the change in stationary responses ($n = 1280$, coefficient of determination $R^2 \simeq 0.33$, $p < 0.001$, Fig. 7e, top; these results were robust to excluding the outliers with high firing rates: $n = 1263$, $R^2 \simeq 0.27$, $p < 0.001$, Fig. S9d), and were orientation tuned (Fig. 7e, bottom).

Discussion

We have shown that a canonical neural network model[8,33] produces cortical-like dynamics when optimized to perform sampling-based probabilistic inference, but not when optimized to perform a non-probabilistic objective, or a non sampling-based probabilistic objective. Importantly, these dynamics played well-defined functional roles in performing inference rather than being mere side products of the particular biological constraints or optimization approach we adopted.

The Gaussian scale mixture model and the stochastic stabilized supralinear network

It was not trivial *a priori* that the canonical network model we used (the stochastic SSN), embodying a set of biologically relevant constraints, would be able to modulate its responses as necessary for successful sampling-based inference under a canonical generative model of visual image patches (the GSM). A hint that this might indeed be possible came from previous studies showing that both in the SSN[8,33] and the GSM[6,43], a range of parameters exists for which the response or posterior mean monotonically increases while the variance decreases with increasing stimulus strength. Empirically, we found a quantitative match that went beyond this coarse, qualitative trend: for example, the SSN was also able capture much of the detailed structure of the GSM posteriors.

Interestingly, the divisive normalization performed by the SSN has been proposed to be a canonical operation implemented throughout the cortex[42]. At the same time, stacked layers of subunits, each with a GSM-like separation of content- and style-like variables, have been suggested to form the basis of probabilistic generative models underlying deep learning[44]. Therefore, our results establishing the SSN as an appropriate recognition model for the GSM suggest that, similarly, a cascade of circuits with SSN-like dynamics could perform efficient inference under more powerful generative models and thus account for computations beyond V1.

Function-optimized neural networks

Our approach extends classical function-first approaches for training neural network models that had shown how various steady-state properties of cortical responses (such as receptive fields,

or trial-averaged activities) emerge from optimizing neural networks for some computationally well-defined objective (such as object recognition, memory, or context-dependent decision making and sensori-motor control)[45–50]. Notably, our sampling-based computational objective required our network to modulate not only the mean but also the variability of its responses in a stimulus-dependent manner. This made the training of networks significantly more challenging than conventional approaches training networks for deterministic targets without explicitly requiring them to modulate their variability[45,49,51,52]. In return, the dynamics of our network exhibited rich, stimulus-modulated patterns of variability, reproducing a variety of ubiquitous features of the trial-by-trial behaviour of cortical responses (noise variability, transients, and oscillations) beyond the steady-state or trial-average properties that could be addressed by previous work. Moreover, previous approaches typically violated at least one of the most salient constraints on the organization of cortical circuits by using networks that were either purely feed-forward[45], utilized neuronal transfer functions that lacked the expansive nonlinearities characteristic of cortical neurons[45,48,50–52], had no separation of E and I cells[45,50–52], or had noiseless dynamics[51]. In contrast, our network respected all of these constraints and thus made predictions which could not have been obtained by using simpler models (such as total inhibitory inputs lagging behind overall excitatory inputs).

Neural representations of uncertainty

Our approach markedly differed from previous work on the neural bases of probabilistic inference. Previous models were typically derived using a *top-down* approach (but see Refs. 49,53), using hand-designed network dynamics that explicitly mimicked specific existing approximate inference algorithms from machine learning based on sampling[39,40,54–56] or other representations[17,25,39,57,58]. As a result, these models also often did not respect some basic biological constraints (e.g. Dale's principle[39,55,57,58]), or had to assume an unrealistically rapid and direct influence of stimuli on network parameters (e.g. synaptic weights[54,55]). In contrast, we used a more *bottom-up* approach, starting from known constraints of cortical circuit organization, and then optimizing their parameters to achieve efficient sampling-based probabilistic inference without pre-specifying the details of the dynamics that needed to be implemented. In turn, our optimized network "discovered" novel algorithmic motifs for speeding up probabilistic inference whose benefits have not been described before (such as the specific use of oscillations and transients beyond the general use of so-called "momentum" variables[59]), or that have only recently appeared in machine learning algorithms (e.g. out-of-equilibrium dynamics[28]). Although some of these motifs have been observed in previous work[54], their function remained unclear as they were built-in by design rather than obtained as a result of optimization, or appeared purely epiphenomenal. In contrast, these motifs served computationally well-defined functions in our network.

Cortical variability, transients, and oscillations

Our work suggests a novel unifying function for three ubiquitous properties of sensory cortical responses: stimulus-modulated variability, transient overshoots, and gamma oscillations. In previous work, these phenomena have traditionally been studied in isolation and ascribed separate functional roles that have been difficult to reconcile – partly because they have not typically been derived normatively, i.e. directly from some functional objective in a principled manner (but see e.g. Ref. 57). As a result, and they have not been normatively derived from some well-defined objective. For example, cortical variability has most often been considered a nuisance, diminishing the accuracy of neural codes[16]. Theories postulating a functional role of variability in probabilistic computations have only considered the steady-state distribution of responses without making specific predictions about their dynamical features[6,17,60]. Conversely, transient responses prominently feature as central ingredients of models of predictive coding, where they signal novelty or deviations between predicted and observed states[57]. However, these theories did not address response variability. Moreover, a prediction of this framework, at least in its classical form, is that neural responses to persistent stimuli should decay to baseline once a percept is formed. This is however often not the case, with neural responses in V1 decaying to non-vanishing stimulus-dependent steady levels after the transient[3], even when stimuli had higher-order structure that efficiently drove higher-level visual areas (V2)[61]. In contrast, our work accounts for both transients and steady-state responses starting from the same principle, using only the equivalents of "internal representation neurons"[62] of predictive coding but without invoking specific prediction error-coding neurons. In particular, our model correctly predicted a specific scaling relationship between transients and steady-state responses which we tested by novel analyses of experimental data (Fig. 7). Furthermore, our mathematical analysis suggested that prediction-error-like signals (more formally, re-

sponses that scale with the magnitude of change in the target distribution, Supplementary Material) are a generic signature of sampling-based dynamics, and will thus not only appear at stimulus onsets but in any situation when predictions change temporally.

Gamma oscillations have also been proposed as a substrate for a number of functional roles in the past, related to how information is encoded, combined, or routed in the brain[11–14,63]. These putative functions need not be mutually exclusive to that played in our network. Nevertheless, some of these functions seem difficult to reconcile with specific experimental findings[3,19,20,64,65]. More generally, theories of gamma oscillations do not typically address transients. Although there are extensions of the predictive coding framework that do account for the presence of gamma oscillations, by attributing to it the representation of prediction errors[66], these theories would also predict a tight coupling between gamma-band synchronization and firing rates (both related to prediction errors) which has not been confirmed experimentally[67]. Moreover, it is unclear whether these theories would also account for properties beyond the mere existence of gamma oscillations, such as the frequency modulation by contrast[3,4] that our model reproduced (Fig. 5), or indeed any aspect of the ubiquitous variability of cortical responses, and its modulation by stimuli, which our model also reproduced as a core feature (Figs. 2, 3 and 5). In contrast, our results show that variability, transients, and gamma oscillations can all emerge from the same functional objective: that neural circuits use an efficient sampling-based representation of uncertainty under time constraints.

The mechanism by which gamma oscillations are generated in the brain, particularly whether it involves interactions between E and I cells ('PING' mechanism) or among I cells only ('ING' mechanism), is a subject of current debate[21]. In our model, voltage-clamping of E cells eliminated gamma oscillations (Fig. S8), pointing to the 'PING' mechanism. However, our network only included a single inhibitory cell type, and heavily constrained connectivity, therefore it remains for future work to study how the precise mechanism of gamma generation depends on such architectural constraints.

Hierarchical processing

Although our model showed salient divisive normalization (Fig. S4) that is thought to underlie many non-classical receptive field effects in V1[42], it still only represented a single idealized V1 hypercolumn and thus could not address layer- and cell-type specific lateral and top-down processing which, for example, predictive coding models can more naturally capture[68]. In particular, recent experimental data shows (movement related) mismatch signals in layer 2/3 of mouse V1[69], suggesting that these neurons may specifically represent errors between bottom-up visual input and top-down predictions. It will be interesting to see whether such neurons would also "automatically" emerge via our approach when optimizing a more complex architecture than what we used here, or if they would require special design constraints or decisions. More generally, a combination of sampling-based representations and predictive coding may be possible[70] and could lead to computationally powerful representations that both encode uncertainty (which most predictive coding models ignore) and are suited for hierarchical processing (which many models of probabilistic representations eschew). Such a hybrid architecture might be able to account for specific forms of variability modulation by stimuli or top-down signals (at which sampling-based models excel[6,26,40]) as well as for prediction error-like signals (naturally captured by predictive coding models[57,68]).

Studying more hierarchical or spatially extended versions of our model may also allow us to study longer-range aspects of gamma oscillations, such as gamma synchronization[63], and the dependence of gamma power on the structure of the stimulus at larger spatial scales[71], which our model of a local hypercolumn could not address. If local features encoded by neurons in different hypercolumns form parts of the same higher-order feature, one expects these neurons to show correlated variability under a sampling-based representation[27], which in turn may lead to the synchronization of gamma oscillations at their respective sites.

Finally, a fully hierarchical version of our model, including layers that directly control decisions or actions, could also allow end-to-end training for behaviorally relevant tasks, rather than the interim goal of representing uncertainty that we used here. This in turn would make it possible to evaluate whether variability is still used to represent uncertainty despite lacking an explicit objective for doing so. Preliminary results on unsupervised training suggest that this may be the case[53]. By providing a direct read-out of predicted behavior, such hierarchical networks will be ideal to study the link between the temporal dynamics of neural and behavioral forms of variability.

Acknowledgements

This work was supported by the Wellcome Trust (New Investigator Award 095621/Z/11/Z and Investigator Award in Science 212262/Z/18/Z to M.L., and Seed Award 202111/Z/16/Z to G.H.), and the Human Frontiers Science Programme (Research Grant RGP0044/2018 to M.L.). We are grateful to A. Ecker, P. Berens, M. Bethge, and A. Tolias for making their data publicly available, to G. Orbán, A. Bernacchia and Y. Ahmadian for useful discussions, and to J.P. Stroud for detailed comments on the manuscript.

Author Contributions

R.E. G.H. and M.L. designed the study. R.E. and G.H. developed the optimization approach, R.E. ran all numerical simulations, R.E. and G.H. analyzed experimental data, all authors performed analytical derivations, R.E. G.H. and M.L. interpreted results and wrote the paper, with comments from L.A.

Competing Interests statement

The authors declare no competing interests.

Online methods

The generative model

Following Refs. 6,31, we adopted the Gaussian scale mixture model (GSM)[30] as the generative model of natural image patches under which the primary visual cortex (V1) performs inference. Thus, an image patch $\mathbf{x} \in \mathbb{R}^{N_x}$ was assumed to be constructed by linearly combining a set of local features, the columns of $\mathbf{A} \in \mathbb{R}^{N_x \times N_y}$, weighted by a set of image-specific feature coefficients, $\mathbf{y} \in \mathbb{R}^{N_y}$, and scaled by a single global (at the scale of the image patch) contrast variable, $z \in \mathbb{R}$, plus additive white Gaussian noise:

$$\mathbf{x}|\mathbf{y}, z \sim \mathcal{N}\left(z\,\mathbf{A}\,\mathbf{y}, \sigma_x^2\,\mathbf{I}\right) \tag{1}$$

where the feature coefficients were assumed to be drawn from a multivariate Gaussian distribution:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}) \tag{2}$$

and z was drawn from a Gamma prior $\mathcal{P}(z)$ (Table S1, but see Ref. 43).

To model inferences in a V1 hypercolumn, we chose the columns of $\mathbf{A}$ (the so-called projective fields of the latent variables) to be oriented Gabor filters that only differed by their orientation (evenly spaced between $-90°$ and $90°$, 4 examples are shown in Fig. 1a, see also Fig. S1). The prior covariance matrix $\mathbf{C}$ was a circulant matrix whose elements varied smoothly as a function of the angular distance between the orientations of the projective fields of the corresponding latent variables, with positive and negative correlations between latent variables with similarly and orthogonally oriented projective fields in $\mathbf{A}$, respectively (Fig. 2d, top left; Fig. S1).

The ideal observer's posterior over latent spatial features $\mathbf{y}$ under the GSM for a given image, $\mathbf{x}$, and a known contrast z, can be written as[43]:

$$\mathbf{y}|\mathbf{x}, z \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3}$$

$$\text{with } \boldsymbol{\mu} = \frac{z}{\sigma_x^2}\,\boldsymbol{\Sigma}\,\mathbf{A}^\mathsf{T}\,\mathbf{x} \tag{4}$$

$$\text{and } \boldsymbol{\Sigma} = \left(\mathbf{C}^{-1} + \frac{z^2}{\sigma_x^2}\,\mathbf{A}^\mathsf{T}\,\mathbf{A}\right)^{-1} \tag{5}$$

Although, in general, z would also need to be inferred, as z is just a single scalar, of which the inference pools information across all pixels in the input, we approximated the posterior over z with a delta distribution at z^*, the true value of z that was used to generate the input[43]. Thus, the final posterior over $\mathbf{y}$, after marginalizing out the unknown z, was approximated by substituting z^* into Eq. 3:

$$\mathcal{P}(\mathbf{y}|\mathbf{x}) \simeq \mathcal{P}(\mathbf{y}|\mathbf{x}, z^*) \tag{6}$$

Following Ref. 6, membrane potentials, $\mathbf{u}$, were taken to represent a weakly non-linear function of visual feature activations $\mathbf{y}$ (Supplementary Material and Fig. S2):

$$u_i = \alpha_{\text{nl}}\,\lfloor y_i + b_{\text{nl}} \rfloor^{n_{\text{nl}}} \tag{7}$$

where $\lfloor \cdot \rfloor$ is the threshold-linear function, and α_{nl}, b_{nl}, and n_{nl} are respectively the scaling, baseline, and power of the transformation (Table S1, Fig. S2a).

Network dynamics and architecture

Following Ref. 8, we modelled the dynamics of a non-linear stochastic excitatory–inhibitory recurrent neural network as:

$$\tau_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = -u_i(t) + h_i(t) + \textstyle\sum_j W_{ij}\,r_j(t) + \eta_i(t) \tag{8}$$

where u_i represents the membrane potential of neuron i, h_i is its feedforward input, η_i is process noise (incorporating intrinsic and extrinsic forms of neural variability), and W_{ij} is the weight of the synapse connecting neuron j to neuron i. The recurrent weight matrix $\mathbf{W}$ was constrained to respect Dale's principle, that is, all outgoing synapses of each cell had the same sign, and we had the same number of excitatory and inhibitory cells (N_y). Process noise was spatially and temporally correlated zero-mean Gaussian:

$$\langle \boldsymbol{\eta}(t) \rangle = 0, \quad \left\langle \boldsymbol{\eta}(t)\,\boldsymbol{\eta}(t+s)^\mathsf{T} \right\rangle = \boldsymbol{\Sigma}^\eta \exp\left(-s/\tau_\eta\right) \tag{9}$$

where $\mathbf{\Sigma}^{\eta}$ is the stationary covariance matrix and τ_{η} is the timescale of the process noise (Table S1). Firing rates r_i were given by a supralinear transformation of membrane potentials:

$$r_i(t) = k\lfloor u_i(t)\rfloor^n \tag{10}$$

where k and n were respectively the scale and exponent of the firing rate nonlinearity (Table S1). As in standard models of V1 simple cells[72], the stimulus-dependent input to each neuron was obtained by applying a linear filter $\mathbf{W}_{\text{ff}}$ to the stimulus followed by a static nonlinearity:

$$h_i(t) = \alpha_{\text{h}}\left\lfloor \beta_{\text{h}} + \textstyle\sum_j \mathbf{W}^{\text{ff}}_{ij}\, x_j(t)\right\rfloor^{\gamma_{\text{h}}} \tag{11}$$

where $\mathbf{x}(t)$ was the stimulus (input image) received at time t, and α_{h}, β_{h}, and γ_{h} were respectively the scale, baseline, and exponent of the input nonlinearity (Table S1, Fig. S3b). The feedforward receptive fields of the neurons were identical (up to a constant factor) to the projective fields of the corresponding latent variables of the GSM, such that each excitatory–inhibitory cell pair received the same input: $\mathbf{W}^{\text{ff}} = [\mathbf{A}\ \ \mathbf{A}]^{\text{T}}/15$, where $\mathbf{A}$ was the same matrix as in the generative model (Eq. 1), and $[\mathbf{A}\ \ \mathbf{A}]$ denotes concatenating $\mathbf{A}$ with itself column-wise.

The network parameters that we optimized were $\mathbf{W}$, $\mathbf{\Sigma}^{\eta}$, and α_{h}, β_{h}, and γ_{h} (Fig. S3). To reduce computational cost, we used a lower-dimensional parameterization of $\mathbf{W}$ and $\mathbf{\Sigma}^{\eta}$, such that each quadrant (E–E, E–I, I–E, and I–I) was a smooth circulant matrix with circular Gaussian falloff around each cell with some characteristic length scale and amplitude (Supplementary Material). This resulted in a total of 15 parameters to be optimized: 8 describing the weight matrix $\mathbf{W}$, 4 describing $\mathbf{\Sigma}^{\eta}$, and 3 specifying the mapping from stimuli to network inputs. See Supplementary Material for further details.

Training and test stimuli

The training set (Fig. 2b) consisted of five images, $\mathbf{x}^{\alpha} = z_{\alpha}\mathbf{A}\bar{\mathbf{y}}$ ($\alpha = 1, \ldots, 5$) with the same dominant orientation (defined by $\bar{\mathbf{y}}$, Supplementary Material) but different contrast levels (z_{α}), together with their corresponding five posterior distributions, $\mathcal{P}^{\alpha} = \mathcal{P}(\mathbf{u}|\mathbf{x}^{\alpha})$, as prescribed by the generative model (Eqs. 3-7).

To test generalization in the network, we generated a set of 500 novel images with the GSM, which were thus not constrained to have a single dominant orientation (as the prior allowed multiple elements of $\mathbf{y}$ with different projective fields to be non-zero, Eq. 2). To be consistent with the training set, we did not include additive noise in $\mathbf{x}$, and added a contrast-dependent baseline to $\mathbf{y}$ so that its mean was modulated by contrast in the same way as in the training set. For each image in the test set, we also computed the corresponding posterior distributions to evaluate the network's test performance.

Network optimization

The cost function $\mathcal{F}$ for which we optimized the network consisted of four terms for each input image α in the training set:

$$\mathcal{F} = \sum_{\alpha}\Big(\epsilon_{\text{mean}}\,\phi^{\alpha}_{\text{mean}} + \epsilon_{\text{var}}\,\phi^{\alpha}_{\text{var}} + \epsilon_{\text{cov}}\,\phi^{\alpha}_{\text{cov}} + \epsilon_{\text{slow}}\,\phi^{\alpha}_{\text{slow}}\Big) \tag{12}$$

where the first three terms penalized deviations of the moments (mean, variance, and covariance) of the network's (across trial) response distribution from the target moments (corresponding to the ideal observer's posterior distributions, $\mathcal{P}^{\alpha}$), averaged over a finite time window starting within 500 ms of stimulus onset, and the last term was an additional slowness cost, penalizing the autocorrelation of network responses (Supplementary Material). The $\epsilon_{\ldots}$ constants (see Table S1) represent the relative weights of these terms. Setting $\epsilon_{\text{slow}} = 0$ did not qualitatively affect our results (not shown). In the first control network (Fig. 5, right column), we set $\epsilon_{\text{var}} = \epsilon_{\text{cov}} = \epsilon_{\text{slow}} = 0$, but kept all other meta-parameters and target means the same. In the second control network (Fig. S7, right), all $\epsilon_{\ldots}$ parameters were the same as for the optimization of the original network, but the target covariances were modified to induce contrast-independent Fano factors (see Supplementary Material).

To minimize the cost function in Eq. 12, we used a novel combination of stochastic and deterministic methods, both involving back-propagation through time[73] (Supplementary Material). The optimizer was written in OCaml and can be found online at LINK.

As the cost function that we used Eq. 12 was strongly non-convex, we checked the robustness of our findings by performing 10 further optimization attempts from random initial conditions. No solutions achieved substantially lower costs, and those whose final cost was comparable to the network presented in the main text behaved qualitatively similarly (in particular, they showed contrast-dependent oscillations and transients).

Numerical experiments after training

To obtain a reliable estimate of the statistical moments of neural responses to a stationary input

(Figs. 2 and 3), a total of $20,000$ independent samples (taken 200 ms apart) were drawn from the network, not including transients, as neural activity evolved according to Eq. 8. Neural activities in Fig. 2a show 1 s of simulated network activity, further convolved with a 20 ms sliding window to match the effects of spike binning to compute average rates in experiments. Neural trajectories in Fig. 2b correspond to the neural activity of two cells in the network with preferred orientations $42°$ (u_i) and $16°$ (u_j), over a post-transient period of 500 ms. To illustrate both the degree of modulation of the posterior covariances and the match between posterior and network covariances in Fig. 3c, the top three PCs of each posterior covariance were computed. Neural activity was then projected onto each PC, and the amount of variance along each direction was computed. The middle plots of Fig. 3c present these posterior PCs scaled by either the square root of the total variance along that direction in the GSM (in green) or in the network (in red).

Autocorrelograms in Fig. 4a were computed in 500 non-overlapping windows of 2 s of simulated neural activity each (subsampled at 0.4 ms) after stimulus onset (not including transients), and then averaged across these windows. Autocorrelograms were first computed for individual cells' membrane potentials and then averaged across all cells. Crosscorrelograms and E–I lags in Fig. 4b–c were computed from a single 400 s-long simulation after stimulus onset and transients (without subsampling). The E–I lag for each cell was determined as the location of the maximum in the anti-symmetric component of the cross-correlogram between its total E and I input. Langevin samplers in Fig. 4a–b corresponded to neural networks with linear, time-reversible dynamics, not constrained to respect Dale's principle (Supplementary Material). As variability in a linear network does not depend on the input (unlike in our nonlinear circuit model), we implemented a separate Langevin sampler for each input. Autocorrelograms and cross-correlograms for the Langevin sampler were computed as for the original network.

Average firing rates in Fig. 5a were computed from the same neural traces used in Fig. 2 to compute $\mathbf{u}$ moments (here taking the average of $\mathbf{r}$ instead of $\mathbf{u}$ in Eq. 8). To compute Fano factors in Fig. 5b we considered a doubly stochastic Gamma process and computed spike-counts over $500,000$ 100 ms windows. The shape parameter k of the Gamma process (Table S1) was chosen to reproduce the experimentally found range of Fano factors which could be less than 1 (Fig. 5b), resulting in more regular spike trains than an inhomogenous Poisson process. Power spectra in Fig. 5c were computed from the (across-cell) average neural activity (membrane potentials), following standard approaches[8], using the same samples as the autocorrelograms of Fig. 4a (see above). Gamma peak frequency was identified as the location of the local maximum (within the gamma band, 20–80 Hz) of the power spectrum. Transients in Fig. 5d correspond to average firing rates across E cells and trials ($n = 100$). These were also averaged over 10 ms windows to mimic the resolution of the experimental results. To account for the response delays observed in experimental data, we simulated a random delay time (truncated Gaussian, with 45 ms mean and 5 ms s.d.) to each E–I cell pair in the network. Conductance changes (relative to baseline) in Fig. 5e correspond to across trial averages ($n = 20$) for a single neuron with preferred orientation aligned to that of the stimulus (see Supplementary Material for further details).

To study the effect of oscillations (Fig. 6a and b) and transient overshoots (Fig. 7a and b) on sampling accuracy, we employed a family of simplified systems which were constructed as 1-dimensional Gaussian processes[74] designed to match the statistics (stationary mean and variance) of single neurons in the network, but allowing to parametrically and independently vary either the degree of oscillatoriness in the system (i.e. the kernel of the Gaussian process), or the temporal profile of the mean response (Supplementary Material).

Autocorrelograms and power spectra of Fig. 6c were computed as in Fig. 4a and Fig. 5c, but for the directions in the space of neural activity that corresponded to the first ten PCs of neural variability. To quantify the degree of oscillatoriness along each PC, we fitted the corresponding autocorrelogram with the function:

$$C(\Delta t) = [1 - \alpha + \alpha \cos(2\pi f \Delta t)] \cdot \frac{\tau_1 e^{-\Delta t/\tau_1} - \tau_2 e^{-\Delta t/\tau_2}}{\tau_1 - \tau_2} \tag{13}$$

containing three free parameters: the degree of oscillatoriness α, and two timescales τ_1, and τ_2. The second term, including the exponentials, describes the effect of the filtering of an input whose temporal correlation is an exponential with time constant τ_1, by the network's intrinsic dynamics characterized by a second time constant τ_2.

Overshoots in Fig. 7c and d were obtained using the same image that was used to train the network at 0.7 contrast, and computed as the maximal across-trial average ($n = 100$) response of

each E cell (membrane potential for c, firing rate for d), minus its stationary mean response, further averaged over 1000 delay configurations in our network (as for Fig. 5d, see above). Steady-state differences denote the magnitude of mean evoked responses of each cell with respect to its mean pre-stimulus response. Fig. 7e shows novel analyses of experimental recordings from awake macaque V1 during the presentation of moving gratings of different orientations[29] (data released in the repository of Ref. 75). Following the same procedure as in Ref. 8, only cells that were significantly tuned (orientation tuning index greater than 0.75) and had an average evoked rate above 1 spike per second were included in the analysis. For each cell and each stimulus, a time-dependent firing rate trace was first obtained by averaging spikes across trials in a 50 ms sliding square window. From these traces, overshoots and steady-state differences were then computed (dots in Fig. 7e top) as average evoked responses excluding transients ($t > 160$ ms after stimulus onset) minus average baseline responses (computed from the 300 ms prior to stimulus presentation). The overshoot size is computed as the maximum of the response trace for $t < 160$ ms after stimulus onset, minus the same average evoked response previously computed. The linear regression was performed using SciPy's 'linregress' function, which reports a two-sided p-value (null hypothesis: zero slope), using a Wald test with a t-distribution of the test statistic. Results in the bottom plots of Fig. 7d and e were computed by averaging stimuli presented at each neuron's preferred orientation ($\pm 30^\circ$) or orthogonal to its preferred orientation ($\pm 30^\circ$). We tested for significance in overshoot tuning using SciPy's 'ttest_ind' function (null hypothesis: identical average).

References

1. MM Churchland, MY Byron, JP Cunningham, LP Sugrue, MR Cohen, GS Corrado, WT Newsome, AM Clark, P Hosseini, BB Scott, et al. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature Neuroscience*, 13(3):369, 2010.
2. B Haider, M Häusser, and M Carandini. Inhibition dominates sensory responses in the awake cortex. *Nature*, 493(7430):97–100, 2013.
3. S Ray and John HR Maunsell. Differences in gamma frequencies across visual cortex restrict their possible use in computation. *Neuron*, 67(5):885–896, 2010.
4. MJ Roberts, E Lowet, NM Brunet, M Ter Wal, P Tiesinga, P Fries, and P De Weerd. Robust gamma coherence between macaque V1 and V2 by dynamic frequency matching. *Neuron*, 78(3):523–536, 2013.
5. IM Finn, NJ Priebe, and D Ferster. The emergence of contrast-invariant orientation tuning in simple cells of cat visual cortex. *Neuron*, 54(1):137–152, 2007.
6. G Orbán, P Berkes, J Fiser, and M Lengyel. Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, 92(2):530–543, 2016.
7. A Ponce-Alvarez, A Thiele, TD Albright, GR Stoner, and G Deco. Stimulus-dependent variability and noise correlations in cortical MT neurons. *Proceedings of the National Academy of Sciences*, 110(32):13162–13167, 2013.
8. G Hennequin, Y Ahmadian, DB Rubin, M Lengyel, and KD Miller. The dynamical regime of sensory cortex: stable dynamics around a single stimulus-tuned attractor account for patterns of noise variability. *Neuron*, 98(4):846–860, 2018.
9. B Haider and DA McCormick. Rapid neocortical dynamics: cellular and network mechanisms. *Neuron*, 62(2):171–189, 2009.
10. G Buzsáki and XJ Wang. Mechanisms of gamma oscillations. *Annual review of neuroscience*, 35:203–225, 2012.
11. C von Der Malsburg. The correlation theory of brain function. In *Models of neural networks*, pages 95–119. Springer, 1994.
12. CM Gray, P König, AK Engel, and W Singer. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature*, 338(6213):334–337, 1989.
13. T Akam and DM Kullmann. Oscillations and filtering networks support flexible routing of information. *Neuron*, 67(2):308–320, 2010.
14. T Masquelier, E Hugues, G Deco, and SJ Thorpe. Oscillations, phase-of-firing coding, and spike timing-dependent plasticity: an efficient learning scheme. *Journal of Neuroscience*, 29(43):13484–13493, 2009.
15. W Schultz, P Dayan, and PR Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.
16. MN Shadlen and WT Newsome. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *Journal of Neuroscience*, 18(10):3870–3896, 1998.
17. WJ Ma, JM Beck, PE Latham, and A Pouget. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438, 2006.

18. P Berkes, G Orbán, M Lengyel, and J Fiser. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331(6013):83–87, 2011.
19. MN Shadlen and JA Movshon. Synchrony unbound: a critical evaluation of the temporal binding hypothesis. *Neuron*, 24(1):67–77, 1999.
20. A Thiele and G Stoner. Neuronal synchrony does not correlate with motion coherence in cortical area MT. *Nature*, 421(6921):366–370, 2003.
21. P Tiesinga and TJ Sejnowski. Cortical enlightenment: are attentional gamma oscillations driven by ing or ping? *Neuron*, 63(6):727–732, 2009.
22. DC Knill and W Richards. *Perception as Bayesian inference*. Cambridge University Press, 1996.
23. P Mamassian, M Landy, and LT Maloney. Bayesian modelling of visual perception. *Probabilistic models of the brain*, pages 13–36, 2002.
24. J Fiser, P Berkes, G Orbán, and M Lengyel. Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, 14(3):119–130, 2010.
25. S Deneve, PE Latham, and A Pouget. Efficient computation and cue integration with noisy population codes. *Nature Neuroscience*, 4(8):826, 2001.
26. RM Haefner, P Berkes, and J Fiser. Perceptual decision-making as probabilistic inference by neural sampling. *Neuron*, 90(3):649–660, 2016.
27. M Bányai, A Lazar, L Klein, J Klon-Lipok, M Stippinger, W Singer, and G Orbán. Stimulus complexity shapes response correlations in primary visual cortex. *Proceedings of the National Academy of Sciences*, 116(7):2723–2732, 2019.
28. KS Turitsyn, M Chertkov, and M Vucelja. Irreversible Monte Carlo algorithms for efficient sampling. *Physica D: Nonlinear Phenomena*, 240(4-5):410–414, 2011.
29. AS Ecker, P Berens, GA Keliris, M Bethge, NK Logothetis, and AS Tolias. Decorrelated neuronal firing in cortical microcircuits. *Science*, 327(5965):584–587, 2010.
30. MJ Wainwright and EP Simoncelli. Scale mixtures of Gaussians and the statistics of natural images. In *Advances in Neural Information Processing Systems*, pages 855–861, 2000.
31. R Coen-Cagli, A Kohn, and O Schwartz. Flexible gating of contextual influences in natural vision. *Nature Neuroscience*, 2015.
32. O Schwartz, TJ Sejnowski, and P Dayan. Perceptual organization in the tilt illusion. *Journal of Vision*, 9(4):19–19, 2009.
33. Y Ahmadian, DB Rubin, and KD Miller. Analysis of the stabilized supralinear network. *Neural computation*, 25(8):1994–2037, 2013.
34. N Kraynyukova and T Tchumatchenko. Stabilized supralinear network can give rise to bistable, oscillatory, and persistent activity. *Proceedings of the National Academy of Sciences*, 115(13):3464–3469, 2018.
35. NJ Priebe and D Ferster. Inhibition, spike threshold, and stimulus selectivity in primary visual cortex. *Neuron*, 57(4):482–497, 2008.
36. C van Vreeswijk and H Sompolinsky. Chaotic balanced state in a model of cortical circuits. *Neural computation*, 10(6):1321–1371, 1998.
37. DJC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.
38. JD Murray, A Bernacchia, DJ Freedman, R Romo, JD Wallis, X Cai, C Padoa-Schioppa, T Pasternak, H Seo, D Lee, et al. A hierarchy of intrinsic timescales across primate cortex. *Nature Neuroscience*, 17(12):1661, 2014.
39. A Grabska-Barwinska, J Beck, A Pouget, and P Latham. Demixing odors-fast inference in olfaction. In *Advances in Neural Information Processing Systems*, pages 1968–1976, 2013.
40. G Hennequin, L Aitchison, and M Lengyel. Fast sampling-based inference in balanced neuronal networks. In *Advances in neural information processing systems*, pages 2240–2248, 2014.
41. M Okun and I Lampl. Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. *Nature Neuroscience*, 11(5):535, 2008.
42. M Carandini and DJ Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51, 2012.
43. R Echeveste, G Hennequin, and M Lengyel. Asymptotic scaling properties of the posterior mean and variance in the Gaussian scale mixture model. *arXiv preprint arXiv:1706.00925*, 2017.
44. AB Patel, T Nguyen, and RG Baraniuk. A probabilistic theory of deep learning. *arXiv preprint arXiv:1504.00641*, 2015.
45. DLK Yamins, H Hong, CF Cadieu, EA Solomon, D Seibert, and JJ DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

46. D Festa, G Hennequin, and M Lengyel. Analog memories in a balanced rate-based network of EI neurons. In *Advances in Neural Information Processing Systems*, pages 2231–2239, 2014.
47. G Hennequin, TP Vogels, and W Gerstner. Optimal control of transient dynamics in balanced networks supports generation of complex movements. *Neuron*, 82(6):1394–1406, 2014.
48. HF Song, GR Yang, and XJ Wang. Training excitatory-inhibitory recurrent neural networks for cognitive tasks: A simple and flexible framework. *PLoS computational biology*, 12(2):e1004792, 2016.
49. AE Orhan and WJ Ma. Efficient probabilistic inference in generic neural networks trained with non-probabilistic feedback. *Nature communications*, 8(1):138, 2017.
50. ED Remington, D Narain, EA Hosseini, and M Jazayeri. Flexible sensorimotor computations through rapid reconfiguration of cortical dynamics. *Neuron*, 98(5):1005–1019, 2018.
51. D Sussillo and LF Abbott. Generating coherent patterns of activity from chaotic neural networks. *Neuron*, 63(4):544–557, 2009.
52. V Mante, D Sussillo, KV Shenoy, and WT Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78, 2013.
53. L Aitchison, G Hennequin, and M Lengyel. Sampling-based probabilistic inference emerges from learning in neural circuits with a cost on reliability. *arXiv preprint arXiv:1807.08952*, 2018.
54. L Aitchison and M Lengyel. The Hamiltonian brain: efficient probabilistic inference with excitatory-inhibitory neural circuit dynamics. *PLoS computational biology*, 12(12):e1005186, 2016.
55. C Savin and S Deneve. Spatio-temporal representations of uncertainty in spiking neural networks. In *Advances in Neural Information Processing Systems*, pages 2024–2032, 2014.
56. 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(11):e1002211, 2011.
57. RPN Rao and DH Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79, 1999.
58. JM Beck, WJ Ma, R Kiani, T Hanks, AK Churchland, J Roitman, MN Shadlen, PE Latham, and A Pouget. Probabilistic population codes for Bayesian decision making. *Neuron*, 60(6):1142–1152, 2008.
59. RM Neal. MCMC using Hamiltonian dynamics. In S Brooks, A Gelman, G Jones, and XL Meng, editors, *Handbook of Markov Chain Monte Carlo*, chapter 5. CRC Press, 2011.
60. PO Hoyer and A Hyvärinen. Interpreting neural response variability as Monte Carlo sampling of the posterior. In *Advances in neural information processing systems*, pages 293–300, 2003.
61. J Freeman, CM Ziemba, DJ Heeger, EP Simoncelli, and JA Movshon. A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, 16(7):974, 2013.
62. GB Keller and TD Mrsic-Flogel. Predictive processing: A canonical cortical computation. *Neuron*, 100(2):424–435, 2018.
63. M Vinck and CA Bosman. More gamma more predictions: gamma-synchronization as a key mechanism for efficient integration of classical receptive field inputs with surround predictions. *Frontiers in systems neuroscience*, 10:35, 2016.
64. PR Roelfsema, VAF Lamme, and H Spekreijse. Synchrony and covariation of firing rates in the primary visual cortex during contour grouping. *Nature Neuroscience*, 7(9):982, 2004.
65. G Buzsáki and EW Schomburg. What does gamma coherence tell us about inter-regional neural communication? *Nature Neuroscience*, 18(4):484, 2015.
66. AM Bastos, WM Usrey, RA Adams, GR Mangun, P Fries, and KJ Friston. Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711, 2012.
67. A Peter, C Uran, J Klon-Lipok, R Roese, S Van Stijn, W Barnes, JR Dowdall, W Singer, P Fries, and M Vinck. Surface color and predictability determine contextual modulation of V1 firing and gamma oscillations. *eLife*, 8:e42101, 2019.
68. MW Spratling. Predictive coding as a model of response properties in cortical area V1. *Journal of Neuroscience*, 30(9):3531–3543, 2010.
69. P Zmarz and GB Keller. Mismatch receptive fields in mouse visual cortex. *Neuron*, 92(4):766–772, 2016.
70. L Aitchison and M Lengyel. With or without you: predictive coding and bayesian inference in the brain. *Current opinion in neurobiology*, 46:219–227, 2017.
71. D Hermes, KJ Miller, BA Wandell, and J Winawer. Stimulus dependence of gamma oscillations in human visual cortex. *Cerebral*

cortex, 25(9):2951–2959, 2015.

72. P Dayan and LF Abbott. *Theoretical neuroscience*, volume 806. Cambridge, MA: MIT Press, 2001.
73. PJ Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.
74. CKI Williams and CE Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT Press Cambridge, MA, 2006.
75. AS Ecker, P Berens, RJ Cotton, M Subramaniyan, GH Denfield, CR Cadwell, SM Smirnakis, M Bethge, and AS Tolias. State dependence of noise correlations in macaque primary visual cortex. *Neuron*, 82(1):235–248, 2014.