
Capturing the diversity of biological tuning curves using generative adversarial networks

Takafumi Arakaki*
Institute of Neuroscience
University of Oregon
Eugene, OR 97403
tarakaki@uoregon.edu

G. Barello*
Institute of Neuroscience
University of Oregon
Eugene, OR 97403
gbarello@uoregon.edu

Yashar Ahmadian
Institute of Neuroscience
University of Oregon
Eugene, OR 97403
yashar@uoregon.edu

* Equal contributions.

Abstract

Tuning curves characterizing the response selectivities of biological neurons often exhibit large degrees of irregularity and diversity across neurons. Theoretical network models that feature heterogeneous cell populations or random connectivity also give rise to diverse tuning curves. However, a general framework for fitting such models to experimentally measured tuning curves is lacking. We address this problem by proposing to view mechanistic network models as generative models whose parameters can be optimized to fit the distribution of experimentally measured tuning curves. A major obstacle for fitting such models is that their likelihood function is not explicitly available or is highly intractable to compute. Recent advances in machine learning provide ways for fitting generative models without the need to evaluate the likelihood and its gradient. Generative Adversarial Networks (GAN) provide one such framework which has been successful in traditional machine learning tasks. We apply this approach in two separate experiments, showing how GANs can be used to fit commonly used mechanistic models in theoretical neuroscience to datasets of measured tuning curves. This fitting procedure avoids the computationally expensive step of inferring latent variables, *e.g.*, the biophysical parameters of individual cells or the particular realization of the full synaptic connectivity matrix, and directly learns model parameters which characterize the *statistics* of connectivity or of single-cell properties. Another strength of this approach is that it fits the entire, joint distribution of experimental tuning curves, instead of matching a few summary statistics picked *a priori* by the user. More generally, this framework opens the door to fitting theoretically motivated dynamical network models directly to simultaneously or non-simultaneously recorded neural responses.

1 Introduction

Neural responses in many brain areas are tuned to external parameters such as stimulus- or movement-related features. Tuning curves characterize the dependence of neural responses on such

parameters, and are a key descriptive tool in neuroscience. Experimentally measured tuning curves often exhibit a rich and bewildering diversity across neurons in the same brain area, which complicates simple understanding [1]. This complexity has given rise to a tendency towards biased selections of minorities of cells which exhibit pure selectivities, and have orderly and easily interpretable tuning curves. As a result the biological richness and diversity of tuning curves in the full neural population is often artificially reduced or ignored. On the theoretical side too, many network models feature homogeneous populations of cells with the same cellular parameters and with regular synaptic connectivity patterns. Neural tuning curves in such models will naturally be regular and have identical shapes.

New theoretical advances, however, have highlighted the computational importance of diverse tuning and mixed selectivity, as observed in biological systems [2, 3]. Furthermore, diversity and heterogeneity can be produced in mechanistic network models which either include cell populations with heterogeneous single-cell parameters (see *e.g.*, [4]), or connectivity that is partly random and irregular despite having regular *statistical* structure (see, *e.g.*, [5, 6, 3, 7, 8, 9]). However, a general effective methodology for fitting such models to experimental data, such as heterogeneous samples of biological tuning curves is lacking.

A related central problem in neural data analysis is that of inferring functional and synaptic connectivity from neural responses and correlations. A rich literature has addressed this problem [10, 11, 12, 13, 14, 15]. However, we see two shortcomings in previous approaches. First, most methods are based on forward models that are primarily inspired by their ease of optimization and fitting to data, rather than by theoretical or biological principles. Second, in the vast majority of approaches, the outcome is the estimate of the particular connectivity matrix between the particular subset of neurons sampled and simultaneously recorded in particular animals [10, 11, 12, 13, 14, 15]. Post-hoc analyses may then be applied to characterize various statistical properties and regularities of connectivity [11, 15]. The latter are what is truly of interest, as they generalize beyond the particular recorded sample. Examples of such statistical regularities are the dependence of connection probability between neurons on their physical distance [16] or preferred stimulus features [17]. Another example is the degree to which neuron pairs tend to be connected bidirectionally beyond chance [18]. A methodology for inferring or constraining such statistics directly from simultaneously or non-simultaneously recorded neural responses is lacking.

Here we propose a methodology that is able to fit theoretically motivated network models to (simultaneously or non-simultaneously) recorded neural responses, by directly optimizing parameters characterizing the *statistics* of connectivity or of single-cell properties. Conceptually, we propose to view network models with heterogeneity and random connectivity as generative models for the observed neural data, *e.g.*, a model that generates diverse tuning curves and hence implicitly models their (high-dimensional) statistical distribution. The generative model is determined by the set of network parameters which specify the distribution of mechanistic circuit variables like synaptic connectivity matrix or single-cell biophysical properties. In this picture, the particular realization of the connectivity matrix or of biological properties of particular neurons are viewed as latent variables. Traditional, likelihood-based approaches such as expectation maximization or related approaches need to optimize or marginalize out (*e.g.*, using variational or Monte Carlo sampling methods) such latent variables conditioned on the particular observed data sample. Such high-dimensional optimizations or integrations are computationally very expensive and often intractable.

Alternatively, one could fit theoretical circuit models by approaches similar to moment-matching, or its Bayesian counterpart, Approximate Bayesian Computation [19, 20]. In such approaches, one *a priori* comes up with a few summary statistics, perhaps motivated on theoretical grounds, which characterize the data objects (*e.g.*, tuning curves). Then one tunes (or in the Bayesian case, samples) the model parameters (but not latent variables) so that the few selected summary statistics are approximately matched between generated tuning curve samples and experimental ones [9]. This approach will, however, generally be biased by the *a priori* choice of statistics to be fit. Furthermore, when model parameters are more than a few (such that grid search becomes impractical), and the chosen statistics are complex and possibly non-differentiable, the fitting procedure will become impractical. By contrast, in the approach we propose here, the fitting of the high-dimensional data distribution can in principle be done in a much more unbiased manner,¹ and without the need to *a priori* choose a few summary statistics.

¹As long as the discriminator network (see Sec. 2.1) used in the adversarial training is sufficiently complex.

A suite of new methods have recently been developed in machine learning for fitting *implicit* generative models [21, 22, 23], *i.e.*, generative models for which a closed or tractable expression for the likelihood or its gradient is not available. Here, we will use one of these methods, namely Generative Adversarial Networks (GANs) [22, 24, 25]. However, other approaches are worth noting, including variational autoencoders [21, 26], and hierarchical and deep implicit models [23]. These other methods are potentially as powerful in their application to neuroscience. Additionally, there is recent progress in unifying these approaches [27, 28, 29] which may clarify future applications of these techniques. This report is organized as follows. In section 2, we introduce the general GAN framework, and the two circuit models adopted from the theoretical neuroscience literature that we use in our experiments. In section 3, we present the results of our GAN-based training of these two models. Finally, in sections 4 and 5 we discuss areas for improvement of our proposed methodology, and other potential future applications of it.

2 Methods

2.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a recently developed approach to fitting generative models to unlabeled data, such as natural images [22, 30, 31]. The GAN approach is powerful because it is applicable to models for which evaluating the likelihood or its gradient are intractable; all that is required is a generative process that, given a random seed, generates a sample data object. In particular, the GAN approach avoids the computationally costly step of inference that is required in, *e.g.*, the expectation maximization algorithm.

In a GAN there are two networks. The first network, the “generator”, implements the generative model, and produces data samples. The generator is defined by a function \mathcal{G} which generates a data sample $\mathcal{G}(\mathbf{z})$ given some sample \mathbf{z} from a fixed, standard distribution. The other network is the “discriminator”, given by a function \mathcal{D} acting on the data space, which is trained to distinguish between the real data and the samples generated by \mathcal{G} . The generator is trained to fool the discriminator. When both \mathcal{D} and \mathcal{G} are differentiable functions, training can be done in a gradient-based way. If the discriminator network is sufficiently powerful, the only way for the generator to fool it is to effectively generate samples that effectively have the same distribution as that of the real data.

GANs are often difficult to optimize and many tricks and techniques have been developed to make learning more robust. In this work we employ the Wasserstein GAN (WGAN) approach which has shown promise in overcoming some of the short-comings of generators learned in the traditional GAN approach [24, 25]. We note, however, that traditional GAN’s could also be used for the types of application we have in mind.² The insight of the WGAN approach is to add a term to the discriminator’s loss which penalizes too small or too large gradients with respect to the inputs, the effect of which is to avoid the vanishing and exploding gradient problem which regular GANs suffer from.³ This approach is mathematically motivated by minimizing the Wasserstein distance (also known as the earth mover’s distance) between the data distribution and the generative model’s distribution [24]. The Wasserstein distance provides a measure of similarity between distributions which exploits metric structure in the data space (by contrast, maximum likelihood fitting minimizes the Kullback-Leibler divergence between the data and model distributions, which is a purely information theoretic measure, blind to metric structure in data space).

Denoting the data samples by \mathbf{x} , with ϵ a uniformly distributed random variable between 0 and 1, the modified WGAN losses [25] for \mathcal{D} and \mathcal{G} are given by

$$\text{Loss}_{\mathcal{D}} = \langle \mathcal{D}(\mathcal{G}(\mathbf{z})) \rangle_{\mathbf{z}} - \langle \mathcal{D}(\mathbf{x}) \rangle_{\mathbf{x}} + \lambda \langle (\|\nabla \mathcal{D}(\epsilon \mathbf{x} + (1 - \epsilon)\mathcal{G}(\mathbf{z}))\|_2 - 1)^2 \rangle_{\mathbf{z}, \mathbf{x}, \epsilon} \quad (1)$$

$$\text{Loss}_{\mathcal{G}} = -\langle \mathcal{D}(\mathcal{G}(\mathbf{z})) \rangle_{\mathbf{z}}, \quad (2)$$

²In its original formulation, unlike in WGAN, the GAN training is equivalent to a zero-sum game with a loss-function equal to the cross-entropy loss (a measure of classification error) for the discriminator. The discriminator parameters are changed to minimize this loss while the generator parameters are changed to maximize it, leading to a minimax or zero-sum game [30].

³In the WGAN approach, the \mathcal{D} network is not necessarily a classifier, but we will nevertheless keep referring to it as the “discriminator”.

respectively. The average over \mathbf{x} denotes the empirical average over the training examples, while \mathbf{z} and ϵ are averaged over their fixed, standard distributions. When necessary to make the dependence of discriminator and generator functions on their learned parameters \mathbf{w} and $\boldsymbol{\theta}$ explicit, we denote them by $\mathcal{D}_{\mathbf{w}}$ and $\mathcal{G}_{\boldsymbol{\theta}}$, respectively. A stochastic gradient descent algorithm on these loss functions is shown in panel 1.

Input: data distribution \mathbb{P}_r , the gradient penalty coefficient λ , the number of critic iterations per generator iteration n_{critic} , the batch size m , Adam hyperparameters $\alpha_{\text{critic}}, \alpha_{\text{generator}}, \beta_1, \beta_2$ and the initial critic \mathbf{w}_0 and generator $\boldsymbol{\theta}_0$ parameters.

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0$;

$\mathbf{w} \leftarrow \mathbf{w}_0$;

while $\boldsymbol{\theta}$ has not converged **do**

repeat

for $i = 1, \dots, m$ **do**

 Sample real data $\mathbf{x} \sim \mathbb{P}_r$, latent variable $\mathbf{z} \sim p(\mathbf{z})$, a random number $\epsilon \sim U[0, 1]$;

$\tilde{\mathbf{x}} \leftarrow \mathcal{G}_{\boldsymbol{\theta}}(\mathbf{z})$;

$\tilde{\mathbf{x}} \leftarrow \epsilon \mathbf{x} + (1 - \epsilon) \tilde{\mathbf{x}}$;

$\text{Loss}_{\mathcal{D}}^{(i)} \leftarrow \mathcal{D}_{\mathbf{w}}(\tilde{\mathbf{x}}) - \mathcal{D}_{\mathbf{w}}(\mathbf{x}) + \lambda(\|\nabla_{\tilde{\mathbf{x}}} \mathcal{D}_{\mathbf{w}}(\tilde{\mathbf{x}})\|_2 - 1)^2$;

end

$\mathbf{w} \leftarrow \text{Adam}(\nabla_{\mathbf{w}} \frac{1}{m} \sum_{i=1}^m \text{Loss}_{\mathcal{D}}^{(i)}, \mathbf{w}, \alpha_{\text{critic}}, \beta_1, \beta_2)$;

until critic is updated n_{critic} times;

 Sample a batch of latent variables $(\mathbf{z}^{(i)})_{i=1}^m \sim p(\mathbf{z})$;

$\boldsymbol{\theta} \leftarrow \text{Adam}(-\nabla_{\boldsymbol{\theta}} (\frac{1}{m} \sum_{i=1}^m \mathcal{D}_{\mathbf{w}}(\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{z}^{(i)})) + \text{Penalty}_{\mathcal{G}}), \boldsymbol{\theta}, \alpha_{\text{generator}}, \beta_1, \beta_2)$;

end

Algorithm 1: Improved WGAN algorithm [25]. Most of the hyper-parameters are taken from [25]: $n_{\text{critic}} = 5, \beta_1 = 0.5, \beta_2 = 0.9$. We use $\alpha_{\text{critic}} = .001, \alpha_{\text{generator}} = .001, m = 30, \text{Penalty}_{\mathcal{G}} = 0$ for the feedforward model and $\alpha_{\text{critic}} = 0.02, \alpha_{\text{generator}} = 0.01, m = 15, \text{Penalty}_{\mathcal{G}} = \text{Eq. 15}$ for the SSN.

2.2 Generative model examples from theoretical neuroscience

In this article, we consider two network models, developed in theoretical neuroscience, as basic examples of generator networks, to illustrate the general GAN-based approach we are proposing for fitting heterogeneous network models to experimental data. In both examples the experimental data take the form of a collection of tuning curves for different neurons from some brain area (tuning curve is the function describing the dependence of a neuron’s trial-averaged response on one or several stimulus or behavioral parameters). But our proposed methodology, based on viewing such models as implicit generative models, is more general and can (with small modifications) also be applied to fit network models to data beyond tuning curves (time-series data from simultaneously recorded neurons in different experimental conditions constitute another important example). Both example networks considered here are scientifically grounded, previously published models of biological neural networks and neural response tunings. In particular, their parameters in principle correspond to physiological and anatomical parameters of biological neural circuits. This is in contrast to the usual case in the GAN literature where the structure of the generator (*e.g.*, a deconvolutional deep feedforward network, generating natural images) has no direct mechanistic interpretation.

In both examples, the random variability occurs in the network structure (connection strengths). By contrast, the external inputs to the network’s neurons, which represent experimental stimuli or conditions used in training data, range over a fixed set of possibilities in all random realizations of the network. To avoid confusion, we will refer to the random variables (denoted by \mathbf{z} as in Sec. 2.1) determining the network structure as “random inputs”, and will refer to the neurons’ external inputs as “stimuli”. We index stimuli ranging over the set of S possibilities used in training data by $s \in \{1, \dots, S\}$. Note that a trained network model can nevertheless be applied to stimuli other than those it is trained on. Let $G(s, \mathbf{z})$ denote the response of a selected neuron (chosen in advance) in a specific realization of the network (determined by \mathbf{z}) to stimulus s . Throughout we will use $\mathcal{G}(\mathbf{z})$ to denote the S -dimensional vector $(G(s, \mathbf{z}))_{s=1}^S$. Thus $\mathcal{G}(\mathbf{z})$ is nothing but the discretized tuning

Parameter	Description
σ_l	Lower bound of receptive field size range
$\delta\sigma$	Width of receptive field size range
J	Scale of connection strengths
ϕ_l	Lower bound of threshold range
$\delta\phi$	Width of threshold range

Table 1: The parameters of the feedforward model, which are fit to tuning curve data.

curve of the selected neuron, *i.e.*, the set of its responses to each of several stimuli differing in one or more stimulus parameters.

2.2.1 Feedforward Model

As our first example we take a feedforward model of primary motor cortex (M1) tuning curves proposed by [9].⁴ This model was proposed by [9] as a plausible mechanism for generating the significant complex nonlinear deviations of observed M1 tuning curves from the classical linear model [32]. The authors fit that model to a collection of experimentally measured M1 tuning curves by matching two specific statistics which were of interest for theoretical reasons. These tuning curves describe the responses of a given neuron in 54 experimental conditions corresponding to the monkey holding its hand, either in supination or pronation, in one location out of a $3 \times 3 \times 3$ cubic grid of spatial locations. In our (numerical) experiment we only used the pronation data, and we will refer to the 27 conditions as “stimuli”, and denote them by $s \in \{1, \dots, 27\}$, while denoting the 3D hand location in condition s by \mathbf{x}_s .

The model is a two-layer feedforward network, with an input layer, putatively corresponding to the parietal reach area or to premotor cortex, and an output layer modeling M1 (see Fig. 1). The activations of the input layer neurons are determined by Gaussian receptive fields on three dimensional space. Across the input layer, the Gaussian receptive fields are centered on a regular cubic grid that extends three times beyond the $3 \times 3 \times 3$ stimulus grid along each dimension. Whereas Lalazar et al. used a network with 100 input-layer receptive fields along each axis, we reduced this number to 40 to allow faster computations; however, changing this resolution beyond 40 only weakly affects the results. The receptive field widths were random across the input layer, and were sampled iid from the uniform distribution on the range $[\sigma_l, \sigma_l + \delta\sigma]$. The feedforward connections from the input to output layer are sparse, with a sparsity of 1%. In our implementation of this model, the strength of the nonzero connections were sampled iid from the uniform distribution on the range $[0, J]$. The response of an output layer neuron is given by a rectified linear response function with a threshold. In the original model of [9] the thresholds were chosen separately for each model neuron such that the coding level of its response (defined as the fraction of stimuli to which a neuron responds with a rate significantly different from zero) matched that of a randomly selected data neuron. In order to have all structural variability in the model in a differentiable form, we instead sampled the threshold uniformly and independently from the range $[\phi_l, \phi_l + \delta\phi]$. We found that with this addition the model was in fact able to match the distribution of neural coding levels in the dataset of [9].

The collection of random inputs, \mathbf{z} , to the generator function for this model determine all input-layer receptive field sizes, individual feedforward connection strengths, and all output-layer neural thresholds for a particular network realization. The output of the network are responses for a single output layer neuron to the 27 stimuli (since the network is feedforward, with independent connections, one could equivalently sample several neurons from the output layer in each realization). The response $G(s)$ of this network given a stimulus s is calculated by

$$G(s; \mathbf{z}) = \left[\sum_{i=1}^{40^3} \mathcal{J}_i h_i(s) - \phi \right]_+ \quad (3)$$

⁴Reference [9] introduced their model in two versions, a basic one, and an “enhanced” version. We have used their enhanced version, with small modifications which allow our approach to be used, as our first example.

where

$$h_i(s) = \frac{1}{Z} \exp\left(-\frac{1}{2\sigma_i^2} \|\mathbf{x}_s - \bar{\mathbf{x}}_i\|^2\right) \quad (4)$$

$$\mathcal{J}_i = J z_i^J M_i \quad (5)$$

$$\phi = \phi_l + z^\phi \delta\phi, \quad (6)$$

$$\sigma_i = \sigma_l + z_i^\sigma \delta\sigma, \quad (7)$$

$$z^\phi, z_i^J, z_i^\sigma \stackrel{iid}{\sim} U[0, 1] \quad (8)$$

where Z is a normalizing factor such that $\sum_i h_i(s) = 1$, $\mathbf{M} = (M_i)_{i=1}^{40^3}$ is the sparse binary mask with sparsity 1%, and $[u]_+ = \max(0, u)$ denotes rectification. The random inputs \mathbf{z} to this model are $\mathbf{z} = (z^\phi, (z_i^\sigma)_{i=1}^{40^3}, (z_i^J)_{i=1}^{40^3}, (M_i)_{i=1}^{40^3})$, and there are five trainable parameters $\boldsymbol{\theta} = (\sigma_l, \delta\sigma, J, \phi_l, \delta\phi)$ (listed in table 1). Crucially, since the network's output, $G(s; \mathbf{z})$, is differentiable with respect to each of the trainable parameters, we can use the output gradient with respect to parameters to optimize the latter using any variant of the stochastic gradient descent algorithm.

2.2.2 Recurrent Model

The second model we consider is a recurrent model of cortical circuitry, the Stabilized Supralinear Network (SSN), which has found broad success in mechanistically explaining the contextual and attentional modulations of neural responses, across multiple areas of sensory and association cortex [33, 7, 34]. The SSN is a model of cortical circuitry (within a given cortical area), in which neurons have a supralinear rectified power-law input/output function, $f(u) = k[u]_+^n$ (where $[u]_+ = \max(0, u)$, $k > 0$, and $n > 1$). We consider a topographically organized version of this model, with a one-dimensional topographic map which could correspond, *e.g.*, to the tonotopic map of the primary auditory cortex (A1), or a one-dimensional reduction of the retinotopic map in primary visual cortex (V1). The network is composed of excitatory (E) and inhibitory (I) neurons, with one neuron of each type at each topographic spatial location. For N topographic locations, the network thus contains $2N$ neurons. For a network with recurrent connectivity matrix \mathbf{W} , the vector of firing rates \mathbf{r} is governed by the differential equation

$$\mathbf{T} \frac{d\mathbf{r}}{dt} = -\mathbf{r} + f(\mathbf{W}\mathbf{r} + \mathbf{I}(s)), \quad f(\mathbf{u}) \equiv (f(u_i))_{i=1}^{2N} = (k[u_i]_+^n)_{i=1}^{2N} \quad (9)$$

where the diagonal matrix $\mathbf{T} = \text{Diag}((\tau_i)_{i=1}^{2N})$ denotes the neural relaxation time constants, and $\mathbf{I}(s)$ denotes the external or stimulus input in condition $s \in \{1, \dots, S\}$. Below, for an arbitrary neuron i , we denote its type by $\alpha(i) \in \{E, I\}$ and its topographic location by x_i . In our SSN example, we let x_i 's range from -4 to 4 on a regular grid. In the SSN example we consider here, we chose to have all random structural variability occur in the connectivity matrix \mathbf{W} . The standard random variables of the GAN formalism, \mathbf{z} , are thus those which determine a concrete realization of \mathbf{W} , and the learned model parameters $\boldsymbol{\theta}$ are those governing the statistics of connectivity (see Eq. (11) below).

For large N and with suitable choices of parameters, the SSN will generically reach a stable fixed point [35]. This fixed point, which we denote by $\hat{\mathbf{r}}$, represents the steady-state response of cortical neurons to stimuli. The steady state response for an *a priori* selected neuron, *i.e.*, a certain component of $\hat{\mathbf{r}}$, with index i_* , is the output of our SSN, viewed as a generative model in the GAN setting. We set $G(s, \mathbf{z}) = \hat{r}_{i_*}$ when the SSN stimulus is $I(s)$ so that $\mathcal{G}(\mathbf{z}) = (G(s, \mathbf{z}))_{s=1}^S$ is the tuning curve of neuron i_* , defined as the set of its steady-state responses in all stimulus conditions.

In our example, we simulated the fitting of an SSN model of V1 to datasets of stimulus size tuning curves of V1 neurons [7]. We let the stimulus input to neuron i be

$$I_i(s) = A \sigma\left(l^{-1}(b_s/2 + x_i)\right) \sigma\left(l^{-1}(b_s/2 - x_i)\right) \quad (10)$$

where $\sigma(u) = (1 + \exp(-u))^{-1}$ is the logistic function, A denotes the stimulus intensity, and $(b_s)_{s=1}^S$ are a set of stimulus sizes. Thus, in condition s , the stimulus targets a central band of width b_s centered on the middle of the topographic grid (see Fig. 1 bottom). The parameter l determines the width of the smoothed edges of this banded stimulus. This stimulus is a simple and reduced model

Parameter	Description
σ_{ab}	connection length scales
J_{ab}	lower bounds of connection strengths
δJ_{ab}	widths of connection strength distribution

Table 2: The parameters of the SSN to be estimated ($a, b \in \{E, I\}$).

of visual input to V1 for gratings of diameter b . The tuning curve $\mathcal{G}(\mathbf{z})$ is then the stimulus size tuning curve of neuron i_* . In experiments, typically the grating stimulus used to measure a neuron’s size tuning curve is centered on the neuron’s receptive field. To model this stimulus centering, we let the selected neuron, the steady-state responses of which form $\mathcal{G}(\mathbf{z})$, be the neuron in the center of the topographic grid, with $x_{i_*} = 0$.⁵

In many cortical areas the statistics of connectivity, such as connection probability and average strength, depend on several factors, including the types of the pre- and post-synaptic neurons, the physical distance between them, or the difference between their preferred stimulus features [16, 17]. A common statistical ensemble adopted in theoretical neuroscience posits a binary distribution for the synaptic weight W_{ij} : the random variable W_{ij} is either zero or, with probability p_{ij} , equal to a value $J_{\alpha(i)\alpha(j)}$. The fixed non-random J_{ab} ’s set the strength and signs of existing connections between neurons of different types. The connection probability p_{ij} can depend on the distance between neurons i and j as well as on their types. In this ensemble, the W_{ij} are, however, not differentiable functions of parameters. We instead chose a statistical ensemble for \mathbf{W} that made the random variables, W_{ij} , differentiable functions of learned model parameters. For simplicity, we assumed that W_{ij} ’s are independent and of uniform distributions with mean and range that depend on the pre- and post-synaptic types and fall off with the distance between the pre- and post-synaptic neurons over characteristic length-scales. More precisely we chose this fall-off to be gaussian, and set

$$W_{ij} = (J_{ab} + z_{ij} \delta J_{ab}) e^{-\frac{(x_i - x_j)^2}{2\sigma_{ab}^2}} \quad a = \alpha(i), b = \alpha(j) \quad (11)$$

$$z_{ij} \stackrel{iid}{\sim} U[0, 1]. \quad (12)$$

Thus $\langle W_{ij} \rangle = \bar{J}_{ab} e^{-\frac{(x_i - x_j)^2}{2\sigma_{ab}^2}}$ where $\bar{J}_{ab} = J_{ab} + \delta J_{ab}/2$, while $\text{Var}[W_{ij}] = \frac{1}{12} \delta J_{ab}^2 e^{-\frac{(x_i - x_j)^2}{\sigma_{ab}^2}}$.

Our connectivity ensemble (and hence our SSN generator) is thus characterized by 12 parameters (enumerated in Table 2), namely the elements of three 2×2 matrices: the connection strength parameters J_{ab} , δJ_{ab} , and characteristic length scales σ_{ab} , where $a, b \in \{E, I\}$. All these parameters satisfy sign constraints: $\sigma_{ab} > 0$, $J_{aE}, \delta J_{aE} \geq 0$, and $J_{aI}, \delta J_{aI} \leq 0$. The latter two sets of constraints ensure that any realization of W_{ij} satisfies Dale’s principle [36, 37]. To enforce this constraint, the parameters we fit (those used to compute gradients) are the logarithm of the positive-definite parameters listed here. To be explicit, the set of 12 learned parameters is $\boldsymbol{\theta} = (\log |J_{ab}|, \log \delta |J_{ab}|, \log \sigma_{ab})_{a,b \in \{E, I\}}$. Finally, a given realization of our SSN model is specified (through its connectivity matrix) by the values of the $(2N)^2$ independent, standard uniform random variables $\mathbf{z} = (z_{ij})_{i,j=1}^{2N}$. To make the dependence of the connectivity matrix on \mathbf{z} explicit, we denote it by $\mathbf{W}^{\mathbf{z}}$ below.

Given a dataset of tuning curves, we would like to find a set of parameters, $\boldsymbol{\theta}$, that produces a matching model distribution of tuning curves. In this paper, as a proof of principle, we used simulated data, by using the same SSN model, with parameters $\boldsymbol{\theta}_0$, to generate size tuning curves used as the training set in the GAN training. During training, according to algorithm 1, every time $\mathcal{G}(\mathbf{z})$ was evaluated, we simulated the SSN until convergence to a fixed point. The gradient $\nabla_{\boldsymbol{\theta}} \mathcal{G}_{\boldsymbol{\theta}}(\mathbf{z}) = (\nabla_{\boldsymbol{\theta}} G_{\boldsymbol{\theta}}(s, \mathbf{z}))_{s=1}^S$, is obtained from the i_* components of the fixed point gradients $\partial \hat{\mathbf{r}} / \partial \boldsymbol{\theta}$. The expression for the latter gradient can be derived from the fixed point equation,

⁵Note that in general, the tuning curves in random SSN’s also show variability across neurons as well as across different realizations of the network for a fixed output neuron. Furthermore, when N is large, one could have approximate ergodicity, in that the variability among the tuning curves of neurons with topographic location $x_i \approx 0$ approximates variability across different \mathbf{z} for neuron i_* . This ergodicity may in principle be exploited for more efficient training and testing of the model.

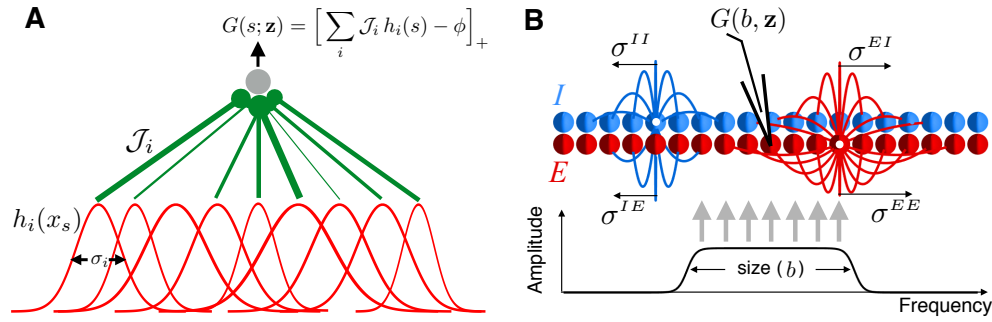


Figure 1: Structure of the feedforward (A) and recurrent SSN (B) models used in our experiments as example generator networks.

$\hat{\mathbf{r}} = f(\mathbf{W}^z \hat{\mathbf{r}} + \mathbf{I}(s))$, using the implicit derivative formula, to yield

$$\frac{d\hat{\mathbf{r}}}{d\theta} = (\mathbf{1} - \Phi \mathbf{W}^z)^{-1} \Phi \frac{\partial \mathbf{W}^z}{\partial \theta} \hat{\mathbf{r}}, \quad (13)$$

where Φ is the diagonal matrix of neural gains defined by $\Phi = \text{Diag}[f'(\mathbf{W}^z \hat{\mathbf{r}} + \mathbf{I}(s))]$.

Enforcing stability

During GAN training for fitting network parameters, the SSN may be pushed to parameter regions in which the network is unstable for some realizations of the randomness inputs \mathbf{z} . Since our framework relies on convergence to a fixed point, and diverging rates are not biological anyway, we avoid the diverging solution by “clipping” the input-output nonlinearity $f(\cdot)$ and redefining it to be

$$f(u) = \begin{cases} k(u)_+^n & \text{if } u \leq V_0 \\ r_0 \left(1 + \frac{r_1 - r_0}{r_0} \tanh \left(n \frac{r_0}{r_1 - r_0} \frac{u - V_0}{V_0} \right) \right) & \text{if } u > V_0 \end{cases} \quad (14)$$

where $r_0 = 200$ Hz, $V_0 = (r_0/k)^{\frac{1}{n}}$, and $r_1 = 1000$ Hz (this makes the nonlinearity concave above the 200 Hz output level, and saturated at 1000 Hz). Note that as long as the generator (SSN) parameters obtained by our fit have fixed points in which all rates are less than r_0 , the model is indistinguishable from that in which the input-output nonlinearity were not modified. The modification Eq. (14) can thus allow the SSN model to venture through unstable and biologically implausible parameter regions during training and yet end up in regions with biologically plausible rates.

Even with the modified input-output relationship Eq. (14), however, the SSN may have a limit cycle or a chaotic attractor. We therefore rejected the parameter points in which the SSN did not converge in 10 seconds of simulated time. To encourage the avoidance of such problematic parameter regions, as well as parameter regions that produce unrealistically large firing rates, we added the differentiable penalty term

$$\text{Penalty}_G = 5000 \sum_{i=1}^{2N} (r_i - 150)_+ \quad (15)$$

to the loss function for the generator.

2.3 Discriminator networks

The most studied application of GANs is in producing highly structured, high-dimensional output such as images, videos, and audio. In these applications it is beneficial to use highly complex discriminators such as deep convolutional networks. It has also been noted that the discriminator network should be sufficiently powerful so that it is capable of fully representing the data and model distributions [30, 38, 39]. In our application, the outputs of the generator are comparatively lower

dimensional objects, with less complex distributions. Correspondingly we used relatively small discriminator networks in order to speed up training time.⁶

For the discriminator network, \mathcal{D} , we used dense feedforward neural networks with two hidden layers of 128 rectified linear units each, and an output layer with a single linear readout unit. The discriminator networks used in the M1 model and SSN model fits were the same, except for the dimensionality of their input layers. Feedforward weights were initialized using the uniform distribution between ± 0.0096 and biases were initialized using a Gaussian distribution with zero mean and standard deviation 0.01. We do not use any kind of normalization or parameter regularization other than the WGAN penalty term in equation 1. We also did not do a substantial search of hyperparameters for the discriminator networks and found that our default options nevertheless performed quite well. This suggests that in our application, the WGAN is relatively insensitive to the detailed structure of the discriminator as long as it is sufficiently complex.

3 Experiments

3.1 Complex M1 Tuning Curves

We fit the feedforward network model introduced in Sec. 2.2.1 using the GAN framework to trial-averaged firing rate data from monkey M1 collected by [9] and available online. The data includes neural responses for trials in which a monkey held its hand, in pronation or supination, in multiple target locations, as described in Sec. 2.2.1. We only used the pronation conditions and pruned the data to include only neurons ($n = 293$) which were measured in all 27 target locations in the pronation condition. We further simplified the dataset by removing spatial scale information in the positions of target locations, which varied slightly between experimental sessions.

Reference [9] fit their model by matching two specific scalar statistics (which were of interest for theoretical reasons) characterizing tuning curves. We will instead adopt a more agnostic approach and let the WGAN fit the full multi-dimensional tuning curve distribution. Moreover, in [9], the model and data tuning curves were individually normalized so that their shape, but not their overall rate scale, was fit to data. As described in Sec. 2.2.1, we instead included a scaling parameter, J , for the feedforward connection strengths and found that with this addition the model, without normalization, is actually able to account for the variability in average firing rate across neurons as well.

The results of our fitting procedure are summarized in Fig. 2. We show the initial and final histograms for four statistics characterizing the tuning curves: R^2 of the linear fit to the tuning curve, complexity score, firing rate (in all conditions), and coding level, *i.e.*, the fraction of conditions with rate significantly greater than zero (which we took to mean larger than 5 Hz). Ref. [9] defined the complexity score of a tuning curve to be the standard deviation of the absolute response differences between nearest neighbor locations on the $3 \times 3 \times 3$ lattice of hand positions (the tuning curve was first normalized so that its responses ranged from -1 to 1). The first two properties, R^2 and complexity score, are the two statistics used by [9] to fit their model; they provide two measures for the degree of nonlinearity of the tuning curve.

We measured the mismatch between the model and data distributions using the Kolmogorov-Smirnov (KS) distance. Figure 2A shows the KS distance between the model and data distributions for the chosen four properties throughout model training using stochastic gradient descent, with the dashed line showing $p < 0.05$. The distribution of complexity (Fig. 2C) is very accurately captured. It is also notable that the trained model fits the distribution of firing rates and coding levels (Fig. 2D-E) quite well: in the work of Ref. [9] the scale of rates was not fit at all (tuning curves were individually normalized) and the fit of coding levels was enforced by hand, as described in Sec. 2.2.1 and below.

The relatively poorer performance in capturing R^2 statistics (Fig. 2B), compared to the enhanced model of [9], may be due to the more flexible way by which they fit individual neuron thresholds (which we did not do). The fact that our distribution of neuron coding levels matches the data

⁶However, note that a \mathcal{D} that is too simple can preclude the fitting of important aspects of the distribution. For example a linear \mathcal{D} used in the WGAN approach would result in a fit that matches only the average tuning curve between model and data, and ignore tuning curve variability altogether.

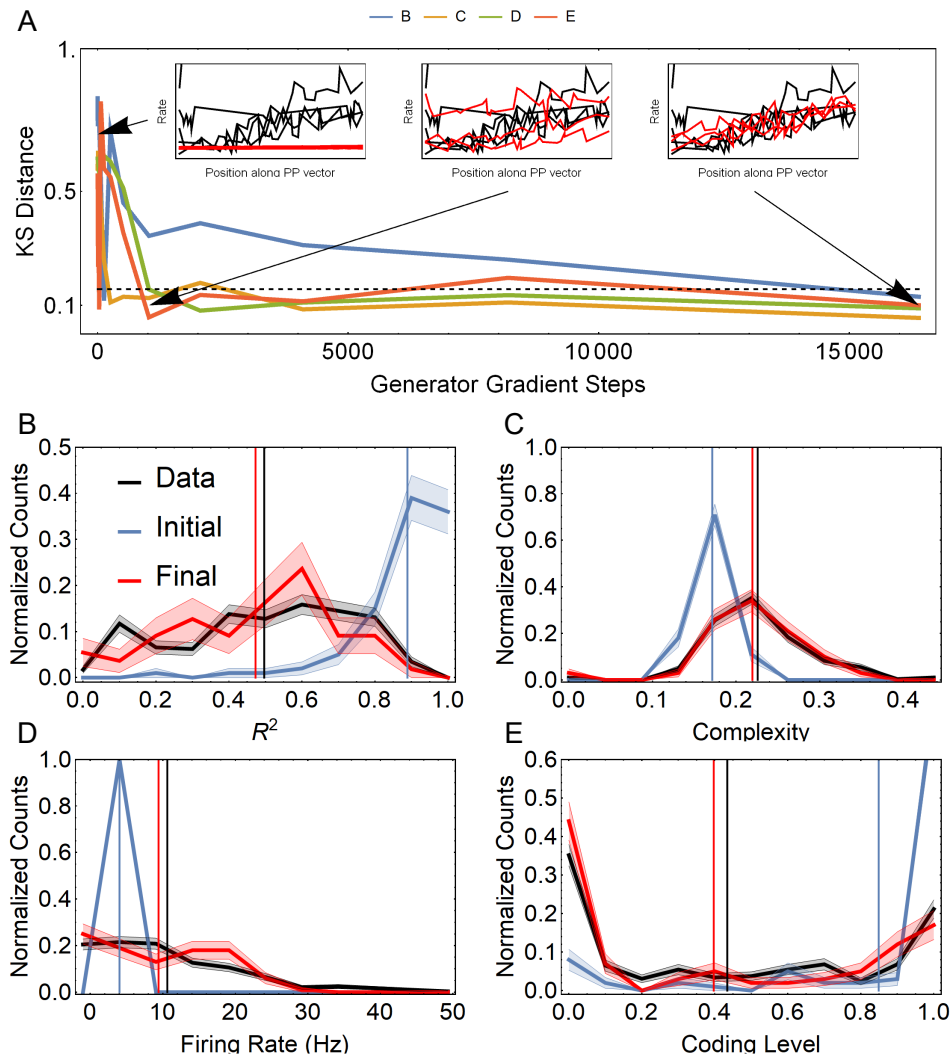


Figure 2: Summary of the feedforward model fit to the M1 tuning curve data of Ref. [9]. **A**: Kolmogorov-Smirnov (KS) distance during training between the data and model distributions, shown in panels B-E, of four summary statistics characterizing the tuning curves. The dashed line shows the KS distance corresponding to significance level $P < 0.05$ between samples of data ($n = 293$) and simulated ($n = 100$) tuning curves. The insets show data and model tuning curves at different points throughout training; the plotted tuning curves are the projections of the 3D tuning curves along the preferred position (PP) vector which is obtained by a linear fit to the 3D tuning curve (see [9]). **B-E**: histograms of four tuning curve statistics (R^2 , complexity score, rates, and coding level) showing the data distribution (black), the initial model distribution (blue) and the final model distribution after fitting (red). Vertical lines show the mean value of each histogram.

quite well suggests that the marginal variability in threshold was nevertheless learned properly by our model. Thresholds in the original work were, however, not independently sampled but were set after other random inputs, \mathbf{z} , so as to match the response coding level of the model neuron to that of a randomly sampled data neuron. Because of this approach, the distribution of thresholds in [9]’s enhanced model was conditioned upon the other random model inputs, \mathbf{z} . If these conditional dependencies are important for fitting the R^2 distribution, the simpler model we implement here is not sufficiently expressive to capture the R^2 distribution in the data.

3.2 Stabilized Supralinear Network

To demonstrate the effectiveness of our proposed approach in its application to complex, dynamical models, we used the GAN framework to train the connectivity parameters of the SSN model described in Sec. 2.2.2. We fit this model to simulated data: 1000 tuning curves generated from an SSN with the connectivity parameters listed in table 3. This set of simulated tuning curves serves as the empirical data distribution for this GAN fit. We fixed all other parameters of both the true and trained SSN and their stimulus inputs as follows: $N = 201$, $k = 0.01$, $n = 2.2$, $\tau_E = 16$ ms, $\tau_I = 2$ ms, $A = 20$, $l = 1/4$, $(b_s)_{s=1}^S = (0.5, 1, 2, 4, 6)$ (for testing purposes, after fitting, we generated tuning curves from the trained SSN using a larger set of stimulus sizes b). To train the learned SSN, we initialized the network parameters $(\sigma_{ab}, J_{ab}, \delta J_{ab})_{a,b \in \{E,I\}}$ to 0.37 times the values of the true parameter. This was done to assure that the SSN began training in a stable regime (see Sec. 4.2 below) and started with parameters that were sufficiently far away from the true parameters so as to make the training non-trivial.

Parameter	EE	EI	IE	II
J_{ab}	0.0957	.0638	.1197	.0479
δJ_{ab}	.7660	.5106	.9575	.3830
σ_{ab}	.6667	.2	1.333	.2

Table 3: The parameter values used for generating the training set tuning curves. The columns correspond to different (a, b) possibilities.

To quantify the goodness-of-fit between the model and data distributions of tuning curves we compare the distribution of four attributes or statistics characterizing the tuning curves: the suppression index, preferred stimulus size, maximum firing rate, and peak width, defined as follows. The suppression index for a neuron (or tuning curve) measures how much the response of that neuron is suppressed below its peak level at large stimulus sizes. It is defined by

$$\text{Suppression Index} = 1 - \frac{\hat{r}(\max(b))}{\max_b(\hat{r}(b))}$$

where $\hat{r}_{i_*}(b)$ is the steady-state response of the neuron to the stimulus with size b . The maximum firing rate is simply $\max_b \hat{r}(b)$, and the preferred size is $\arg \max_b \hat{r}(b)$. Finally, the peak width is a variant of the inverse participation ratio [40] that has the correct units, and limiting behavior, to be interpreted as the width of the tuning curve peak, and is defined by

$$\text{Peak Width} = \left(\sum_b \left(\frac{\hat{r}(b)}{\sum_{b'} \hat{r}(b')} \right)^2 \right)^{-1}. \quad (16)$$

Figures 3 B-E provide comparisons of the initial and final distributions of these tuning curve attributes. As above, we measured the mismatch of these distributions under the trained and true SSN using the KS distance. The KS distance for all distributions decreased during learning, indicating that the fit of the model distributions to the data were in fact improving, and the final fit of all distributions is very good (Figure 3A).⁷

4 Possible issues

4.1 Optimization difficulties

As with any gradient based fit it is possible for a GAN or WGAN to become stuck in sub-optimal local minima for the generator. Furthermore, it is an open question whether GAN training will always converge [30, 41, 42]. As research in GANs and non-convex optimization advances this

⁷Note that the KS-test is very stringent: in the limit where the number of samples grows large even quite similar distributions will be reported as significantly different under this measure, however KS distance alone provides a convenient way to quantify improvement during training, even though the associated p-value may be misleading.

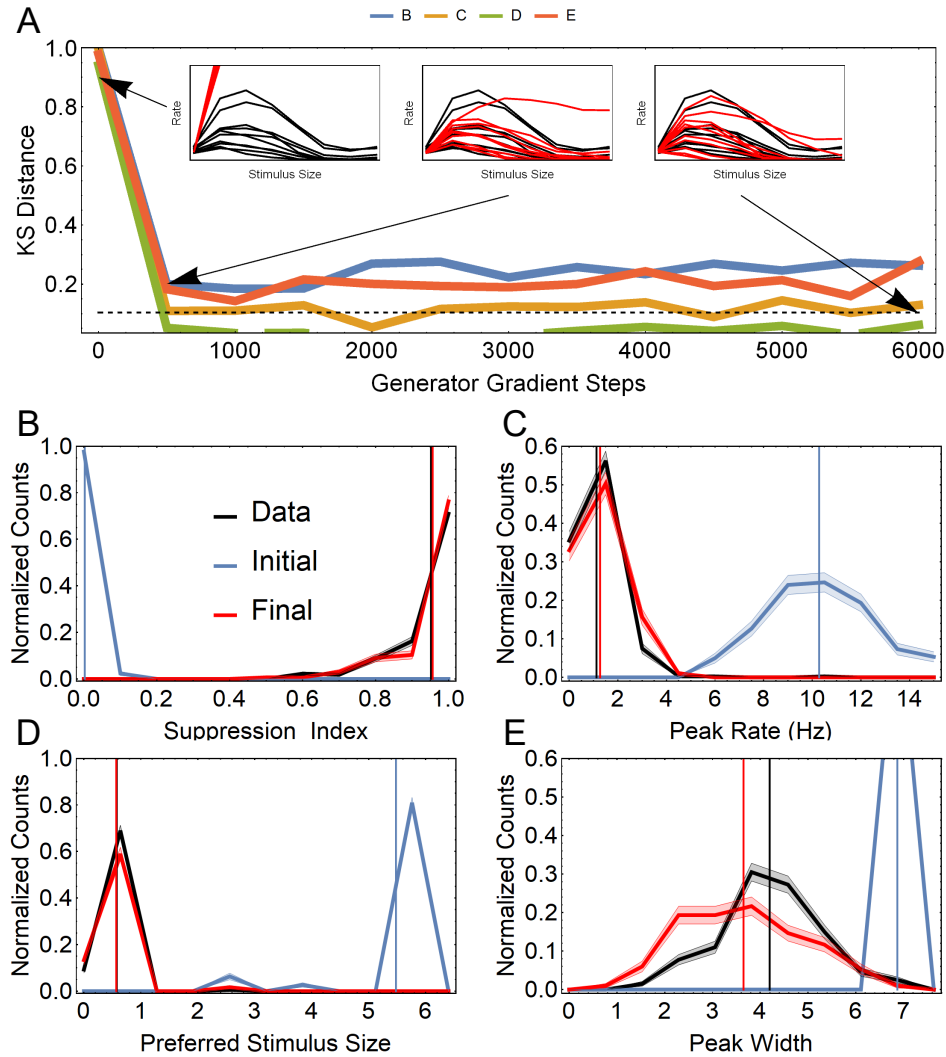


Figure 3: Summary of the recurrent SSN model fit to simulated V1 tuning curve data. **A:** KS distance during training between the data and model distributions, shown in panels B-E, of four summary statistics characterizing the tuning curves. The dashed line shows the KS distance corresponding to significance level $P < 0.05$ between a sample of data ($n = 400$) and simulated ($n = 300$) tuning curves. The insets show data and model tuning curves at different points throughout training. **B-E:** histograms of four tuning curve statistics (suppression index, peak rate, preferred size, and peak width) showing the data distribution (black), the initial model distribution (blue) and the final model distribution after fitting (red). Vertical lines show the mean value of each histogram.

issue will be improved. For now avoiding this pitfall will be a matter of the user judging the quality of fit after the fit has converged. Starting the gradient descent with several different initial conditions for θ (parameters of \mathcal{G}) can also help, with some initializations leading to better final fits.

4.2 Dynamical stability in recurrent generative models

When the generator is a recurrent neural network (RNN) with its output based on a fixed point of the RNN (as was the case in our SSN experiment), an important potential issue is lack of convergence to a stable fixed point for some choices of recurrent network parameters. In the experiment with SSN we initialized the generator network in such a way that it had a stable fixed point for almost all realizations of \mathbf{z} . For the SSN this would generically be the case when recurrent excitation (which has destabilizing effects) is sufficiently weak. Hence initializations with small J_{EE} and δJ_{EE} are good choices. In addition, a relatively large SSN size N improves the stability issue because random relative fluctuations in total incoming synaptic weights are small when the number of post-synaptic neurons is large. Thus, when N is large, for a given choice of network parameters, θ , either the network converges to a stable fixed point for almost all \mathbf{z} , or almost never does. Finally, to avoid entering parameter regions leading to instability during training, we used the tricks introduced at the end of Sec. 2.2.2.

The issue of fixed point stability is reminiscent of the recurrent weight initialization problem in learning of RNNs. In a non-back-propagation based paradigm called the Echo State Networks (ESN) [43], the recurrent weights are configured in such a way that the network dynamics always settle in a fixed point with any stationary input. Jaeger [43] showed that this can be achieved by tuning the spectral radius of the recurrent weight matrix to be smaller than 1 for the $\tanh(x)$ nonlinearity. This helps the network to continuously flush out the memory of old inputs and become ready to store new incoming signals. Although the recurrent weights are not updated in the original ESN formalism, Sutskever et al. [44] applied this idea also to the initialization of recurrent weights learned using the back-propagation through time (BPTT) algorithm. Since an overly small spectral radius is also problematic for the BPTT algorithm due to vanishing gradient [45], they found that choosing the spectral radius to be just above instability (specifically, 1.1) achieves a good balance between stability and speed of training. The spectral radius slightly less than 1 (0.95) also has been used [45]. Unlike those RNNs which learn general dynamics, we do not have the vanishing gradient problem in our SSN example, because in that case the spectral radius is less than 1 with the gradient calculated without BPTT, using Eq. (13). Thus, we are led to believe that starting learning from a very stable dynamical regime, which is analogous to a small spectral radius, is useful in general when the output of the generator is the fixed point of a dynamical system.

4.3 Unidentifiability of parameters

In fitting generative models it is possible for models with widely different parameters to result in nearly identical output distributions. In our case, this corresponds to cases in which networks with widely divergent connectivity or single-cell parameters nevertheless generate very similar tuning curve distributions. In such cases it would not be possible to use our approach to make precise inferences about network parameters (*e.g.*, connectivity parameters) using tuning curve data. In the case of our SSN experiment, for example, we trained the generative model using only tuning curves with respect to stimulus size (measured at a few different sizes). In fact, there is no general reason why a low-dimensional output of the network, such as size tuning curves, would provide sufficient information for constraining all model parameters. Even in such cases, however, a richer dataset of tuning curves can allow for further constraining of model parameters. Fortunately there is nothing in our approach that would prevent one from using joint tuning curves with respect to several stimulus parameters (*e.g.*, joint tuning curves with respect to both stimulus size and stimulus intensity). Generally, it should be the case that the more high-dimensional the tuning curves used in training, *i.e.*, the larger the number of stimulus parameters and conditions they span, the more identifiable the network parameters. Thus with sufficiently rich datasets, our GAN-based method provides a promising way to infer biophysical networks parameters, such as those governing connectivity statistics. Furthermore, our approach can in principle be used to *design* experiments, *i.e.*, optimally choose the stimulus conditions and quantities to be recorded, to maximize the identifiability of the parameters of a given model.

5 Conclusion

Developing theoretically grounded models that can capture the diversity and heterogeneity of neural response selectivities is an important task in computational neuroscience. The ability to fit such models directly to neural data, *e.g.*, to datasets of heterogeneous tuning curves from some brain area, can greatly facilitate this pursuit. However, the statistical fitting of biologically motivated mechanistic models of brain circuitry to neural data, using likelihood-based approaches, is often intractable. A practical solution used in the past has been to instead use models that are designed primarily to have tractable likelihoods and to be easy to fit. The elements and parameters of such models may not have mechanistic, biological interpretations.

Here, we demonstrated that Generative Adversarial Networks enable the fitting of theoretically grounded models to the full distribution of neural data. Conceptually, we view mechanistic network models with randomness in their structure or biophysical parameters as generative models for diverse tuning curves. In Sec. 2 we reviewed the basic GAN setup and the WGAN algorithm, as developed in the machine learning literature. We used this method to successfully fit two representative example models from theoretical neuroscience to tuning curve data, demonstrating that this technique is applicable to a wide range of network models including feedforward and recurrent models of cortical networks. In Sec. 4, we reviewed some of the potential pitfalls of our approach in an attempt to make the path clearer for other applications of this approach.

There are several directions in which the current work can be extended. For example, the methodology can be generalized such that it is able to handle cases of datasets with missing data; it is often the case in experiments that not all neurons are recorded in all stimulus conditions. As another extension, similar methodology can be used to fit mechanistic recurrent network models to simultaneously recorded neural activity, to capture, *e.g.*, the distribution of noise correlations between pairs of neurons in some cortical area. More generally, the potential for the application of newly developed methods for fitting implicit generative models in neuroscience is broad.

Acknowledgments

The authors thank Daniel Lowd and Caleb Holt for inspiring conversations.

References

- [1] Ringach DL, Shapley RM, Hawken MJ (2002) Orientation selectivity in macaque V1: diversity and laminar dependence. *J Neurosci* 22: 5639–5651.
- [2] Rigotti M, Barak O, Warden MR, Wang XJ, Daw ND, et al. (2013) The importance of mixed selectivity in complex cognitive tasks. *Nature* 497: 585–590.
- [3] Barak O, Rigotti M, Fusi S (2013) The sparseness of mixed selectivity neurons controls the generalization-discrimination trade-off. *J Neurosci* 33: 3844–3856.
- [4] Persi E, Hansel D, Nowak L, Barone P, van Vreeswijk C (2011) Power-law input-output transfer functions explain the contrast-response and tuning properties of neurons in visual cortex. *PLoS Comput Biol* 7: 1–21.
- [5] Roxin A, Brunel N, Hansel D, Mongillo G, Van Vreeswijk C (2011) On the Distribution of Firing Rates in Networks of Cortical Neurons. *The Journal of Neuroscience* 31: 16217–16226.
- [6] Litwin-Kumar A, Doiron B (2012) Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nature Neuroscience* 15: 1498–1505.
- [7] Rubin DB, Van Hooser SD, Miller KD (2015) The stabilized supralinear network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron* 85: 402–417.
- [8] Ahmadian Y, Fumarola F, Miller KD (2015) Properties of networks with partially structured and partially random connectivity. *Phys Rev E Stat Nonlin Soft Matter Phys* 91: 012820.
- [9] Lalazar H, Abbott LF, Vaadia E (2016) Tuning curves for arm posture control in motor cortex are consistent with random connectivity. *PLOS Comp Bio* .
- [10] Schneidman E, Berry M, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440: 1007–1012.
- [11] Shlens J, Field GD, Gauthier JL, Grivich MI, Petrusca D, et al. (2006) The Structure of Multi-Neuron Firing Patterns in Primate Retina. *J Neurosci* 26: 8254–8266.

- [12] Pillow J, Shlens J, Paninski L, Sher A, Litke A, et al. (2008) Spatiotemporal correlations and visual signaling in a complete neuronal population. *Nature* 454: 995–999.
- [13] Tang A, Jackson D, Hobbs J, Chen W, Smith JL, et al. (2008) A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. *Journal of Neuroscience* 28: 505–518.
- [14] Shlens J, Field GD, Gauthier JL, Greschner M, Sher A, et al. (2009) The structure of large-scale synchronized firing in primate retina. *Journal of Neuroscience* 29: 5022–31.
- [15] Yatsenko D, Josic K, Ecker AS, Froudarakis E, Cotton RJ, et al. (2015) Improved estimation and interpretation of correlations in neural circuits. *PLoS Comput Biol* 11: e1004083.
- [16] Perin R, Berger TK, Markram H (2011) A synaptic organizing principle for cortical neuronal groups. *Proceedings of the National Academy of Sciences* : 1–12.
- [17] Ko H, Hofer S, Pichler B, Buchanan K, Sjöström P, et al. (2011) Functional specificity of local synaptic connections in neocortical networks. *Nature* 473: 87–91.
- [18] Song S, Sjöström P, Reigl M, Nelson S, Chklovskii D (2005) Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS biology* 3: e68.
- [19] Marin JM, Pudlo P, Robert CP, Ryder R (2011). Approximate Bayesian computational methods. *arXiv*: 1101.0955.
- [20] Beaumont MA (2010) Approximate Bayesian computation in evolution and ecology. *Annu Rev Ecol Evol Syst* 41: 379–406.
- [21] Kingma DP, Welling M (2013). Auto-encoding variational Bayes. *arXiv*:1312.6114.
- [22] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, et al. (2014). Generative adversarial networks. *arXiv*:1406.2661.
- [23] Tran D, Ranganath R, Blei DM (2017). Deep and hierarchical implicit models. *arXiv*:1702.08896.
- [24] Arjovsky M, Chintala S, Bottou L (2017). Wasserstein GAN. *arXiv*:1701.07875.
- [25] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A (2017). Improved training of Wasserstein GANs. *arXiv*:1704.00028.
- [26] Doersch C (2016). Tutorial on variational autoencoders. *arXiv*:1606.05908.
- [27] Mescheder L, Nowozin S, Geiger A (2017). Adversarial variational Bayes: Unifying variational autoencoders and generative adversarial networks. *arXiv*:1701.04722.
- [28] Hu Z, Yang Z, Salakhutdinov R, Xing EP (2017). On unifying deep generative models. *arXiv*:1706.00550.
- [29] Rosca M, Lakshminarayanan B, Warde-Farley D, Mohamed S (2017). Variational approaches for auto-encoding generative adversarial networks. *arXiv*:1706.04987.
- [30] Goodfellow I (2016). NIPS 2016 tutorial: Generative adversarial networks. *arXiv*:1701.00160.
- [31] Radford A, Metz L, Chintala S (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv*:1511.06434.
- [32] Georgopoulos AP, Kalaska JF, Caminiti R, Massey JT (1982) On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *J Neurosci* 2: 1527–1537.
- [33] Ahmadian Y, Rubin DB, Miller KD (2013) Analysis of the stabilized supralinear network. *Neural Comput* 25: 1994–2037.
- [34] Hennequin G, Ahmadian Y, Rubin DB, Lengyel M, Miller KD (2016). Stabilized supralinear network dynamics account for stimulus-induced changes of noise variability in the cortex. *bioRxiv*:094334.
- [35] Ahmadian Y, Rubin DB, Miller KD (2013) Analysis of the stabilized supralinear network. *Neural Computation* 25: 1994–2037.
- [36] Dale H (1935) Pharmacology and nerve-endings (walter ernest dixon memorial lecture): (section of therapeutics and pharmacology). *Proc R Soc Med* 28: 319–332.
- [37] Strata P, Harvey R (1999) Dale’s principle. *Brain Res Bull* 50: 349–350.
- [38] Arora S, Zhang Y (2017). Do GANs actually learn the distribution? an empirical study. *arXiv*: 1706.08224.
- [39] Arora S, Ge R, Liang Y, Ma T, Zhang Y (2017). Generalization and equilibrium in generative adversarial nets (gans). *arXiv*:1703.00573.
- [40] Mirlin AD (2000) Statistics of energy levels and eigenfunctions in disordered systems. *Physics Reports* 326: 259 - 382.
- [41] Nagarajan V, Kolter JZ (2017). Gradient descent gan optimization is locally stable. *arXiv*:1706.04156.

- [42] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Klambauer G, et al. (2017). GANs trained by a two time-scale update rule converge to a nash equilibrium. *arXiv:1706.08500*.
- [43] Jaeger H, Haas H (2004) Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. *Science* 304: 78–80.
- [44] Sutskever I, Martens J, Dahl G, Hinton G (2013) On the importance of initialization and momentum in deep learning. In: PMLR. pp. 1139–1147. URL <http://proceedings.mlr.press/v28/sutskever13.html>.
- [45] Pascanu R, Mikolov T, Bengio Y (2013) On the difficulty of training recurrent neural networks. In: Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28. pp. III-1310–III-1318. URL <http://dl.acm.org/citation.cfm?id=3042817.3043083>.