Statistical inference in theoretical models of cognition

Sean R. Bittner, *the DSN alliance*, John P. Cunningham

1 Abstract

Theoretical neuroscientists often design circuit models of neural activity with the hopes of producing some emergent property observed in data. However, the standard inference toolkit is designed to condition on data points collected in an experiment, rather than the abstract notion of an emergent property. We introduce a novel machine learning methodology called degenerate solution networks (DSNs), which learn a distribution of generative model parameterizations that produces the emergent property of interest, and is otherwise as random as possible. We use DSNs to advance theoretical understanding of the stomatogastric ganglion (STG), primary visual cortex (V1), superior colliculus (SC), and low rank recurrent neural networks (RNNs). As models continue to increase in complexity and become less tractable in terms of conventional analytic and theoretical approaches, DSNs will be a useful tool for theorists to interrogate their models

2 Introduction

Developing a theory for a neural computation requires a.) a parameterized model of the brain area(s) executing the computation, b.) a mathematical definition of the emergent properties of the model signifying the computation, and then c.) a characterization of the model parameters that produce these emergent properties. Since the advent of theoretical neuroscience, parameterized models of neural systems (a) have become ubiquitous in neuroscience [1]. The emergent properties of interest (b) in these models is predicated by the nervous system function being studied. Prevalent examples include biophysical circuit firing frequency (cite review), memory capacity (cite review), holistic response properties of sensory areas (cite a few), and task execution (cite a few).

In idealized practice of theory, scientists analytically derive parametric solutions to their model that produce the emergent property in focus (c). These derivations often rely on modeling strategies from physics (e.g. [2, 3]). Unfortunately, such gold standard theoretical practices are not always feasible in neuroscience. The desire to make a neural model more biologically realistic and interpretable is at odds with the tractability of its analysis. In cases of such realistic modeling, theorists search for structure in simulated activity [**?**] (cite a bunch here). This approach becomes computationally

intractable as the number of parameters increases, and glaringly lacks a probabilistic treatment of parameter space.

Classically, the probabilistic treatment of model parameter spaces, statistical inference, is designed to condition on an experimentally observed set of data points. This is appropriate in many scientific contexts, but is often unsuitable for theoretical neuroscience. Contrary to the common narrative, theoreticians rarely attempt to directly reproduce experimental data. They work with abstracted mathematical definitions of the requisite neural circuit properties for computation (albeit, these are motivated by experimental findings). Thus, in practice, theoreticians want to condition their interpretable neural circuit models on the abstract notion of an emergent property of computation, *not* a data set. Here, we present a novel machine learning method, degenerate solution networks (DSNs), which are particularly useful for theoretical neuroscience research. DSNs learn a distribution of theoretical model parameterizations that produces some statistically constrained behavior, and is otherwise as random as possible.

DSNs are a tool designed for theorists, that enables a new class of model analyses relying on the full distribution of generative parameters that result in some statistically specified activity. In this study, we use DSNs to advance theory of interpretable models of neural computation ranging from biophysical conductance-based circuits to low-rank RNNs. We begin by introducing the method along with an exploratory analysis of the stomatogastric ganglion (STG) circuit. Then, we apply the same hessian analysis to an inhibitory subtype population model of primary visual cortex (V1) to characterize the possible strategies of inhibition stabilization. By conditioning on task execution with DSNs, we are able to identify a sufficient mechanism of task learning in an interpretable model of superior colliculus, and characterize the role of chaos and limit cycle geometry in the stability of oscillating RNNs. Equipped with this method, we readily attain novel insights into cutting-edge models, suggesting that DSNs should be an integral piece of technology for theoretical neursocience moving forward.

3 Results

3.1 Degenerate solution networks

We have a host of methods from Bayesian machine learning that prescribe how to go from data points through a likelihood model and choice of prior to a posterior distributions on likely parameterizations to have produced such data. But, how do we condition on emergent properties of

behavior that we prefer to define statistically? DSNs combine ideas from likelihood-free variational inference (cite Tran et al) and maximum entropy flow networks (cite Gabe) to make this possible. A maximum entropy flow network is used as a deep generative model for the parameter distribution of the theoretical model, while these samples are passed through a differentiable dynamics simulator, which can 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. Neural circuit models usually have known sampling procedures for simulating activity given a circuit parameterization, yet often lack an explicit likelihood function for the neural activity due to having nonlinear dynamics. DSNs learn a distribution on parameters z, that yields a behavior of interest $\mathcal{B}$,

$$\mathcal{B} : E_{z\sim q_\theta}\left[E_{x\sim p(x|z)}\left[T(x)\right]\right] = \mu \tag{1}$$

by making a deep generative approximation $q_\theta(z)$ to $p(z \mid \mathcal{B})$. So, over the degenerate solution distribution $q_\theta(z)$ of the model for behavior $\mathcal{B}$, the emergent properties (or sufficient statistics) $T(x)$ are constrained in expectation to μ.

In deep generative models, a simple random variable $\omega \sim q_0$ 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 and some behavior of interest, DSNs (Fig. 1A) are trained by optimizing parameters θ to find the best approximation $q^*_\theta(z)$ within the deep generative variational family Q to $p(z \mid \mathcal{B})$ by differentiating through a calculation of the emergent properties given the parameters. This is done by maximizing the entropy of the learned distribution $q^*_\theta(z)$ such that the constraints in $\mathcal{B}$ are satisfied (Fig. 1B):

$$q^*_\theta(z) = \arg\max_{q_\theta \in Q} H(q_\theta(z))$$
$$\text{s.t. } E_{z\sim q_\theta}\left[E_{x\sim p(x|z)}\left[T(x)\right]\right] = \mu$$

(– paragraph about the STG –)

3.2 Exploratory analysis of a theoretical model

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 V1. When

an inhibition stabilized network (ISN, the I population stabilizes an otherwise unstable E population), these models exhibit the paradoxical effect [4], selective amplification [5], surround suppression [6], and sensory integrative properties [7]. Since almost all I neurons fall into one of three classes (parvalbumin (P)-, somatostatin (S)-, and vasointestinal peptide (V)-expressing neurons) [8, 9], theoretical neuroscientists look to extend these dynamical models to four populations [10]. A current challenge in theoretical neuroscience is understanding the distributed role of inhibition stabilization across these inhibitory subtypes.

These four populations exhibit neuron-type specific connectivity (Fig. 1A) [11], in which some populations do not project to others. Since S and V are the only populations that mutually inhibit each other, a popular conceptualization is that S and V have winner-take-all dynamics. In fact, evidence in mice suggests that V silences S when presented with large stimuli, and S silences V for small stimuli [12]. Here, we use DSNs to understand the possible sources of inhibition stabilization in this V1 model, when either S or V is inactive, selecting the weight matrix parameters as the free parameters of the DSN. The behavior of the DSN sampled models is constrained to produce two things: 1.) a mean-zero distribution of ISN coefficients $\gamma(W) = 1 - f'(f^{-1}(r_E(W)))W_{EE}$ with some variance, and 2.) α-population silencing $r_\alpha(W) = 0$, for $\alpha \in \{S, V\}$. When $\gamma < 0$ the network is ISN, and not ISN otherwise. Constraining the DSN behavior to a zero-mean distribution of ISN coefficients gives us samples of both ISN and non-ISN networks, optimizied to have greatest variety of stabilization motifs.

(– paragraph about V1 analyses –)

3.3 Identifying sufficient mechanisms of task learning

In behavioral neuroscience, model organisms are studied while performing tasks in order to investigate the underlying neural computation. As the animal is trained, its accuracy improves via a learning process in the brain. A central challenge for theoreticians is to describe sufficient changes of model parameters that drive task performance, since such changes may indicate how the learning brain adapts. We show that when a data-motivated, restricted dynamical model is proposed, we can use DSNs to clearly identify sufficient changes in network connectivity for task learning.

In a rapid task switching experiment, where rats are 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, 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 [13]. We trained five DSNs on a 4-neuron model of SC proposed by Duan et al. (Fig. 1A, see Methods), constraining the task performance in both the pro and anti tasks to an accuracy p with some allowed variance fixed across chosen p. We constrained the network to emit Bernoulli responses (approximately 0.0 or 1.0 on a given trial), and have winner-take-all behavior between the pro neuron populations of each hemisphere. Altogether, these DSNs, optimized to be expansive and unbiased, learned posteriors of SC model weight matrix parameters, $z = W$, conditioned on different regimes of rapid task switching performance denoted by $\mathcal{B}(p)$ (see Methods).

A convenient property of this dynamical model is that the weight matrix always has the same Schur modes (Fig. 1B), albeit variable eigenvalues for each mode. These Schur modes have intuitive roles with respect to processing in this task, and are accordingly named the *all*, *side*, *task*, and *diag* modes. The degree of amplification of each processing mode in a task performance regime can be examined via $q(S_\alpha(z) \mid \mathcal{B}(p))$, where $S_\alpha(z)$, $\alpha \in \{\text{all}, \text{side}, \text{task}, \text{diag}\}$, is the eigenvalue of the matching Schur mode (Fig. 1C).

As learning progresses, the task mode is increasingly amplified, indicating the criticality of a distributed task representation at the time of stimulus presentation, (Fig. 1D, purple). Stepping from task-naive 50% networks to task-performing 60% networks, there is a switch from amplified (pos. eigs.) to suppressed side mode (neg. eigs.) (Fig. 1D, orange). Side mode suppression is also found in the regimes of greater accuracy, revealing the importance of side mode suppression in allowing a distributed task representation to exist. Across all learning regimes, the diag mode is amplified (Fig. 1D, cyan), and the all mode is suppressed (Fig. 1D, black), which can be seen as signatures of Bernoulli winner-take-all networks. We can conclude that side mode suppression allows rapid task switching, and that greater task-mode representation increases accuracy.

3.4 Conditioning on computation with interpretable models of RNNs

4 Discussion

Still need to write this.

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] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Mario Dipoppa, Adam Ranson, Michael Krumin, Marius Pachitariu, Matteo Carandini, and Kenneth D Harris. Vision and locomotion shape the interactions between neuron types in mouse visual cortex. *Neuron*, 98(3):602–615, 2018.

[13] 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.

[14] Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, 2013.

[15] Gamaleldin F Elsayed and John P Cunningham. Structure in neural population recordings: an expected byproduct of simpler phenomena? *Nature neuroscience*, 20(9):1310, 2017.

[16] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.