Interrogating theoretical models of neural computation with deep learning

Sean R. Bittner, Agostina Palmigiano, Alex T. Piet, Chunyu A. Duan, Francesca Mastrogiuseppe, Srdjan Ostojic, Carlos D. Brody, Kenneth D. Miller, and John P. Cunningham.

1 Abstract

The cornerstone of theoretical neuroscience is the circuit model: a system of equations that captures a hypothesized neural mechanism of scientific importance. At its best, such a model will give rise to an experimentally observed phenomenon – whether behavioral or in terms of neural activity – and thus can offer insight into neural computation. The operation of these circuits, like all models, critically depends on the choices of model parameters. Historically, the gold standard has been to analytically derive the relationship between model parameters and computational properties. However, this enterprise quickly becomes infeasible as biologically realistic constraints are included into the model, often resulting in *ad hoc* approaches to understanding the relationship between model and computation. We bring recent machine learning techniques – the use of deep generative models for probabilistic inference – to bear on this problem, learning distributions of parameters that produce the specified properties of computation. Importantly, the techniques we introduce offer a logical and unbiased means to understand the implications of model parameter choices on computational properties of interest. We motivate this methodology with a worked example analyzing sensitivity in the stomatogastric ganglion. We then use it to generate insights into neuron-type input-responsivity in primary visual cortex, a new understanding of rapid task switching in superior colliculus models, and improved attribution of bias in low-rank recurrent neural networks. More generally, this work moves us away from the tradeoff of biological realism vs analytical tractability, and towards the use of modern machine learning for sophisticated interrogation of biologically relevant models.

2 Introduction

The fundamental practice of theoretical neuroscience is to use a mathematical *model* to understand neural computation, whether that computation enables perception, action, or some intermediate processing [1]. In this field, a neural computation is systematized with a set of equations – the model – and these equations are motivated by biophysics, neurophysiology, and other conceptual considerations. The function of this system is governed by the choice of model parameters, which

when configured in some special way, give rise to some measurable signature of a computation. The work of analyzing a model then becomes the inverse problem: given a computation of interest, how can we reason about these special parameter configurations – their likely values, their uniquenesses and degeneracies, their attractor states and phase transitions, and more?

Consider the idealized practice: a theorist considers a model carefully and analytically derives how model parameters govern the computation. Seminal examples of this gold standard include our field's understanding of memory capacity in associative neural networks [2], chaos and autocorrelation timescales in random neural networks [3], and the paradoxical effect in excitatory/inhibitory networks [4]. Unfortunately, as circuit models include more biological realism, theory via analytic derivation becomes intractable. This fact creates an unfavorable tradeoff for the theorist. On the one hand, one may tractably analyze systems of equations with unrealistic assumptions (for example symmetry or gaussianity), producing accurate inferences about parameters of a too-simple model. On the other hand, one may choose a more biologically relevant model at the cost of *ad hoc* approaches to analysis (simply examining simulated activity), producing questionable or partial inferences about parameters of an appropriately complex, scientifically relevant model.

Of course, this same tradeoff has been confronted in many scientific fields and engineering problems characterized by the need to do inference in complex models. In respose, the machine learning community has made remarkable progress in recent years, via the use of deep neural networks as a powerful inference engine: a flexible function family that can map observed phenomena (in this case the measurable signal of some computation) back to probability distributions quantifying the likely parameter configurations. One celebrated example of this approach from the machine learning community, from which we draw key inspiration for this work, is the variational autoencoder [5, 6], which uses a deep neural network to induce an (approximate) posterior distribution on hidden variables in a latent variable model, given data. Indeed, these tools have been used to great success in neuroscience as well, in particular for interrogating parameters (sometimes treated as hidden states) in models of both cortical population activity [7, 8, 9, 10] and animal behavior [11, 12, 13]. These works have used deep neural networks to expand the expressivity and accuracy of statistical models of neural data [14].

However, these inference tools have not significantly influenced the study of theoretical neuroscience models for at least three reasons. First, at a practical level, the nonlinearities and dynamics of many theoretical models are such that conventional inference tools (for example mean field variational inference) typically produce a narrow set of insights into these models [15]. Indeed, only in the last

few years has the deep learning toolkit expanded to a point of relevance to this class of problem. Second, the object of interest from a theoretical model is not typically data itself, but rather a qualitative phenomenon – inspection of model behavior, or better, a measurable signature of some computation – an *emergent property* of the model. Third, because theoreticians work carefully to construct a model that has biological relevance, such a model as a result often does not fit cleanly into the framing of a statistical model. Technically, because many such models stipulate a noisy system of differential equations that can only be sampled or realized through forward simulation, they lack the explicit likelihood and priors central to the probabilistic modeling toolkit.

To address these three challenges, we developed an inference methodology – 'emergent property inference' – which learns a distribution over parameter configurations in a theoretical model. Critically, this distribution is such that draws from the distribution (parameter configurations) correspond to systems of equations that give rise to a specified emergent property. First, we stipulate a deep neural network that induces a flexible family of probability distributions over model parameterizations. We will insist on being able to quantify the probability of various model parameter configurations, and thus we choose the deep neural network to be of the (bijective) normalizing flow class [16]. Second, we quantify the notion of emergent properties as a set of moment constraints on datasets generated by the model. Thus an emergent property is not a single data realization, but a phenomenon or a feature of the model, which is the central object of interest to the theorist (unlike say the statistical neuroscientist). The requirement to condition on an emergent property requires the adaptation of deep probabilistic inference methods, and we extend recent tools to do so [17]. Third, because we can not assume the theoretical model has explicit likelihood on data or the emergent property of interest, we use stochastic gradient techniques in the spirit of likelihood free variational inference [18]. Taken together, emergent property inference (EPI) provides a methodology for inferring and then reasoning about parameter configurations that give rise to particular emergent phenomena in theoretical models. Emergent property inference is described schematically in Fig 1A.

Equipped with this methodology, we investigated four models of historical and current importance in theoretical neuroscience. These models were chosen to demonstrate generality through ranges of biological realism (conductance-based biophysics to recurrent neural networks), neural system function (pattern generation to abstract cognitive function), and network scale (four to infinite neurons). First, to motivate the contribution of emergent property inference, we investigated network syncing in a classic model of the stomatogastric ganglion [19]. Second, we coducted a

exploratory analysis of a four neuron-type dynamical model of primary visual cortex resulting in testable predictions of nonliear population input-responsivity. Third, we demonstrated how the systematic application of EPI to levels of behavioral accuracy can generate experimentally testable hypotheses regarding connectivity in superior colliculus. Fourth, we leveraged the flexibility of EPI to uncover the sources of bias in a low-rank recurrent neural network executing Bayesian inference. The novel scientific insights offered by EPI contextualize and clarify the previous studies exploring these models [19, 20, 21, 22] and more generally offer a quantitative grounding for theoretical models going forward, pointing a way to how rigorous statistical inference can enhance theoretical neuroscience at large.

We note that, during our preparation and early presentation of this work [23, 24], another work has arisen with broadly similar goals: bringing statistical inference to mechanistic models of neural circuits [25]. We are excited by this broad problem being recognized by the community, and we emphasize that these works offer complementary neuroscientific contributions and use different technical methodologies. Scientifically, our work has focused primarily on systems-level theoretical models, while their focus is on lower-level cellular models. Secondly, there are several key technical differences in the approaches (see Section A.1.4) perhaps most notably is our focus on the emergent property – the measurable signal of the computation in question, vs their focus on observed datasets; both certainly are worthy pursuits. The existence of these complementary methodologies emphasizes the increased importance and timeliness of both works.

3 Results

3.1 Motivating emergent property inference of theoretical models

Consideration of the typical workflow of theoretical modeling clarifies the need for emergent property inference. First, the theorist designs or chooses an existing model that, it is hypothesized, captures the computation of interest. To ground this process in a well-known example, consider the stomatogastric ganglion (STG) of crustaceans, a small neural circuit which generates multiple rhythmic muscle activation patterns for digestion [**?**]. A standard model for the STG is shown schematically in Figure 1A, and note that the behavior of this model will be critically dependent on its parameterization – the choices of conductance parameters $z = [g_{\text{el}}, g_{\text{synA}}]$. Second, the theorist defines the emergent property, the measurable signal of scientific interest. To continue our running STG example, one such emergent property is the phenomenon of *network syncing* – in

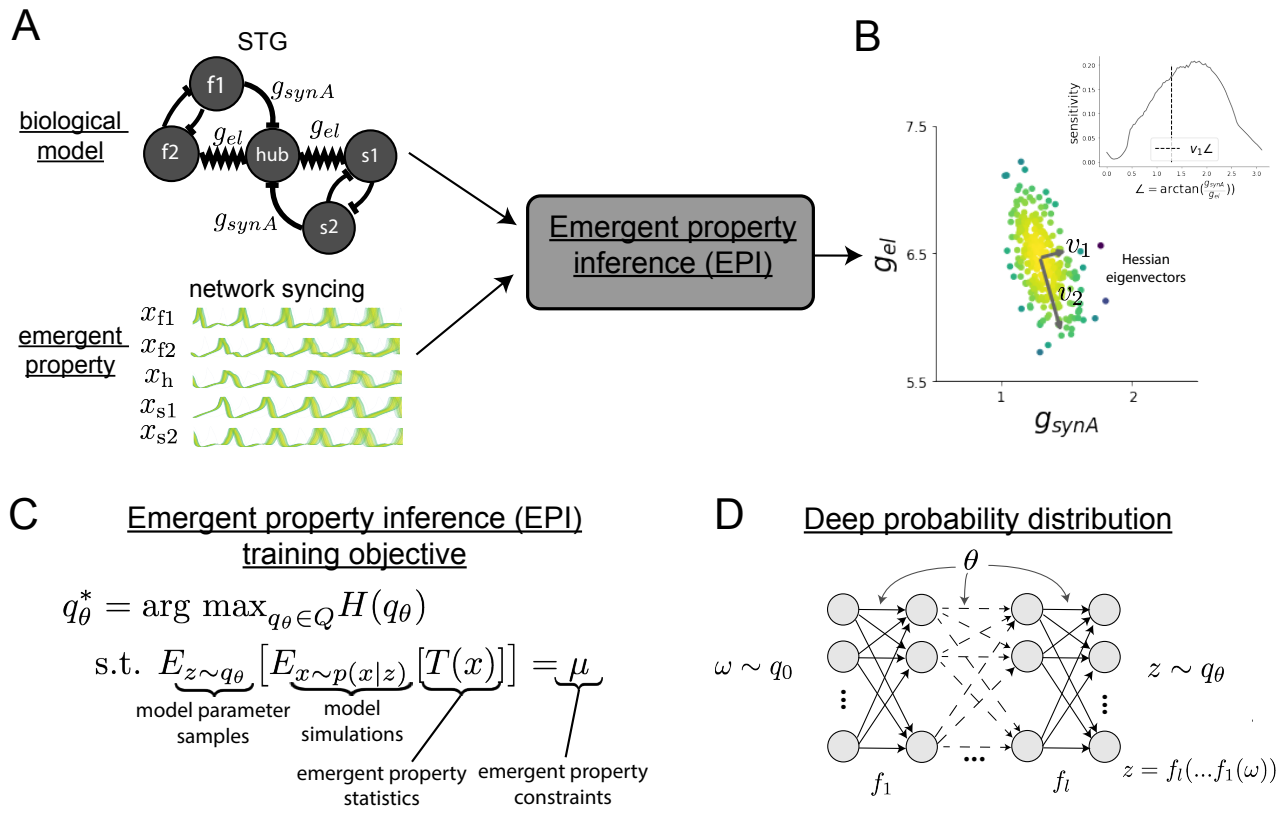

Figure 1: A. For a choice of model (STG) and emergent property (network syncing), emergent property inference (EPI) learns a posterior distribution of the model parameters $z = [g_{\text{el}}, g_{\text{synA}}]^\top$ conditioned on network syncing. B. An EPI distribution of STG model parameters producing network syncing. (Inset) Sensitivity of the system with respect to network syncing along all dimensions of parameter space away from the mode. C. EPI solves a a constrained stochastic optimization, in which emergent property statistics $T(x)$ are fixed in expectation over model simulations $x \sim p(x \mid z)$ and parameter distribution samples $z \sim q_\theta(z)$ to be a particular value μ. EPI distributions maximize randomness via entropy, although other measures are sensible. D. Degenerate solution networks (DSNs) are deep probability distributions $q_\theta(z)$ of theoretical model parameterizations that produce emergent properties of interest. The stochasticity of a deep probability distribution comes from a simple random variable $\omega \sim q_0$, where q_0 is often chosen to be an isotropic gaussian, and the structure comes from the deterministic transformation made by the deep neural network with optimized parameters θ.

certain parameter regimes, the activity of the hub neuron x_h matches that of the fast (x_{f1}, x_{f2}) and slow (x_{s1}, x_{s2}) populations. This emergent property is shown in Figure 1A. Third, qualtitative parameter analysis ensues: since precise mathematical analysis is intractable in this model, a brute force sweep of parameters is done to describe different parameter configurations that lead to the emergent property. In this last step lies the opportunity for modern machine learning. We can precisely quantify the emergent property as a statistical feature of the model, and we can infer a probability distribution over parameter configurations that produce this emergent property.

Before presenting technical details (in the following section), let us understand emergent property inference schematically: the black box in Figure 1A takes, as input, the model and the specified emergent property, and produces as output the parameter distribution shown in Figure 1B. This distribution – represented for clarity as samples from the distribution – is then a scientifically meaningful and mathematically tractable object. It conveys parameter regions critical to the emergent property, directions in parameter space that will be invariant (or not) to that property, and more. In the STG model, this distribution can be specifically queried to determine the prototypical parameter configuration for network syncing (the mode; Figure 1B star), and then how quickly network syncing will decay based on changes away that mode (Figure 1B, inset). Of course, validation of this distribution is critical: we can also plot a metric of network syncing that is independent of this inference process, to show that this EPI distribution indeed captures that structure (Figure 1B, contour lines). Taken together, bringing careful inference to theoretical models offers deeper insight into the behavior of these models, and the opportunity to make rigorous this last analysis step of the practice of theoretical neuroscience.

3.2 A deep generative modeling approach to emergent property inference

Emergent property inference (EPI) systematizes the three-step procedure of the previous section. First, we consider the model as a coupled set of differential (and potentially stochastic) equations. In the running STG example the dynamical state $x = \left[x_{\mathrm{f1}}, x_{\mathrm{f2}}, x_{\mathrm{hub}}, x_{\mathrm{s1}}, x_{\mathrm{s2}}\right]^\top$ is the membrane potential for each neuron, which evolves according to the biophysical conductance-based equation:

$$C_m \frac{\partial x}{\partial t} = -\left[h_{leak} + h_{Ca} + h_K + h_{hyp} + h_{elec} + h_{syn}\right] \tag{1}$$

where C_m=1nF, and h_{leak}, h_{Ca}, h_K, h_{hyp}, h_{elec}, h_{syn} are the leak, calcium, potassium, hyperpolarization, electrical, and synaptic currents, all of which have their own complicated dependence on x and θ (see Methods) [THIS NEEDS TO BE CLARIFIED]. Second, we define the emergent

property, which as above is network syncing: the phase locking of the population and its oscillation at an intermediate frequency (Figure 1A bottom) [REPETITIVE]. It is worth noting that theoretical work has characterized how model parameters such as the electrical conductance g_{el} and synaptic conductance g_{synA} govern the production of STG rhythms [19]. In this 5-neuron model, two fast neurons ($f1$ and $f2$) mutually inhibit one another, and oscillate at a faster frequency than the slow neurons ($s1$ and $s2$), which also mutually inhibit each other (Figure 1A). Quantifying this phenomenon is straightforward: we define [XXXXXXXXX statistics, values, and we insist on these being the measurable signatures of the EP... End up with a constraint function $E(T(X)) = \mu$ ....XXXXXXXXXXXXXX]. Third, having rationalized the above components, we can introduce deep generative modeling for performing emergent property inference. We seek a distribution over parameter configurations z, and we insist that samples from this distribution produce the emergent property; in other words, they obey the constraints introduced in Equation 1. Thus results the following optimization program:

$$\begin{aligned} &\underset{q_\theta \in \mathcal{Q}}{\operatorname{argmax}} \, H(q_\theta(z)) \\ &\text{s.t. } E_{z\sim q_\theta}\left[E_{x\sim p(x|z)}\left[T(x)\right]\right] = \mu \end{aligned} \tag{2}$$

The purpose of each element in this program is detailed in Figure 1D. Stating such a problem is easy enough; finding a tractable and suitably flexible family of probability distributions ($\mathcal{Q}$) is hard. We use norm flow deep learning [DETAIL HERE], inducing the deep distribution Fig 1E. Finally, we recognize that many distributions will satisfy the emergent property, so we require a normative principle to select amongst them. This principle is captured in Equation [XXX] by the primal objective H. Here we chose Shannon entropy as it has been well used for a variety of things [Elsayed, MEFN, Jaynes, Savin review], but the EPI methods (not the results) offered here are largely unaffected by this choice.

[A few sentences about ... we run optimization... learning.... and then what results is the EPI distribution voila.] Armed with this distribution, we now prove out the value of this technology by investigating a range of models and using EPI to produce novel scientific insights.

3.3 Comprehensive input-responsivity in a nonlinear sensory system

First, dynamical models with two populations (excitatory (E) and inhibitory (I) neurons) of visual processing have been used to reproduce a host of experimentally documented phenomena in primary visual cortex (V1). In particular regimes of excitation and inhbition, these models exhibit the

paradoxical effect [4], selective amplification [26], surround suppression [27], and sensory integrative properties [28]. Since inhibitory neurons mostly fall into one of three classes (parvalbumin (P)-, somatostatin (S)-, and vasointestinal peptide (V)-expressing neurons) [29, 30], theorists look to extend these dynamical models to four populations [20] (Fig. 3A).

The dynamical state of this model is the firing rate of each neuron type population $x = \begin{bmatrix} x_E & x_P & x_S & x_V \end{bmatrix}^\top$, which evolves according to rectified, exponentiated dynamics:

$$\tau \frac{dx}{dt} = -x + [Wx + h]_+^n \tag{3}$$

with effective connectivity weights W and input h. We set the time constant $\tau = 20ms$ and dynamics coefficient $n = 2$. We obtained an informative estimate of the effective connectivities between these neuron types in mice by multiplying their probability of connection by the synaptic efficacy [**?**] (see Section A.2.2). Given these fixed parameter choices of W, n, and τ and a baseline input b, we asked what differential inputs $dh = \begin{bmatrix} dh_E & dh_P & dh_S & dh_V \end{bmatrix}^\top$ cause each neuron type population to increase its firing rate.

Yet, we were at least able to derive the linearized response of the system $\frac{dx_{ss}}{dh}$ at fixed points x_{ss}. While this linear prediction is accurate for small differential inputs (Fig. 3B, left), it is often misleading in such nonlinear models as differential input strength increases (Fig. 3B, right). In fact, for a baseline input of $b = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}^\top$ the linearly predicted response for $dh = \begin{bmatrix} 0.5 & 0.5 & 0.5 & 0.5 \end{bmatrix}^\top$ was actually in the opposite direction of the true response for the V-population (Fig. 3B, right, green).

To get a more comprehensive picture of the input-reponsivity of each neuron type, we used EPI to learn a distribution of inputs dh that cause the rate of each neuron-type population α to increase by a value y with some allowed variance. We denote this emergent property of neuron-type responsivity as $\mathcal{B}(\alpha, y)$. In Fig. 3C, each column visualizes the approximate posterior dh corresponding to a specific neuron-type increase, while each row corresponds to amounts of increase 0.1 and 0.5. Akin to the practice of exploratory analysis of neural data sets, in which we search for meaningful structure, we can do an exploratory analysis of this *model* yielding a meaningful understanding of its operation.

To visualize these four-dimensional distributions, we pairploted the two-dimensional marginal densities, which yield an insightful picture. As expected, the inferred distributions revealed that each neuron-type's rate is sensitive to its direct input. The E- and P-population are largely unaffected by dh_V, and the S-population by dh_P for the rate increases examined. Additionally, we observed

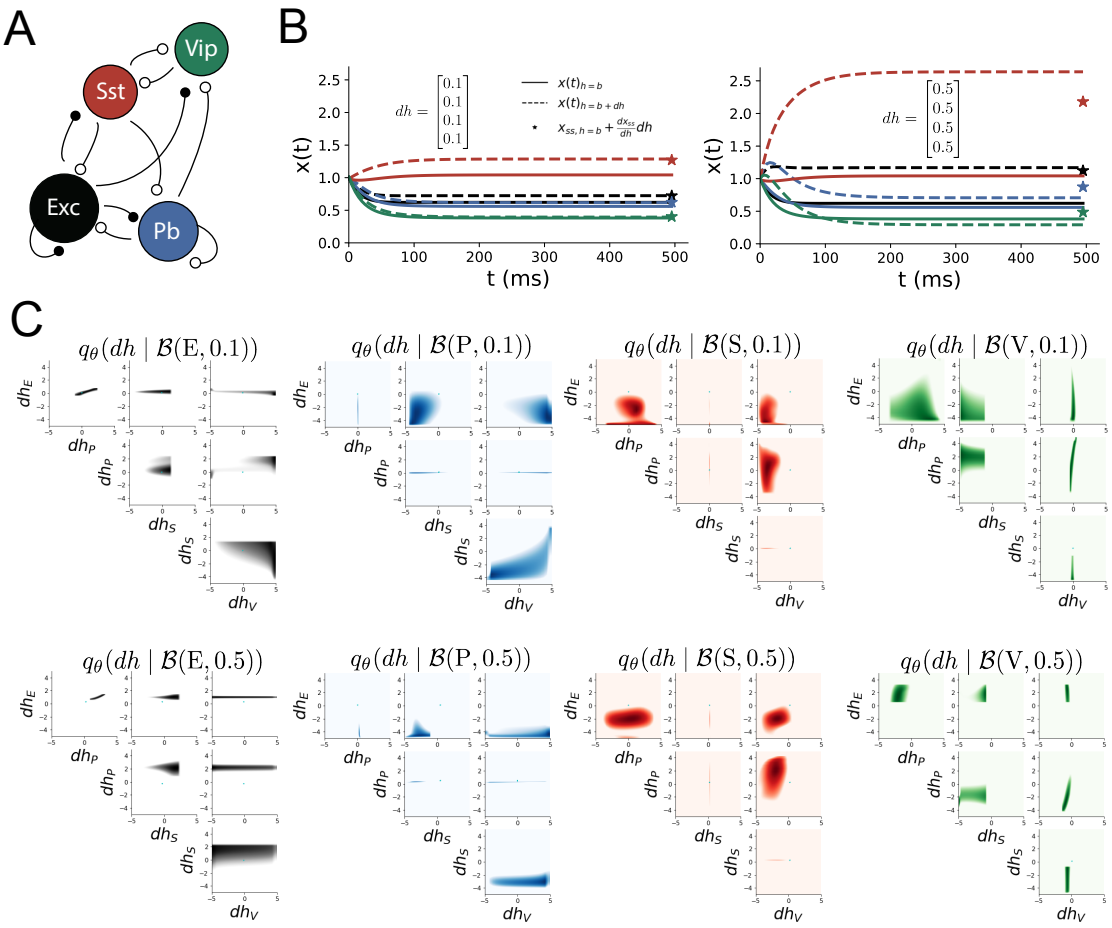

Figure 2: A. Four-population model of primary visual cortex with excitatory (black), parvalbumin (blue), somatostatin (red), and vip (green) neurons. Some neuron types largely do not form synaptic projections to others Excitatory and inhibitory projections filled and unfilled, respectively. B. Linear response predictions become innacurate with greater input strength. V1 model simulations for input (solid) $h = b$ and (dashed) $h = b + dh$ with $b = [1, 1, 1, 1]\top$ and (left) $dh = [0.1, 0.1, 0.1, 0.1]\top$ (right) $dh = [0.5, 0.5, 0.5, 0.5]\top$. Stars indicate the linear response prediction. C. EPI distributions on dh conditioned on neuron type α rate increases of y: $\mathcal{B}(\alpha, y)$ (see Section A.2.2).

that the P-population likely stabilizes the E- and V-populations (the two it projects to besides itself). This deduction is supported by the tight positive correlation between dh_P and dh_E for E-population increases and also between dh_P and dh_V for V-population increases. Finally, EPI showed that negative dh_E results in small inhibitory firing rate increases. However, for a larger increase from the V-population, a positive dh_E is required.

All of this insight was gained beyond what the analytic linear prediction told us (cyan lines that you can't see). These analyses can generate experimentally testable hypotheses, that if confirmed, can be used to build a theory of V1 circuit operation. For example, one could test the P-population's role in stabilization by optogenetically stimulating the E- and V-populations and measuring the P-population response. Additionally, one would predict that at some point, the V-population's response will flip sign as input to the E-population increases.

3.4 Identifying neural mechanisms of behavioral learning.

A key challenge for theorists modeling neural circuits underlying cognitive behavior is the description of sufficient changes to biologically meaningful parameters that result in improved behavoir. Identifying measurable biological changes that should occur for increased performance is critical for neuroscience, since they may indicate how the learning brain adapts. We used EPI to learn connectivities distributions consistent with various levels of rapid task switching accuracy, resulting in a clear picture of connectivity changes which improve rapid task switching. Furthermore, this analysis produced experimentally testable predictions regarding effective connectivity throughout learning of this behavioral paradigm.

In a rapid task switching experiment, where rats were to respond right (R) or left (L) to the side of a light stimulus in the pro (P) task, and oppositely in the anti (A) task predicated by an auditory cue (Fig. 4A), neural recordings exhibited two population of neurons in each hemisphere of superior colliculus (SC) that simultaneously represented both task condition and motor response: the pro/contra and anti/ipsi neurons [21]. Duan et al. proposed a four-population dynamical model of superior colliculus with a Pro- and Anti-population in each hemisphere, where activities were bounded from 0-1, and a high output of the Pro population in a given hemisphere corresponds to the contralateral response. The connectivity matrix is parameterized by the geometry of the population arrangement (Fig. 4B).

We ran EPI to learn appproximate posteriors of SC model weight matrix parameters $z = W$ condi-

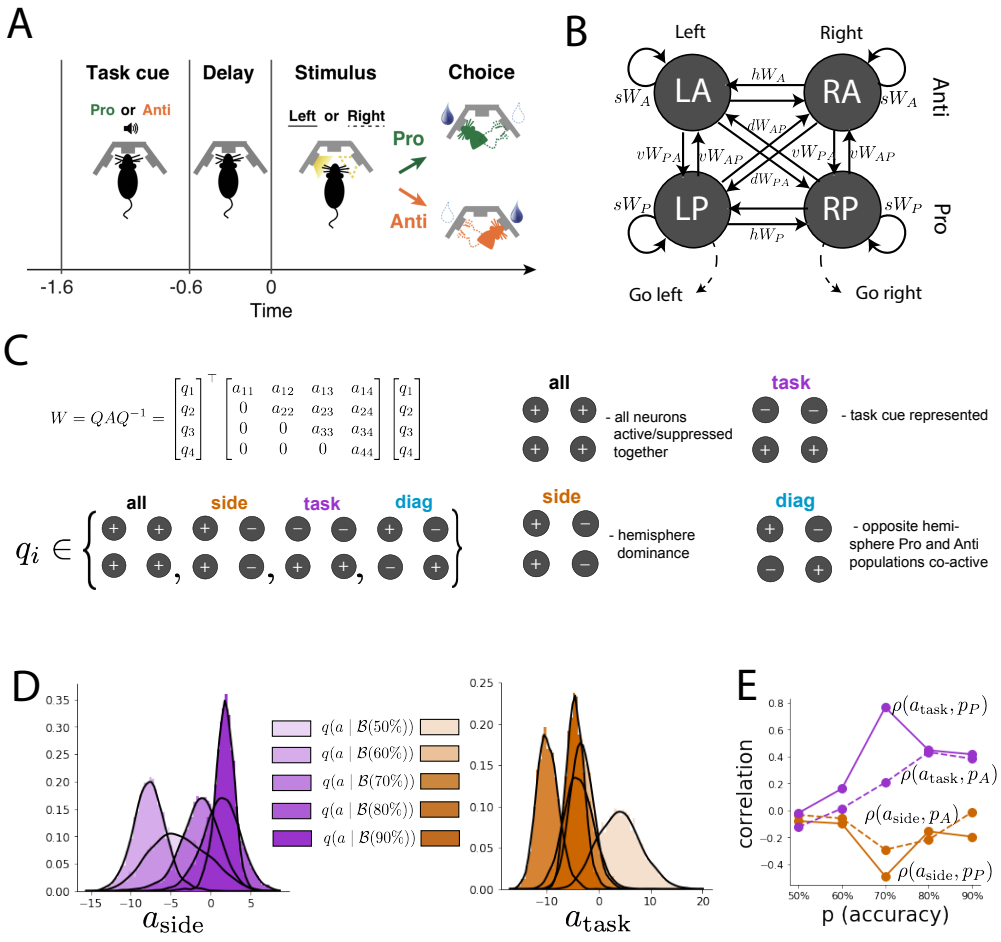

Figure 3: A. Rapid task switching behavioral paradigm. In the Pro (Anti) condition indicated by an auditory cue, the rats are to respond to the same (opposite) side as the light stimulus that is provided after a delay to receive a reward. B. Model of superior collicullus (SC). Neurons: LP - left pro, RP - right pro, LA - left anti, RA - right anti. Parameters: sW - self, hW - horizontal, vW - vertical, dW - diagonal weights. C. The Schur decompostion of the weight matrix. D. The marginal EPI distributions of the Schur eigenvalues at each level of task accuracy. E. The correlation of Schur eigenvalue with task performance.

tioned on of various levels of rapid task switching accuracy $\mathcal{B}(p)$ for $p \in \{50\%, 60\%, 70\%, 80\%, 90\%\}$ (see Section A.2.3). The Schur decomposition of the weight matrix, a unique decomposition revealing the underlying directed structure between modes (the eigenvectors), has the same eigenvectors for all Ws under this symmetric parameterization (Fig. 4C). These consistent Schur eigenvectors have intuitive roles in processing for this task, and are accordingly named the *all*, *side*, *task*, and *diag* modes. The corresponding eigenvalues a of each mode (which change according to W indicate the amplification or supression of activity along that mode.

As learning progresses, the task mode is increasingly amplified, indicating the criticality of a distributed task representation at the time of stimulus presentation, (Fig. 4D, purple). Stepping from task-naive 50% networks to task-performing 60% networks, there is a marked suppression of the side mode (Fig. 4D, orange). Such side mode suppression remains in the regimes of greater accuracy, revealing its importance towards the existence of a distributed task representation. There were no interesting trends with learning in the all or diag mode. We can conclude that side mode suppression allows rapid task switching, and that greater task-mode representation increases accuracy (Fig. 4E). These findings motivate experimental predictions, in which we would expect the effective connectvity between these populations to change throughout learning in a way that increases the task mode and decreses the side mode eigenvalues.

3.5 Characterizing the sources of bias during approximate inference in RNNs

At a more abstract level, recurrent neural networks (RNNs) are high-dimensional models of computation, which have become increasingly popular in systems neuroscience research [32]. Typically, RNNs are trained to do a task from a systems neuroscience experiment, and then the latent factors of the trained RNN are compared to recorded neural activity. Recent theoretical work extends dynamic mean field theory (DMFT) from random [3] to low rank RNNs [22]. This theory establishes a link between interpretable, geometric parameterizations of the RNN connectivity with the emerging dynamics. We used this theory along with EPI to characterize the mechanistic sources of bias during approximate Bayesian inference in rank-1 RNNs.

The connectivity of a rank-1 RNN is the sum of a random component with strength determined by g, and a structured component determined by the outer product of vectors m and n:

$$J = g\chi + \frac{1}{N} m n^\top \tag{4}$$

where $\chi_{ij} \sim \mathcal{N}(0, \frac{1}{N})$ and $m_i \sim \mathcal{N}(M_m, 1)$ and $n_i \sim \mathcal{N}(M_n, 1)$. The rank-1 RNNs were to produce

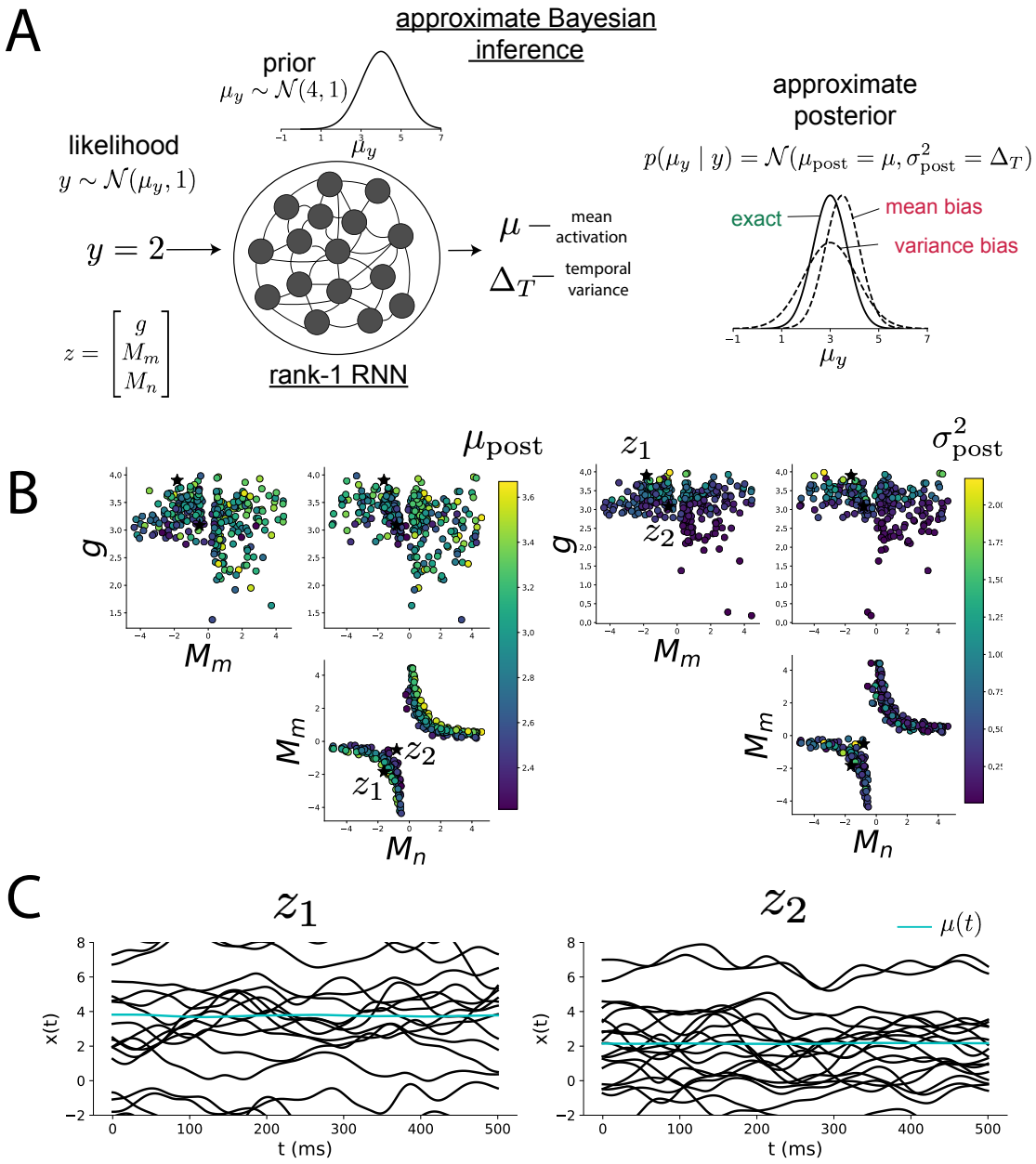

Figure 4: A. We examined a rank-1 RNN running approximate Bayesian inference on μ_y assuming a gaussian likelihood variance of 1 and a prior of $\mathcal{N}(4,1)$. B. Distribution of rank-1 RNNs executing approximate Bayesian inference. Samples are colored by (left) posteior mean $\mu_{\text{post}} = \mu$ and (right) posterior variance $\sigma^2_{\text{post}} = \Delta_T$ C. Finite size realizations agree with the DMFT theory.

the posterior mean μ_{post} and variance σ^2_{post} in their mean activity μ and temporal variance Δ_T (see Section A.2.4). The Bayesian inference problem was to produce the gaussian posterior of the mean μ_y of a gaussian likelihood of observations $y \sim \mathcal{N}(\mu_y, 1)$ given a single observation of $y = 2$ and a prior of $\mu_y \sim \mathcal{N}(4, 1)$ (Fig. 5A). The true posterior to this problem is $\mu_x \sim \mathcal{N}(\mu_{\text{post}} = 3, \sigma^2_{\text{post}} = 0.5)$, although different parameterizations of the connectivity $z = \begin{bmatrix} g & M_m & M_n \end{bmatrix}^\top$ result in approximate inference procedures of varying biases in μ_{post} and σ^2_{post}.

Weran EPI on rank-1 RNNs solving this Bayesian inference problem, whie allowing a substantial amount of variability in the second moment constraints of the network mean μ and temporal variance Δ_T. This allowed us to study the mechanistic sources of bias in the sampled rank-1 RNNs executing the Bayesian inference computation inexactly. The posterior distribution was roughly symmetric in the M_m-M_n plane which structure suggesting there is a degeneracy in the product of M_m and M_n (Fig. 5B). The product of M_m and M_n almost completely determines the posterior mean (Fig. 5B, left), and the random strength g is the most influential variable on the temporal variance (Fig. 5B, right). Neither of these observations were obvious from the consistency equations afforded by DMFT, the solvers of which we took gradients through to run EPI.

It is good practice to check that finite-size realizations of these infinite-size networks match the predictions from DMFT. A 2,000-neuron network with parameters z_1 produced an overestimate of th posterior mean and variance (mean activity (cyan), temporal variance of traces Fig. 5C, left), while a 2,000-neuron network with parameter z_2 produced underestimates (Fig. 5C, right). This novel procedure of doing inference in interpretable parameterizations of RNNs conditioned on abstract cognitive tasks can be generally executed on other tasks like noisy integration and context-dependent decision making (Fig. S1).

3.6 EPI is a general tool for all theoretical neuroscience

There are many considerations when assessing the feasibility of classical statistical inference in neuroscience such as conjugacy, likelihood tractability, scalability to large data-sets. EPI is no exception, but we emphasize that it is practicable in most settings of theoretical neuroscience. Models of close biological fidelity often have complex nonlinear differential equations, making traditional statistical inference intractable. In contrast, EPI is capable of learning distributions of such model parameters producing low-level signatures of computation as we have shown through our analysis of the STG. This approach is not specific to models of such biological realism, as we have demonstrated its utility studying abstract models like RNNs. We are able to condition both deterministic

and stochastic models on all sorts of emergent properties from membrane potential firing to execution of approximate inference. In sum, EPI is feasible when the emergent property statistics are continuously differentiable with respect to the model parameters, which is very often the case. This gradient does not even need to be derived by hand, since we use automatic differentiation tools available in tensorflow[33] and other similar software packages.

In this study, we have focused on applying EPI to low dimensional parameter spaces of models with low dimensional dynamical state. These choices were made to present the reader with a series of interpretable conclusions, which is more challenging in high dimensional spaces. In fact, EPI scales well to high dimensional parameter spaces, as the underlying technology has produced state-of-the-art performance on texture generation [17]. However, increasing the dimensionality of the dynamical state of the model makes optimization more expensive, and there is a practical limit there as with any machine learning approach. For systems with high dimensional state, we recommend using theoretical approaches (e.g. [22]) to reason about reduced parameterizations of such high-dimensional (even infinite dimensional) systems.

There are additional technical considerations when assessing the suitability of EPI for a particular modeling question. First, one should consider how computationally expensive the gradient of the emergent propety statistic is with respect to the model parameters. In the best circumstance, there is a simple, closed form expression (e.g. Section A.1.1) for the emergent property statistic given the model parameters. On the other end of the spectrum, you may require a large number of simulation iterations before a high quality measurement of the emergent property statistic is available (e.g. Section A.2.1). In such cases, optimization will be expensive, and it is worth considering an alternative methodology (see Section A.1.4). Secondly, the defined emergent property should always be appropriately conditioned. Of course, learning can not occur when under- or overconstrained. When underconstrained, the posterior grows (in entropy) unstabily unless mapped to a finite support. If overconstrained, and there is no support producing the emergent property, EPI optimization will never converge.

4 Discussion

Draft in progress:

Machine learning has played an effective, multifaceted role in neuroscientific progress. Primarily, has revealed structure in large scale neural data sets [34, 35, 36, 37, 38, 39] (see review, [14]).

Secondarily, trained algorithms of varying degrees of biological relevace serve as fully-observable computational systems that are compared to the brain [40, **?**]. Theoretical neuroscientists may currently be too focused on this secondary role, and thus missing out on the primary benefit statistical machine learning can offer. Specifically, deep learning for probabilistic inference has matured to a level where theorists can use it to understand their *models* rather than the experimental *data sets* for which it is traditionally purposed.

For example, consider the fact that we do not yet understand just a four-dimensional, deterministic model of V1. This should not be surprising, since analytic approaches to studying nonlinear dynamical systems complexify greatly when stepping from two-dimensional to three- or four-dimensional systems in the absence of restrictive simplifying assumptions [31]. We are not suggesting to forego the development or application of challenging analytic procedures in theoretical neuroscience. However, we suggest the judicious recognition of arduous mathematical challenges, and alternatively using deep learning through EPI to gain the desired insights. In Section **??**, we showed that EPI was far more informative about neuron-type input responsivity than the predictions afforded through analysis. By flexibly conditioning this V1 model on different emergent properties, we performed an exploratory analysis of a *model* rather than a data set, which generated a set of testable predictions.

Exploratory analyses in theoretical neuroscience can certainly be less agnostic to the eventual hypothesis generated. When interested in the mechanistic changes that occur in the brain throughout learning, one can use EPI to condition on various levels of an emergent property statistic indicative of performance like task accuracy in a behavioral paradigm (see Section **??**). This analysis identified experimentally testable predictions of changes in connectivity in SC throughout rapid task switching learning. Precisely, we would predict an initial reduction in side mode eigenvalue, and a steady increase in task mode eigenvalue of such effective connectivity matrices.

While experimentally testable predictions are highly valuable, sometimes it is prohibitively challenging to design a biologically realistic model of a neural computation. Thusly, RNNs have become an increasingly popular tool in systems neuroscience research. The scientific philosophy is as follows: optimize an RNN to execute a task from behavioral neuroscience, compare the activity of this optimized system to brain activity from a model organism doing the same task, and leverage the full observability of the trained RNN to generate hypotheses of the neural mechanisms of computation. While fixed point identification and jacobian measurement yield intuitive, consistent portraits of the implemented computational algorithm [**?**], there is dizzying degeneracy in the RNN connectivity matrix with respect to these characterizations. Since neural activity generally lies on

a low dimensional manifold [**?**], we may attain an understanding of the neural mechanisms at play in cortical processing by working in a reduced, interpretable parameter setting of these powerfully general models [42].

In our final analysis, we present a novel procedure for doing statistical inference on interpretable parameterizations of RNNs executing tasks from behavioral neuroscience. This methodology relies on recently extended theory of responses in random neural networks with minimal structure [22]. This theory makes a direct like between a geometric description of the connectivity and the emerging dynamics. These emerging dynamics in response to various inputs can be cast to perform noisy detection, context-dependent evidence integration and more. With this methodology, we can finally open the probabilistic model selection toolkit reasoning about the connectivity of RNNs solving tasks.

Some statements to inject somewhere

- If theory/practice of deep learning improves to a point where we converge to global optima more regularly, we could do cool reasoning about models (the hypothetico-deductive stuff from Gelman/Shalizi) using these max-ent distributions.
- We can think of the probability in these models as deviation from the mean constraint in the sufficient statistics. See this proved out with the 2D LDS example.

References

[1] Larry F Abbott. Theoretical neuroscience rising. *Neuron*, 60(3):489–495, 2008.

[2] John J Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences*, 81(10):3088–3092, 1984.

[3] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural networks. *Physical review letters*, 61(3):259, 1988.

[4] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Paradoxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*, 17(11):4382–4388, 1997.

[5] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.

[6] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and variational inference in deep latent gaussian models. *International Conference on Machine Learning*, 2014.

[7] Yuanjun Gao, Evan W Archer, Liam Paninski, and John P Cunningham. Linear dynamical neural population models through nonlinear embeddings. In *Advances in neural information processing systems*, pages 163–171, 2016.

[8] Yuan Zhao and Il Memming Park. Recursive variational bayesian dual estimation for nonlinear dynamics and non-gaussian observations. *stat*, 1050:27, 2017.

[9] Gabriel Barello, Adam Charles, and Jonathan Pillow. Sparse-coding variational auto-encoders. *bioRxiv*, page 399246, 2018.

[10] Chethan Pandarinath, Daniel J O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky, Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods*, page 1, 2018.

[11] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta. Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.

[12] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.

[13] Eleanor Batty, Matthew Whiteway, Shreya Saxena, Dan Biderman, Taiga Abe, Simon Musall, Winthrop Gillis, Jeffrey Markowitz, Anne Churchland, John Cunningham, et al. Behavenet: nonlinear embedding and bayesian neural decoding of behavioral videos. *Advances in Neural Information Processing Systems*, 2019.

[14] Liam Paninski and John P Cunningham. Neural data science: accelerating the experiment-analysis-theory cycle in large-scale neuroscience. *Current opinion in neurobiology*, 50:232–241, 2018.

[15] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[16] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *International Conference on Machine Learning*, 2015.

[17] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. Maximum entropy flow networks. *International Conference on Learning Representations*, 2017.

[18] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-free variational inference. In *Advances in Neural Information Processing Systems*, pages 5523–5533, 2017.

[19] Gabrielle J Gutierrez, Timothy O'Leary, and Eve Marder. Multiple mechanisms switch an electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators. *Neuron*, 77(5):845–858, 2013.

[20] Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and visual coding in cortical circuits with multiple interneuron subtypes. *Journal of neurophysiology*, 115(3):1399–1409, 2016.

[21] Chunyu A Duan, Marino Pagan, Alex T Piet, Charles D Kopec, Athena Akrami, Alexander J Riordan, Jeffrey C Erlich, and Carlos D Brody. Collicular circuits for flexible sensorimotor routing. *bioRxiv*, page 245613, 2018.

[22] Francesca Mastrogiuseppe and Srdjan Ostojic. Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron*, 99(3):609–623, 2018.

[23] Sean R Bittner, Agostina Palmigiano, Kenneth D Miller, and John P Cunningham. Degenerate solution networks for theoretical neuroscience. *Computational and Systems Neuroscience Meeting (COSYNE), Lisbon, Portugal*, 2019.

[24] Sean R Bittner, Alex T Piet, Chunyu A Duan, Agostina Palmigiano, Kenneth D Miller, Carlos D Brody, and John P Cunningham. Examining models in theoretical neuroscience with degenerate solution networks. *Bernstein Conference*, 2019.

[25] Jan-Matthis Lueckmann, Pedro Goncalves, Chaitanya Chintaluri, William F Podlaski, Giacomo Bassetto, Tim P Vogels, and Jakob H Macke. Amortised inference for mechanistic models

of neural dynamics. In *Computational and Systems Neuroscience Meeting (COSYNE), Lisbon, Portugal*, 2019.

[26] Brendan K Murphy and Kenneth D Miller. Balanced amplification: a new mechanism of selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009.

[27] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. Inhibitory stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–592, 2009.

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

[29] Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*, 5(10):793, 2004.

[30] Bernardo Rudy, Gordon Fishell, SooHyun Lee, and Jens Hjerling-Leffler. Three groups of interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neurobiology*, 71(1):45–61, 2011.

[31] Steven H Strogatz. Nonlinear dynamics and chaos: with applications to physics. *Biology, Chemistry, and Engineering (Studies in Nonlinearity), Perseus, Cambridge, UK*, 1994.

[32] Omri Barak. Recurrent neural networks as versatile tools of neuroscience research. *Current opinion in neurobiology*, 46:1–6, 2017.

[33] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[34] Robert E Kass and Valérie Ventura. A spike-train probability model. *Neural computation*, 13(8):1713–1720, 2001.

[35] Emery N Brown, Loren M Frank, Dengda Tang, Michael C Quirk, and Matthew A Wilson. A statistical paradigm for neural spike train decoding applied to position prediction from

ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18(18):7411–7425, 1998.

[36] Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15(4):243–262, 2004.

[37] M Yu Byron, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. In *Advances in neural information processing systems*, pages 1881–1888, 2009.

[38] Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science*, 349(6244):184–187, 2015.

[39] Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable continuous-time models of latent stochastic dynamical systems. *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[40] David Sussillo and Omri Barak. Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural computation*, 25(3):626–649, 2013.

[41] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356, 2016.

[42] Kenji Doya. Universality of fully connected recurrent neural networks. *Dept. of Biology, UCSD, Tech. Rep*, 1993.

[43] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

[44] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[45] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[46] Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons. *Nature neuroscience*, 16(8):1068, 2013.

A Methods

A.1 Emergent property inference (EPI)

Draft in progress:

Emergent property inference (EPI) learn distributions of theoretical model parameters that produce emergent properties of interest. They combine ideas from likelihood-free variational inference [18] and maximum entropy flow networks [17]. A maximum entropy flow network is used as a deep probability distribution for the parameters, while these samples are passed through a differentiable model simulator, which may lack a tractable likelihood function.

Consider model parameterization z and data x generated from some theoretical model simulator represented as $p(x \mid z)$, which may be deterministic or stochastic. Theoretical models usually have known sampling procedures for simulating activity given a circuit parameterization, yet often lack an explicit likelihood function due to the nonlinearities and dynamics. With EPI, a distribution on parameters z is learned, that yields a behavior of interest $\mathcal{B}$,

$$\mathcal{B} : E_{z \sim q_\theta} \left[E_{x \sim p(x|z)} \left[T(x) \right] \right] = \mu \tag{5}$$

by making an approximation $q_\theta(z)$ to $p(z \mid \mathcal{B})$ (see Section A.1.5). So, over the DSN distribution $q_\theta(z)$ of model $p(x \mid z)$ for behavior $\mathcal{B}$, the emergent properties $T(x)$ are constrained in expectation to μ.

In deep probability distributions, a simple random variable $w \sim p_0$ is mapped deterministically via a function f_θ parameterized by a neural network to the support of the distribution of interest where $z = f_\theta(\omega) = f_l(..f_1(\omega))$. Given a theoretical model $p(x \mid z)$ and some behavior of interest $\mathcal{B}$, the deep probability distributions are trained by optimizing the neural network parameters θ to find a good approximation q_θ^* within the deep variational family Q to $p(z \mid \mathcal{B})$.

In most settings (especially those relevant to theoretical neuroscience) the likelihood of the behavior with respect to the model parameters $p(T(x) \mid z)$ is unknown or intractable, requiring an alternative to stochastic gradient variational bayes [5] or black box variational inference[43]. These types of methods called likelihood-free variational inference (LFVI, cite Tran) skate around the intractable likelihood function in situations where there is a differentiable simulator. Akin to LFVI, DSNs are optimized with the following objective for a given generative model and statistical constraints on its produced activity:

$$q_\theta^*(z) = \underset{q_\theta \in Q}{\operatorname{argmax}} H(q_\theta(z)) \quad \text{s.t. } E_{z\sim q_\theta}\left[E_{x\sim p(x|z)}\left[T(x)\right]\right] = \mu \tag{6}$$

TODO expand on this in terms of Fig. 2, in a way that complements 3.1.

A.1.1 Example: 2D LDS

Draft in progress:

To gain intuition for EPI, consider two-dimensional linear dynamical systems, $\tau\dot{x} = Ax$ with

$$A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}$$

that produce a band of oscillations. To do EPI with the dynamics matrix elements as the free parameters $z = [a_1, a_2, a_3, a_4]$, and fixing $\tau = 1$, such that the posterior yields a band of oscillations, the emergent property statistics $T(x)$ are chosen to contain the first- and second-moments of the oscillatory frequency Ω and the growth/decay factor d of the oscillating system (the real part of the complex conjugate pairs of eigenvalues). To learn the distribution of real entries of A that yield a distribution of d with mean zero with variance 1, and oscillation frequency $\Omega$with mean 1 Hz with variance 1, the emergent property values would be set to:

$$\mu = E\begin{bmatrix} d \\ \Omega \\ (d-0)^2 \\ (\Omega-1)^2 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 1.0 \\ 1.0 \\ 1.025 \end{bmatrix} \tag{7}$$

To obtain a differntiable estimate of the oscillation frequency with respect to the dynamics matrices, we could simuilate system activity x from $z = A$ for some finite number of time steps, and estimate Ω by e.g. taking the peak of the discrete Fourier transform. Instead, the emergent property statistics for this oscillating behavior are computed through a closed form function $g(z)$ by taking the eigendecomposition of the dynamics matrix

$$g(z) = E_{x\sim p(x|z)}\left[T(x)\right] = \begin{bmatrix} \text{real}(\lambda_1) \\ \frac{\text{imag}(\lambda_1)}{2\pi} \\ \text{real}(\lambda_1)^2 \\ (\frac{\text{imag}(\lambda_1)}{2\pi})^2 \end{bmatrix} \tag{8}$$

$$\lambda = \frac{(\frac{a_1+a_4}{\tau}) \pm \sqrt{(\frac{a_1+a_4}{\tau})^2 + 4(\frac{a_2a_3-a_1a_4}{\tau})}}{2} \tag{9}$$

where λ_1 is the eigenvalue of $\frac{1}{\tau}A$ with greatest real part. Even though $E_{x\sim p(x|z)}\left[T(x)\right]$ is calculable directly via $g(z)$, we cannot derive the distribution q^*_θ, since the backward mapping from the mean parameters μ to the natural parameters η of his exponential family is unknown [44]. Instead, we can use EPI to learn the linear system parameters producing such a band of oscillations (Fig. S2B).

Even this relatively simple system has nontrivial (though intuitively sensible) structure in the parameter distribution. The contours of the probability density can be derived from the emergent property statistics and values (Fig. S3). In the a_1-a_4 plane, is a black line at $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} = 0$, a dotted black line at the standard deviation $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} \pm 1$, and a grey line at twice the standard deviation $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} \pm 2$ (Fig. S3A). Here the lines denote the set of solutions at fixed behaviors, which overlay the posterior obtained through EPI. The learned DSN distribution precisely reflects the desired statistical constraints and model degeneracy in the sum of a_1 and a_4. Intuitively, the parameters equivalent with respect to emergent property statistic $\text{real}(\lambda_1)$ have similar log densities.

To explain the structure in the bimodality of the DSN posterior, we can look at the imaginary component of λ_1. When $\text{real}(\lambda_1) = \frac{a_1+a_4}{2} = 0$, we have

$$\text{imag}(\lambda_1) = \begin{cases} \sqrt{\frac{a_1a_4-a_2a_3}{\tau}}, & \text{if } a_1a_4 < a_2a_3 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

When $\tau = 1$ and $a_1a_4 > a_2a_3$ (center of distribution above), we have the following equation for the the other two dimensions:

$$\text{imag}(\lambda_1)^2 = a_1a_4 - a_2a_3 \tag{11}$$

Since we constrained $E_{q_\theta}\left[\text{imag}(\lambda)\right] = 2\pi$ (with $\omega = 1$), we can plot contours of the equation $\text{imag}(\lambda_1)^2 = a_1a_4 - a_2a_3 = (2\pi)^2$ for various a_1a_4 (Fig. S3A). If $\sigma_{1,4} = E_{q_\theta}(|a_1a_4 - E_{q_\theta}[a_1a_4]|)$, then we plot the contours as $a_1a_4 = 0$ (black), $a_1a_4 = -\sigma_{1,4}$ (black dotted), and $a_1a_4 = -2\sigma_{1,4}$ (grey dotted) (Fig. S3B). We take steps in negative standard deviation of a_1a_4 (dotted and gray lines), since there are few positive values a_1a_4 in the posterior. More subtle model-behavior combinations will have even more complexity, further motivating the use of EPI for understanding these systems.

For futher validation of the underlying technology, see recovery of ground truth distributions with maximum entorpy flow networks [17].

A.1.2 Augmented Lagrangian optimization

Draft in progress:

To optimize $q_\theta(z)$ in equation 1, the constrained optimization is performed using the augmented Lagrangian method. The following objective is minimized:

$$L(\theta; \alpha, c) = -H(q_\theta) + \alpha^\top \delta(\theta) + \frac{c}{2}||\delta(\theta)||^2 \tag{12}$$

where $\delta(\theta) = E_{z\sim q_\theta}\left[E_{x\sim p(x|z)}\left[T(x) - \mu\right]\right]$, $\alpha \in \mathcal{R}^m$ are the Lagrange multipliers and c is the penalty coefficient. For a fixed (α, c), θ is optimized with stochastic gradient descent. A low value of c is used initially, and increased during each augmented Lagrangian epoch – a period of optimization with fixed *alpha* and c for a given number of stochastic optimziation iterations. Similarly, α is tuned each epoch based on the constraint violations. For the linear 2-dimensional system (Fig. S4C) optimization hyperparameters are initialized to $c_1 = 10^{-4}$ and $\alpha_1 = 0$. The penalty coefficient is updated based on a hypothesis test regarding the reduction in constraint violation. The p-value of $E[||\delta(\theta_{k+1})||] > \gamma E[||\delta(\theta_k)||]$ is computed, and c_{k+1} is updated to βc_k with probability $1 - p$. Throughout the project, $\beta = 4.0$ and $\gamma = 0.25$ is used. The other update rule is $\alpha_{k+1} = \alpha_k + c_k \frac{1}{n}\sum_{i=1}^{n}(T(x^{(i)}) - \mu)$. In this example, each augmented Lagrangian epoch ran for 5,000 iterations. We consider the optimization to have converged when a null hypothesis test of constraint violations being zero is accepted for all constraints at a significance threshold 0.05. This is the dotted line on the plots below depicting the optimization cutoff of EPI optimization for the 2-dimensional linear system. If the optimization is left to continue running, entropy usually decreases, and structural pathologies in the distribution may be introduced.

The intention is that c and λ start at values encouraging entropic growth early in optimization. Then, as they increase in magnitude with each training epoch, the constraint satisfaction terms are increasingly weighted, resulting in a decrease in entropy. Rather than using a naive initialization, before EPI, we optimize the deep probability distribution parameters to generate samples of an isotropic gaussian of a selected variance, such as 1.0 for the 2D LDS example. This provides a convenient starting point, whose level of entropy is controlled by the user.

A.1.3 Normalizing flows

Draft in progress:

Since we are optimizing parameters θ of our deep probability distribution with respect to the entropy, we will need to take gradients with respect to the log-density of samples from the deep probability distribution.

$$H(q_\theta(z)) = \int -q_\theta(z) \log(q_\theta(z)) dz = E_{z\sim q_\theta}\left[-\log(q_\theta(z))\right] = E_{\omega\sim q_0}\left[-\log(q_\theta(f_\theta(\omega)))\right] \tag{13}$$

$$\nabla_\theta H(q_\theta(z)) = E_{\omega\sim q_0}\left[-\nabla_\theta \log(q_\theta(f_\theta(\omega)))\right] \tag{14}$$

Deep probability models typically consist of several layers of fully connected neural networks. When each neural network layer is restricted to be a bijective function, the sample density can be calculated using the change of variables formula at each layer of the network. For $z' = f(z)$,

$$q(z') = q(f^{-1}(z')) \left|\det \frac{\partial f^{-1}(z')}{\partial z'}\right| = q(z) \left|\det \frac{\partial f(z)}{\partial z}\right|^{-1} \tag{15}$$

However, this computation has cubic complexity in dimensionality for fully connected layers. By restricting our layers to normalizimg flows [16] – bijective functions with fast log determinant jacobian computations, we can tractably optimize deep generative models with objectives that are a function of sample density, like entropy. Most of our analyses use real NVP [45], which have proven effective in our architecture searches, and have the advantageous features of fast sampling and fast density evaluation.

A.1.4 Related work

Draft in progress:

A.1.5 Emergent property inference as variational inference in an exponential family

Draft in progress:

Consider the goal of doing variational inference (VI) in with an exponential family posterior distribution $p(z \mid x)$. We'll use the following abbreviated notation to collect the base measure and

sufficient statistics into $\tilde{T}(z)$ and likewise concatenate a 1 onto the end of the natural parameter $\tilde{\eta}(x)$. The log normalizing constant $A(\eta(x))$ will remain unchanged.

$$p(z \mid x) = b(z) \exp\Big(\eta(x)^\top T(z) - A(\eta(x))\Big) = \exp\left(\begin{bmatrix}\eta(x)\\ 1\end{bmatrix}^\top \begin{bmatrix}T(z)\\ b(z)\end{bmatrix} - A(\eta(x))\right) \\ = \exp\Big(\tilde{\eta(x)}^\top \tilde{T}(z) - A(\eta(x))\Big) \tag{16}$$

VI looks with an exponential family posterior distribution uses optimization to minimize the following divergence [15]:

$$q_\theta^* = \underset{q_\theta \in Q}{\operatorname{argmin}} KL(q_\theta \mid\mid p(z \mid x)) \tag{17}$$

$q_\theta(z)$ is the variational approximation to the posterior with variational parameters θ. We can write this KL divergence in terms of entropy of the variational approximation.

$$KL(q_\theta \mid\mid p(z \mid x)) = E_{z\sim q_\theta}\left[\log(q_\theta(z))\right] - E_{z\sim q_\theta}\left[\log(p(z \mid x))\right] \tag{18}$$

$$= -H(q_\theta) - E_{z\sim q_\theta}\Big[\tilde{\eta}(x)^\top \tilde{T}(z) - A(\eta(x))\Big] \tag{19}$$

As far as the variational optimization is concerned, the log normalizing constant is independent of $q_\theta(z)$, so it can be dropped.

$$\underset{q_\theta \in Q}{\operatorname{argmin}} KL(q_\theta \mid\mid p(z \mid x)) = \underset{q_\theta \in Q}{\operatorname{argmin}} -H(q_\theta) - E_{z\sim q_\theta}\Big[\tilde{\eta}(x)^\top \tilde{T}(z)\Big] \tag{20}$$

Further, we can write the objective in terms of the first moment of the sufficient statistics $\mu = E_{z\sim p(z|x)}\left[T(z)\right]$.

$$= \underset{q_\theta \in Q}{\operatorname{argmin}} -H(q_\theta) - E_{z\sim q_\theta}\Big[\tilde{\eta}(x)^\top \Big(\tilde{T}(z) - \mu\Big)\Big] + \tilde{\eta(x)}^\top \mu \tag{21}$$

$$= \underset{q_\theta \in Q}{\operatorname{argmin}} -H(q_\theta) - E_{z\sim q_\theta}\Big[\tilde{\eta}(x)^\top \Big(\tilde{T}(z) - \mu\Big)\Big] \tag{22}$$

In emergent property inference (EPI), we're solving the following problem.

$$q_\theta^*(z)y = \underset{q_\theta \in Q}{\operatorname{argmax}} H(q_\theta(z)), \text{ s.t. } E_{z\sim q_\theta}\left[E_{x\sim p(x|z)}\left[T(x)\right]\right] = \mu \tag{23}$$

The lagrangian objective is

$$q_\theta^* = \underset{q_\theta \in Q}{\operatorname{argmin}} -H(q_\theta) + \alpha^\top \Big(E_{z\sim q_\theta}\Big[\tilde{T}(z)\Big] - \mu\Big) \tag{24}$$

As the lagrangian optimization proceeds, α should converge to $\tilde{\eta}(x))$ through its adaptations in each epoch. More formally, $\tilde{\eta}(x) \leftrightarrow \tilde{\eta}(\mu)$ is referred to as the backward mapping, and is formally hard to identify [44]. Since this backward mapping is deterministic, conceptually, we can replace $p(z \mid x)$ with $p(z \mid \mu)$. More commonly, we write $p(z \mid \mathcal{B})$ for clarity where $\mathcal{B}$ more explicitly captures the moment constraints of the sufficient statistics.

A.2 Theoretical models

Draft in progress:

In this study, we used emerent property inferrence to examine several models relevant to theoretical neuroscience. Here, we provide the details of each model and the related analyses.

A.2.1 Stomatogastric ganglion

Draft in progress:

Each neuron's membrane potential is the solution of the following differential equation.

$$C_m \frac{\partial x_m}{\partial t} = -\left[h_{leak} + h_{Ca} + h_K + h_{hyp} + h_{elec} + h_{syn}\right] \tag{25}$$

The membrane potential of each neuron is a affected by the leak, Calcium, Potassium, hyperpolarization, electrical and synaptic currents, respectively. The capictance of the circuit is set to $C_m = 1nF$. Each current has an associated reversal potential: $V_{leak} = -40mV$, $V_{Ca} = 100mV$, $V_K = -80mV$, $V_{hyp} = -20mV$, and $V_{syn} = -75mV$. Each current is a function of the difference in membrane and reversal potential multiplied by a conductance:

$$h_{leak} = g_{leak}(x_m - V_{leak}) \tag{26}$$

$$h_{elec} = g_{el}(x_m^{post} - x_m^{pre}) \tag{27}$$

$$h_{syn} = g_{syn} S_\infty^{pre}(x_m^{post} - V_{syn}) \tag{28}$$

$$h_{Ca} = g_{Ca} M_\infty (x_m - V_{Ca}) \tag{29}$$

$$h_K = g_K N (x_m - V_K) \tag{30}$$

$$h_{hyp} = g_h H (x_m - V_{hyp}) \tag{31}$$

where g_{el} and g_{syn} are DSN-focused parameters, $g_{leak} = 1 \times 10^{-4} \mu S$, and g_{Ca}, g_K, and g_{hyp} have different values based on fast, intermediate (hub) or slow neuron. Fast: $g_{Ca} = 1.9 \times 10^{-2}$, $g_K = 3.9 \times 10^{-2}$, and $g_{hyp} = 2.5 \times 10^{-2}$. Intermediate: $g_{Ca} = 1.7 \times 10^{-2}$, $g_K = 1.9 \times 10^{-2}$, and $g_{hyp} = 8.0 \times 10^{-3}$. Intermediate: $g_{Ca} = 8.5 \times 10^{-3}$, $g_K = 1.5 \times 10^{-2}$, and $g_{hyp} = 1.0 \times 10^{-2}$.

The Calcium, Potassium, and hyperpolarization channels have time-dependent gating dynamics dependent on steady-state gating varibles M_∞, N_∞ and H_∞, respectively.

$$M_\infty = 0.5\left(1 + \tanh\left(\frac{x_m - v_1}{v_2}\right)\right) \tag{32}$$

$$\frac{\partial N}{\partial t} = \lambda_N(N_\infty - N) \tag{33}$$

$$N_\infty = 0.5\left(1 + \tanh\left(\frac{x_m - v_3}{v_4}\right)\right) \tag{34}$$

$$\lambda_N = \phi_N \cosh\left(\frac{x_m - v_3}{2v_4}\right) \tag{35}$$

$$\frac{\partial H}{\partial t} = \frac{(H_\infty - H)}{\tau_h} \tag{36}$$

$$H_\infty = \frac{1}{1 + \exp\left(\frac{x_m + v_5}{v_6}\right)} \tag{37}$$

$$\tau_h = 272 - \left(\frac{-1499}{1 + \exp\left(\frac{-x_m + v_7}{v_8}\right)}\right) \tag{38}$$

where $v_1 = 0mV$, $v_2 = 20mV$, $v_3 = 0mV$, $v_4 = 15mV$, $v_5 = 78.3mV$, $v_6 = 10.5mV$, $v_7 = -42.2mV$, $v_8 = 87.3mV$, $v_9 = 5mV$, and $v_{th} = -25mV$.

Finally, there is a synaptic gating variable as well:

$$S_\infty = \frac{1}{1 + \exp\left(\frac{v_{th} - x_m}{v_9}\right)} \tag{39}$$

When the dynamic gating variables are considered, this is actually a 15-dimensional nonlinear dynamical system.

In order to measure the frequency of the hub neuron during EPI, the STG model was simulated for $T = 500$ time steps of $dt = 25ms$. In EPI, since gradients are taken throught the simulation process, the number of time steps are kept modest if possible. The chosen dt and T were the most computationally convenient choices yielding accurate frequency measurement.

Our original approach to measuring frequency was to take the max of the fast Fourier transform (FFT) of the simulated time series. There are a few key considerations here. One is resolution in frequency space. Each FFT entry will correspond to a signal frequency of $\frac{F_s k}{N}$, where N is the number of samples used for the FFT, $F_s = \frac{1}{dt}$, and $k \in [0, 1, ..., N-1]$. Our resolution is improved by increasing N and decreasing dt. Increasing N = T-b, where b is some fixed number of buffer burn-in inititaliziation samples, necessitates an increase in simulation time steps T, which directly increases computational cost. Increasing F_s (decreasing dt) increases system approximation accuracy, but requires more time steps before a full cycle is observed. At the level of $dt = 0.025$, thousands of temporal samples were required for resolution of .01Hz. These challenges in frequency resolution with the discrete Fourier transform motivated the use of an alternative basis of complex

exponentials. Instead, we used a basis of complex exponentials with frequencies from 0.0-1.0 Hz at 0.01Hz resolution, $\Phi = [0.0, 0.01, ..., 1.0]^\top$

Another consideration is that the the frequency spectra of the hub neuron has several peaks. This is due to high-frequency sub-threshold activity. The maximum frequency was often not the firing frequency. Accordingly, subthreshold activity was set to zero, and the whole signal was low-pass filtered with a moving average window of length 20. The signal is subsequently mean centered. After this pre-processing, the maximum frequency in filter bank accurately reflected the firing frequency.

Finally, to differentiate through the maximum frequency identificiation step, we used a sum-of-powers normalization strategy: Let $\mathcal{X}_i \in \mathcal{C}^{|\Phi|}$ be the complex exponential filter bank dot products with the signal $x_i \in \mathcal{R}^N$, where $i \in \{\text{f1}, \text{f2}, \text{hub}, \text{s1}, \text{s2}\}$. The "frequency identification" vector is

$$u_i = \frac{|\mathcal{X}_i|^\alpha}{\sum_{k=1}^{N} |\mathcal{X}_i(k)|^\alpha}$$

. The frequency is then calculated as $\Omega_i = u_i^\top \Phi$ with $\alpha = 100$.

Network syncing, like all other emergent properties in this work, are defined by the emergent property statistics and values. The emergent property statistics are the first- and second-moments of the firing frequencies. The first moments are set to 0.55Hz, while the second moments are set to 0.025Hz^2.

$$E \begin{bmatrix} \Omega_{\text{f1}} \\ \Omega_{\text{f2}} \\ \Omega_{\text{hub}} \\ \Omega_{\text{s1}} \\ \Omega_{\text{s2}} \\ (\Omega_{\text{f1}} - 0.55)^2 \\ (\Omega_{\text{f2}} - 0.55)^2 \\ (\Omega_{\text{hub}} - 0.55)^2 \\ (\Omega_{\text{s1}} - 0.55)^2 \\ (\Omega_{\text{s2}} - 0.55)^2 \end{bmatrix} = \begin{bmatrix} 0.55 \\ 0.55 \\ 0.55 \\ 0.55 \\ 0.55 \\ 0.025^2 \\ 0.025^2 \\ 0.025^2 \\ 0.025^2 \\ 0.025^2 \end{bmatrix} \tag{40}$$

For EPI in Fig 2C, we used a real NVP architecture with two coupling layers. Each coupling layer had two hidden layers of 10 units each.

A.2.2 Primary visual cortex

Draft in progress:

The dynamics of each neural populations average rate $x = \begin{bmatrix} x_E \\ x_P \\ x_S \\ x_V \end{bmatrix}$ are given by:

$$\tau \frac{dx}{dt} = -x + [Wx + h]_+^n \tag{41}$$

Some neuron types largely lack synaptic projections to other neuron types [46], and it is popular to only consider a subset of the effective connectivities [20].

$$W = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} & 0 \\ W_{PE} & W_{PP} & W_{PS} & 0 \\ W_{SE} & 0 & 0 & W_{SV} \\ W_{VE} & W_{VP} & W_{VS} & 0 \end{bmatrix} \tag{42}$$

(TODO: ask Ken about how to introduce these values and what to ref). Estimates of the of the probability of connection and strength of connection from the Allen institute result in an estimate of the effective connectivity:

$$W = \begin{bmatrix} 0.0576 & 0.19728 & 0.13144 & 0 \\ 0.58855 & 0.30668 & 0.4285 & 0 \\ 0.15652 & 0 & 0 & 0.2 \\ 0.13755 & 0.0902 & 0.4004 & 0 \end{bmatrix} \tag{43}$$

We look at how this four-dimensional nonlinear dynamical model of V1 responds to different inputs, and compare the predictions of the linear response to the approximate posteriors obtained through EPI. The input to the system is the sum of a baseline input $b = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}^\top$ and a differential input dh:

$$h = b + dh \tag{44}$$

All simluations of this system had $T = 100$ time points, a time step $dt = 5$ms, and time constant $\tau = 20$ms. And the system was initialized to a random draw $x(0)_i \sim \mathcal{N}(1, 0.01)$.

We can describe the dynamics of this system more generally by

$$\dot{x}_i = -x_i + f(u_i) \tag{45}$$

where the input to each neuron is

$$u_i = \sum_j W_{ij} x_j + h_i \tag{46}$$

Let $F_{ij} = \gamma_i \delta(i,j)$, where $\gamma_i = f'(u_i)$. Then, the linear response is

$$\frac{\partial x_{ss}}{\partial h} = F(W \frac{\partial x_{ss}}{\partial h} + I) \tag{47}$$

which is calculable by

$$\frac{\partial x_{ss}}{\partial h} = (F^{-1} - W)^{-1} \tag{48}$$

The emergent property we considered was the first and second moments of the change in rate dr between the baseline input $h = b$ and $h = b + dh$. We use the following notation to indicate that the emergent property statistics were set to the following values:

$$\mathcal{B}(\alpha, y) \leftrightarrow E \begin{bmatrix} dr_\alpha \\ (dr_\alpha - y)^2 \end{bmatrix} = \begin{bmatrix} y \\ 0.01^2 \end{bmatrix} \tag{49}$$

For each $\mathcal{B}(\alpha, y)$ with $\alpha \in \{E, P, S, V\}$ and $y \in \{0.1, 0.5\}$, we ran EPI with five different random initial seeds using an architecture of four coupling layers, each with two hidden layers of 10 units. We set $c_0 = 10^5$.

A.2.3 Superior colliculus

Draft in progress:

There are four total units: two in each hemisphere corresponding to the PRO/CONTRA and ANTI/IPSI populations. Each unit has an activity (x_i) and internal variable (u_i) related by

$$x_i(t) = \left(\frac{1}{2} \tanh \left(\frac{u_i(t) - \epsilon}{\zeta} \right) + \frac{1}{2} \right) \tag{50}$$

$\epsilon = 0.05$ and $\zeta = 0.5$ control the position and shape of the nonlinearity, repsectively.

We can order the elements of x_i and u_i into vectors x and u with elements

$$x = \begin{bmatrix} x_{LP} \\ x_{LA} \\ x_{RP} \\ x_{RA} \end{bmatrix} \qquad u = \begin{bmatrix} u_{LP} \\ u_{LA} \\ u_{RP} \\ u_{RA} \end{bmatrix} \tag{51}$$

The internal variables follow dynamics:

$$\tau \frac{\partial u}{\partial t} = -u + Wx + h + \sigma \partial B \tag{52}$$

with time constant $\tau = 0.09s$ and gaussian noise $\sigma \partial B$ controlled by the magnitude of $\sigma = 1.0$. The weight matrix has 8 parameters sW_P, sW_A, vW_{PA}, vW_{AP}, hW_P, hW_A, dW_{PA}, and dW_{AP} (Fig. 4B).

$$W = \begin{bmatrix} sW_P & vW_{PA} & hW_P & dW_{PA} \\ vW_{AP} & sW_A & dW_{AP} & hW_A \\ hW_P & dW_{PA} & sW_P & vW_{PA} \\ dW_{AP} & hW_A & vW_{AP} & sW_A \end{bmatrix} \tag{53}$$

The system receives five inputs throughout each trial, which has a total length of 1.8s.

$$h = h_{\text{rule}} + h_{\text{choice-period}} + h_{\text{light}} \tag{54}$$

There are rule-based inputs depending on the condition,

$$h_{\text{P,rule}}(t) = \begin{cases} I_{\text{P,rule}} \begin{bmatrix} 1 & 0 & 0 & 1 \end{bmatrix}^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \tag{55}$$

$$h_{\text{A,rule}}(t) = \begin{cases} I_{\text{A,rule}} \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix}^\top, & \text{if } t \leq 1.2s \\ 0, & \text{otherwise} \end{cases} \tag{56}$$

a choice-period input,

$$h_{\text{choice}}(t) = \begin{cases} I_{\text{choice}} \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}^\top, & \text{if } t > 1.2s \\ 0, & \text{otherwise} \end{cases} \tag{57}$$

and an input to the right or left-side depending on where the light stimulus is delivered.

$$h_{\text{light}}(t) = \begin{cases} I_{\text{light}} \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}^\top, & \text{if } t > 1.2s \text{ and Left} \\ I_{\text{light}} \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix}^\top, & \text{if } t > 1.2s \text{ and Right} \\ 0, & t \leq 1.2s \end{cases} \tag{58}$$

The inpput parameterization was fixed to $I_{\text{P,rule}} = 10$, $I_{\text{A,rule}} = 10$, $I_{\text{choice}} = 2$, and $I_{\text{light}} = 1$

TODO: this is probably a good place to explain the intuition behind the naming of the Schur eigenmodes.

To produce a Bernoulli rate of p_{LP} in the Left, Pro condition (we can generalize this to either cue, or stimulus condition), let $\hat{p}_i$ be the empirical average steady state (ss) response (final V_{LP} at end of task) over M=500 gaussian noise draws for a given SC model parameterization z_i:

$$\hat{p}_i = E_{\sigma\partial B}\left[x_{LP,\text{ss}} \mid s = L, c = P, z_i\right] = \frac{1}{M}\sum_{j=1}^{M} x_{LP,\text{ss}}(s = L, c = P, z_i, \sigma\partial B_j) \tag{59}$$

For the first constraint, the average over posterior samples (from $q_\theta(z)$) to be p_{LP}:

$$E_{z_i\sim q_\phi}\left[E_{\sigma\partial B}\left[x_{LP,\text{ss}} \mid s = L, c = P, z_i\right]\right] = E_{z_i\sim q_\phi}\left[\hat{p}_i\right] = p_{LP} \tag{60}$$

We can then ask that the variance of the steady state responses across gaussian draws, is the Bernoulli variance for the empirical rate $\hat{p}_i$.

$$E_{z\sim q\phi}\left[\sigma^2_{err}\right] = 0 \tag{61}$$

$$\sigma^2_{err} = Var_{\sigma\partial B}\left[x_{LP,\text{ss}} \mid s = L, c = P, z_i\right] - \hat{p}_i(1 - \hat{p}_i) \tag{62}$$

We have an additional constraint that the Pro neuron on the opposite hemisphere should have the opposite value. We can enforce this with a final constraint:

$$E_{z\sim q\phi}\left[d_P\right] = 1 \tag{63}$$

$$E_{\sigma\partial W}\left[\left(x_{LP,\text{ss}} - x_{RP,\text{ss}}\right)^2 \mid s = L, c = P, z_i\right] \tag{64}$$

We refer to networks obeying these constraints as Bernoulli, winner-take-all networks. Since the maximum variance of a random variable bounded from 0 to 1 is the Bernoulli variance ($\hat{p}(1-\hat{p})$), and the maximum squared difference between to variables bounded from 0 to 1 is 1, we do not need to control the second moment of these test statistics. In reality, these variables are dynamical system states and can only exponentially decay (or saturate) to 0 (or 1), so the Bernoulli variance error and squared difference constraints can only be undershot. This is important to be mindful of when evaluating the convergence criteria. Instead of using our usual hypothesis testing criteria for convergence to the emergent property, we set a slack variable threshold for these technically infeasible constraints to 0.05.

Training DSNs to learn distributions of dynamical system parameterizations that produce Bernoulli responses at a given rate (with small variance around that rate) was harder to do than expected. There is a pathology in this optimization setup, where the learned distribution of weights is bimodal attributing a fraction p of the samples to an expansive mode (which always sends x_{LP} to 1), and a fraction $1-p$ to a decaying mode (which always sends x_{LP} to 0). This pathology was avoided using an inequality constraint prohibiting parameter samples that resulted in low variance of responses across noise.

771 In total, the emergent property of rapid task switching accuracy at level p was defined as

$$\mathcal{B}(p) \leftrightarrow \begin{bmatrix} \hat{p}_P \\ \hat{p}_A \\ (\hat{p}_P - p)^2 \\ (\hat{p}_A - p)^2 \\ \sigma^2_{P,err} \\ \sigma^2_{A,err} \\ d_P \\ d_A \end{bmatrix} = \begin{bmatrix} p \\ p \\ 0.15^2 \\ 0.15^2 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \tag{65}$$

772 For each accuracy level p, we ran EPI for 10 different random seeds and selected the maximum
773 entropy solution using an architecture of 10 planar flows with $c_0 = 2$.

774 A.2.4 Rank-1 RNN

775 **Draft in progress**:

776 The network dynamics of neuron i's rate x evolve according to:

$$\dot{x}_i(t) = -x_i(t) + \sum_{j=1}^{N} J_{ij}\phi(x_j(t)) + I_i \tag{66}$$

777 where the connectivity is comprised of a random and structured component:

$$J_{ij} = g\chi_{ij} + P_{ij} \tag{67}$$

778 The random all-to-all component has elements drawn from $\chi_{ij} \sim \mathcal{N}(0, \frac{1}{N})$, and the structured
779 component is a sum of r unit rank terms:

$$P_{ij} = \sum_{k=1}^{r} \frac{m_i^{(k)} n_j^{(k)}}{N} \tag{68}$$

780 We use this theory to compute $T(x)$ when running EPI.

781 Rank-1 vectors m and n have elements drawn

$$m_i \sim \mathcal{N}(M_m, \Sigma_m)$$

782

$$n_i \sim \mathcal{N}(M_n, \Sigma_n)$$

783 The current has the following statistics:

$$I = M_I + \frac{\Sigma_{mI}}{\Sigma_m} x_1 + \frac{\Sigma_{nI}}{\Sigma_n} x_2 + \Sigma_\perp h$$

where x_1, x_2, and h are standard normal random variables.

The $\ddot{\Delta}$ equation is broken into the equation for Δ_0 and Δ_∞ by the autocorrelation dynamics assertions.

$$\ddot{\Delta(\tau)} = -\frac{\partial V}{\partial \Delta}$$

$$\ddot{\Delta} = \Delta - \{g^2\langle[\phi_i(t)\phi_i(t+\tau)]\rangle + \Sigma_m^2\kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2\}$$

We can write out the potential function by integrating the negated RHS.

$$V(\Delta, \Delta_0) = \int \mathcal{D}\Delta \frac{\partial V(\Delta, \Delta_0)}{\partial \Delta}$$

$$V(\Delta, \Delta_0) = -\frac{\Delta^2}{2} + g^2\langle[\Phi_i(t)\Phi_i(t+\tau)]\rangle + (\Sigma_m^2\kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2)\Delta + C$$

We assume that as time goes to infinity, the potential relaxes to a steady state.

$$\frac{\partial V(\Delta_\infty, \Delta_0)}{\partial \Delta} = 0$$

$$\frac{\partial V(\Delta_\infty, \Delta_0)}{\partial \Delta} = -\Delta + \{g^2\langle[\phi_i(t)\phi_i(t+\infty)]\rangle + \Sigma_m^2\kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2\} = 0$$

$$\Delta_\infty = g^2\langle[\phi_i(t)\phi_i(t+\infty)]\rangle + \Sigma_m^2\kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2$$

$$\Delta_\infty = g^2 \int \mathcal{D}z \left[\int \mathcal{D}x \phi(\mu + \sqrt{\Delta_0 - \Delta_\infty}x + \sqrt{\Delta_\infty}z\right]^2 + \Sigma_m^2\kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2$$

Also, we assume that the energy of the system is perserved throughout the entirety of its evolution.

$$V(\Delta_0, \Delta_0) = V(\Delta_\infty, \Delta_0)$$

$$-\frac{\Delta_0^2}{2} + g^2\langle[\Phi_i(t)\Phi_i(t)]\rangle + (\Sigma_m^2\kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2)\Delta_0 + C = -\frac{\Delta_\infty^2}{2} + g^2\langle[\Phi_i(t)\Phi_i(t)]\rangle + (\Sigma_m^2\kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2)\Delta_\infty + C$$

$$\frac{\Delta_0^2 - \Delta_\infty^2}{2} = g^2\left(\langle[\Phi_i(t)\Phi_i(t)]\rangle - \langle[\Phi_i(t)\Phi_i(t)]\rangle\right) + (\Sigma_m^2\kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2)(\Delta_0 - \Delta_\infty)$$

$$\frac{\Delta_0^2 - \Delta_\infty^2}{2} = g^2\left(\int \mathcal{D}z \Phi^2(\mu + \sqrt{\Delta_0}z) - \int \mathcal{D}z \int \mathcal{D}x \Phi(\mu + \sqrt{\Delta_0 - \Delta_\infty}x + \sqrt{\Delta_\infty}z)\right)$$
$$+(\Sigma_m^2\kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2)(\Delta_0 - \Delta_\infty)$$

Consistency equations:

$$
\begin{aligned}
\mu &= F(\mu, \kappa, \Delta_0, \Delta_\infty) = M_m \kappa + M_I \\
\kappa &= G(\mu, \kappa, \Delta_0, \Delta_\infty) = M_n \langle [\phi_i] \rangle + \Sigma_{nI} \langle \left[\phi_i'\right] \rangle \\
\frac{\Delta_0^2 - \Delta_\infty^2}{2} &= H(\mu, \kappa, \Delta_0, \Delta_\infty) = g^2 \left(\int \mathcal{D}z \Phi^2(\mu + \sqrt{\Delta_0} z) - \int \mathcal{D}z \int \mathcal{D}x \Phi(\mu + \sqrt{\Delta_0 - \Delta_\infty} x + \sqrt{\Delta_\infty} z) \right) \\
&\quad + (\Sigma_m^2 \kappa^2 + 2\Sigma_{mI} \kappa + \Sigma_I^2)(\Delta_0 - \Delta_\infty) \\
\Delta_\infty &= L(\mu, \kappa, \Delta_0, \Delta_\infty) = g^2 \int \mathcal{D}z \left[\int \mathcal{D}x \phi(\mu + \sqrt{\Delta_0 - \Delta_\infty} x + \sqrt{\Delta_\infty} z \right]^2 + \Sigma_m^2 \kappa^2 + 2\Sigma_{mI} \kappa + \Sigma_I^2
\end{aligned}
\tag{69}
$$

We can solve these equations by simulating the following Langevin dynamical system.

$$
\begin{aligned}
x(t) &= \frac{\Delta_0(t)^2 - \Delta_\infty(t)^2}{2} \\
\Delta_0(t) &= \sqrt{2x(t) + \Delta_\infty(t)^2} \\
\dot{\mu}(t) &= -\mu(t) + F(\mu(t), \kappa(t), \Delta_0(t), \Delta_\infty(t)) \\
\dot{\kappa}(t) &= -\kappa + G(\mu(t), \kappa(t), \Delta_0(t), \Delta_\infty(t)) \\
\dot{x}(t) &= -x(t) + H(\mu(t), \kappa(t), \Delta_0(t), \Delta_\infty(t)) \\
\dot{\Delta_\infty}(t) &= -\Delta_\infty(t) + L(\mu(t), \kappa(t), \Delta_0(t), \Delta_\infty(t))
\end{aligned}
\tag{70}
$$

Then, the temporal variance is simply calculated via

$$
\Delta_T = \Delta_0 - \Delta_\infty \tag{71}
$$

TODO Need to explain the warm starting for the afficionados.

TODO explain the density network architectures used.

A.3 Supplementary Figures

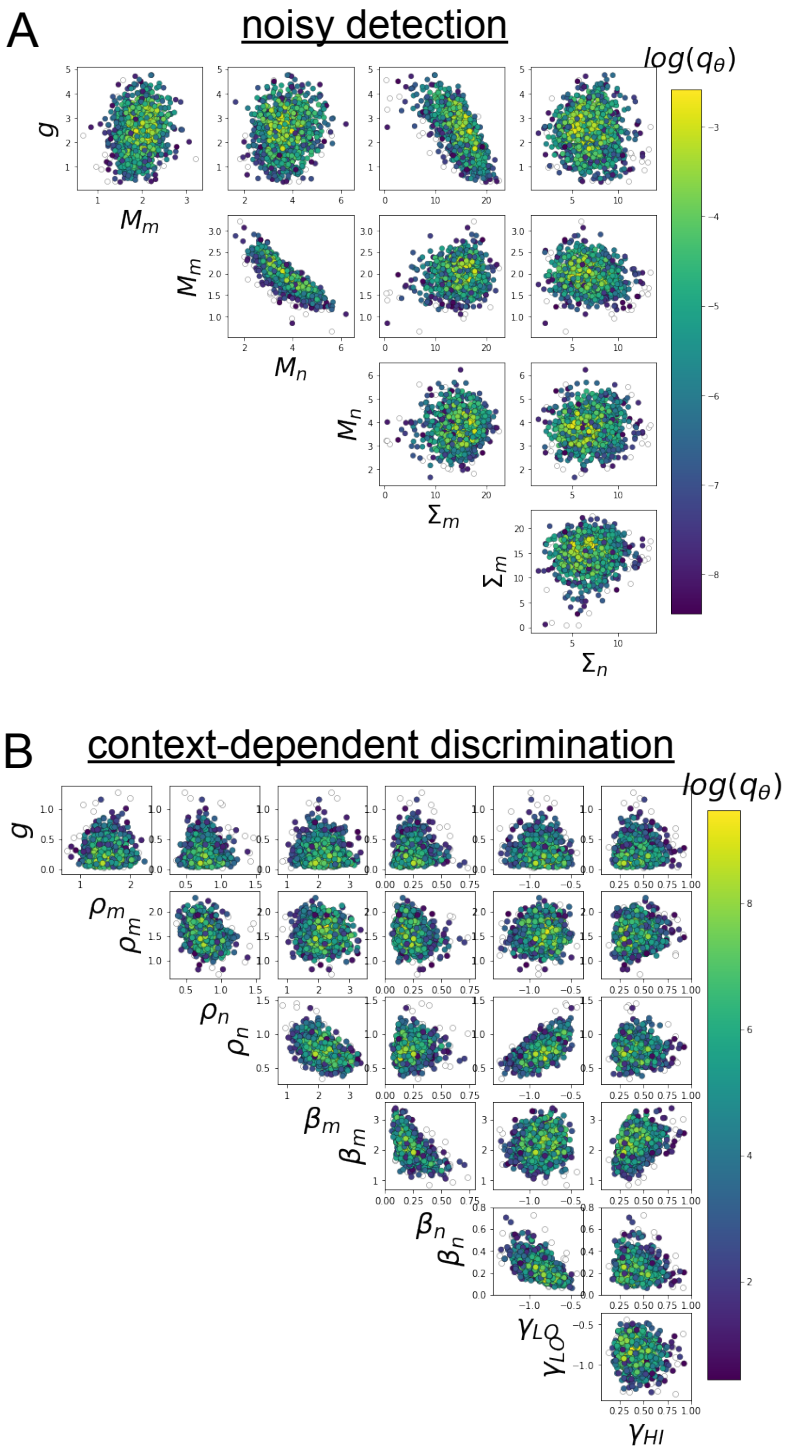

Fig. S1: A. EPI for rank-1 networks doing discrimination. B. EPI for rank-2 networks doing context-dependent discrimination.

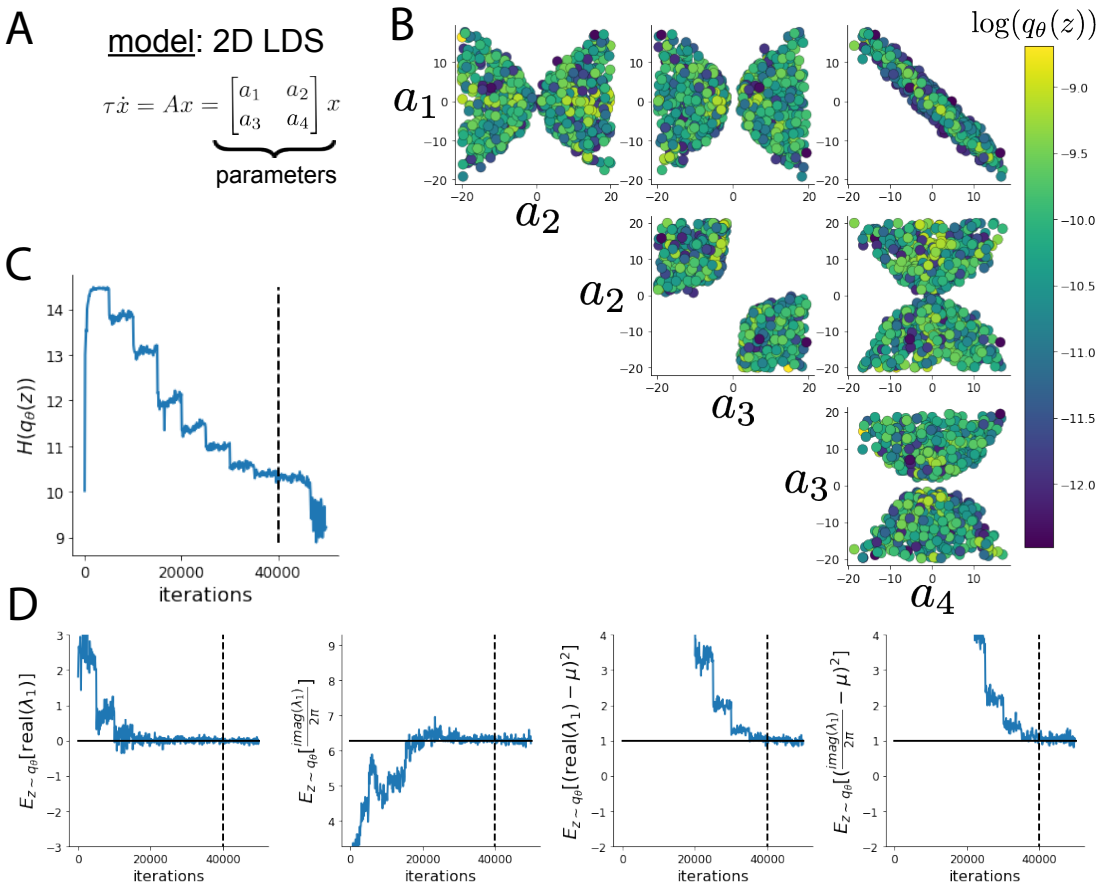

Fig. S2: A. Two-dimensional linear dynamical system model, where real entries of the dynamics matrix A are the parameters. B. The DSN distribution for a 2D LDS with $\tau = 1$ that produces an average of 1Hz oscillations with some small amount of variance. C. Entropy throughout the optimization. At the beginning of each augmented lagrangian epoch (5,000 iterations), the entropy dips due to the shifted optimization manifold where emergent property constraint satisfaction is increasingly weighted. D. Emergent property moments throughout optimization. At the beginning of each augmented lagrangian epoch, the emergent property moments move closer to their constraints.

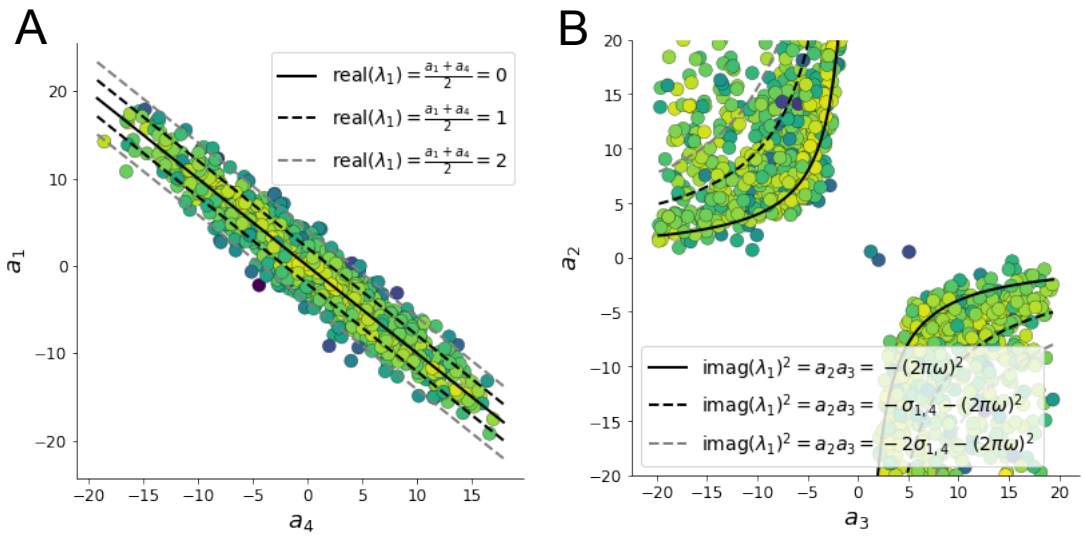

Fig. S3: A. Probability contours in the $a_1 - a_4$ plane can be derived from the relationship to emergent property statistic of growth/decay factor. B. Probability contours in the $a_2 - a_3$ plane can be derived from relationship to the emergent property statistic of oscillation frequency.