REINFORCEMENT LEARNING-GUIDED OPTOGENETIC STIMULATION POLICIES FOR ROBUST FUNCTIONAL NETWORK DISCOVERY

Shoutik Mukherjee[1,2] *Peter Jendrichovksy*[3] *Patrick O. Kanold*[3,4] *Behtash Babadi*[1,2]

[1]Department of Electrical & Computer Engineering
[2]Institute for Systems Research
University of Maryland, College Park
{smukher2, behtash}@umd.edu

[3]Department of Biomedical Engineering
[4]Kavli Neuroscience Discovery Institute
Johns Hopkins University
{pjendri1, pkanold}@jhu.edu

ABSTRACT

Optogenetic stimulation has opened up a new avenue to probe neuronal circuitry at high spatiotemporal resolutions. A key challenge in optogenetic stimulation is deciding which subset out of thousands of neurons should be stimulated to elicit a desired network activation or affect behavior. In this work, we introduce a reinforcement learning approach to adaptively narrow down the multitude of stimulation possibilities and robustly identify Granger causal networks that underlie neuronal activity. We use realistic simulations with different underlying circuitry to show the effectiveness of reinforcement learning in identifying an optimal policy for selecting stimulation targets.

***Index Terms*—** Neuronal Signal Processing, Reinforcement Learning, Optogenetic Stimulation, Granger Causal Networks

1. INTRODUCTION

The development of optogenetic stimulation paradigms has enabled neuroscientists to investigate neuronal circuitry at high spatiotemporal resolutions by perturbing network activity [1, 2, 3]. The applications of optogenetic stimulation experiments are diverse. For instance, optogenetic stimuli have been used to investigate principles of neural coding, such as the role of spike timing in sensory perception in the visual [4] and olfactory [5, 6] systems. The interrogation of ensemble neural codes is also possible by stimulating population responses. Optogenetic stimulation of neurons active during behavior has been shown to elicit ensemble activity that correlates strongly with natural stimulus responses, shedding light on latent neural dynamics that drive behavior [6, 7, 8, 9]. The population-wide effects of individual neuronal activity in part motivates functional network analysis of neuronal assemblies, either from responses evoked by sensory stimuli as in the aforementioned examples or from spontaneous activity.

In the preceding examples, stimulated neurons were known to be active during behavior; however, selecting stimulation targets is nontrivial without prior knowledge about the network. Moreover, the statistical confidence of functional network estimates from spontaneous recordings is encumbered by low population activity that necessitate longer recordings. Optogenetic stimulation can be used to induce population activity, but exhaustive stimulation of every possible target is intractable due to the size of the neuronal assembly.

Viewing the neurons as agents competing for a shared resource (i.e. to be stimulated), the search for a stimulation target selection policy lends itself to a multi-agent reinforcement learning (MARL) paradigm [10, 11, 12]. Reinforcement learning (RL) approaches require an agent to adapt its actions based on rewards received from its environment; as the agent accumulates rewards for its actions, its policy adapts to maximize the cumulative reward [13]. By extension, MARL considers several agents in the same environment competing for shared resources, and identifies a policy that satisfies an equilibrium at which the cumulative reward for any agent cannot be increased without decreasing that of another. Because MARL scales poorly with the number of agents, mean-field approximations to the interactions between agents have been developed to tractably identify optimal policies [11, 12, 14].

In this work, we adapt the reinforcement-learning mean-field game (RL-MFG) approach developed in [12] to adaptively identify optimal policies for selecting stimulation targets. In a realistic simulation setting, we employ this approach to identify optimal policies for four neuronal networks with differing circuitry, then use Granger Causality [15, 16] analysis to estimate the network structure from spiking responses generated by stimulating neurons according to the optimal policies. Contrasting these networks with those estimated from spontaneous activity of the same duration, we find that RL-MFG optimized stimulation enables accurate network discovery with less data.

Several applications of RL in neuroscience have been explored recently, including: deep brain stimulation controllers [17, 18]; and relating neural activity to motor control [19] or cognitive constructs such as representation and memory [20, 21]. However, to the best of our knowledge, neither MARL nor its mean-field approximations have been previously used in neuroscience applications. Thus, while this work draws on existing work (primarily [12], which relates closely to [22]), the application considered here constitutes a novel contribution to current research in network neuroscience.

This work was supported in part by National Science Foundation award no. ECCS2032649, and National Institutes of Health awards no. 1U19NS107464-01 and 5R01DC017785-03.

2. METHODS

2.1. Reinforcement Learning Algorithm Development

To find a policy for selecting optogenetic stimulation targets, we first formulate the search as a MARL problem where the N neurons in a network are viewed as Markov agents [10]. Supposing that spiking activity is observed for T samples, the state of neuron n is defined as its activity over the observed duration, i.e. its spike train $\boldsymbol{s}^{(n)} = [s_1^{(n)}, \ldots, s_T^{(n)}]$, with $\boldsymbol{s}^{(n)} \in \mathcal{S} := \{0,1\}^T$. The action of a neuron is denoted by $a^{(n)} \in \mathcal{A} := \{0,1\}$, where $a^{(n)} = 1$ indicates that neuron n is stimulated. The policy for stimulating neuron n is defined as $\pi^{(n)} := \mathbb{P}[a^{(n)} = 1]$.

Finding the optimal set of policies $\{\pi^{(n)}\}_{n=1:N}$ involves modeling interactions between the agents [10, 11]. Not only does this scale poorly with N, but is limited by the inherent lack of information about the network structure. Mean-field approximations have been used to address tractability [11, 12, 14]. With a mild assumption that only a proportion p of all possible links in a network exist, we employ a mean-field approximation that also circumvents knowledge of the latent network structure. We adapt the natural actor-critic (NAC) algorithm of Mao et al. [12] to identify an optimal policy for this mean-field game. The following sections recap essential elements of the method as adapted to the present setting.

2.1.1. Mean-Field Approach Preliminaries

In contrast to the N-agent Markov game considered in MARL, the mean-field approximation considers an infinite number of identical agents where the population's collective behavior is described by the mean-field state. It suffices to consider one representative agent with state $\boldsymbol{s}_k$ in relation to the mean-field $\boldsymbol{\mu}_k$, the subscript k denoting the state of the agent after the k^{th} stimulation pattern. This simplifies a multi-agent policy optimization problem to a single-agent problem; i.e., we are now searching for the probability of stimulating one neuron, and will repeat the forthcoming procedure for all neurons in the network.

Given a population response $\boldsymbol{\mu}_k$, an action a_k is selected according to the stimulation policy. The policy of the agent should be adjusted based on the population response to stimulation, so a reward $r(a_k, \boldsymbol{\mu}_k)$ is computed, and the states of the agent and mean-field are updated according to the transition:

$$\begin{aligned}\mathbb{P}\left[s_{k+1,t} = 1 | a_k\right] &= \text{logit}^{-1}\Big(\nu + p\boldsymbol{\omega}^\top \boldsymbol{\mathcal{H}}_t^{\boldsymbol{\mu}} \\ &\qquad + U\mathbb{1}\{a_k = 1, t = 1\}\Big), \\ \mu_{k+1,t} &= \text{logit}^{-1}\Big(\nu + p\boldsymbol{\omega}^\top \boldsymbol{\mathcal{H}}_t^{\boldsymbol{\mu}} \\ &\qquad + U\mathbb{P}\left[a_k = 1\right]\mathbb{1}\{t = 1\}\Big).\end{aligned} \tag{1}$$

Note the population response's effect on the state of the agent, the cross-history coefficients $\boldsymbol{\omega}$, is weighted by the link probability p, representing the average network effect on the agent. Here, $\boldsymbol{\mathcal{H}}_t^{\boldsymbol{\mu}}$ denotes the recent history of the mean-field response; for early time indices, this includes overlap with the previous mean-field state vector $\boldsymbol{\mu}_k$.

Additionally, note that $\boldsymbol{\mu}_{k+1,t} = \mathbb{E}_{a_k}\left[\mathbb{P}\left[s_{k+1,t} = 1 | a_k\right]\right]$. The action $a_k = 1$ with probability $\pi(\boldsymbol{\mu}_k; \theta) = \text{logit}^{-1}(\theta\phi_k)$, where $\phi_k := \boldsymbol{\Gamma}^\top \boldsymbol{\mu}_k$ is a feature measuring the population response that weights the time indices shortly after potential stimulation. The response kernel $\boldsymbol{\Gamma}$ is composed of uniform samples of a Γ-density whose mode is at time index 4, matching the latency of the cross-history kernel $\boldsymbol{\omega}$. The reward function is also defined in terms of ϕ_k: $r(a_k, \boldsymbol{\mu}_k) := \mathbb{1}\{a_k = 1\}\exp\left(\phi^2\right)$. For notational compactness, the reward will be denoted as r_k.

2.1.2. Policy Update via Natural Actor-Critic Algorithm

Policies were updated by maximizing the entropy-regularized value function [22] with respect to the parameter θ each time after the mean-field is updated (i.e. after the population response to a stimulation pattern is observed) using a natural actor-critic (NAC) algorithm proposed by Mao et al. [12]. Essential details are recapitulated here. For some initial mean-field $\boldsymbol{\rho}$, the regularized value function is defined as

$$V_{\boldsymbol{\mu}_k}^{\pi,\beta}(\boldsymbol{\rho}) := \mathbb{E}\left[\sum_{l=1}^{\infty}\gamma^l\Big(r_k - \beta\big(a\log(\pi_\theta) + (1-a)\log(1-\pi_\theta)\big)\Big)\,\Big|\,\boldsymbol{s}_0 \sim \boldsymbol{\rho}\right], \tag{2}$$

where the first term is the unregularized value function with discount factor γ, and the second term is the Shannon entropy of the policy with regularization coefficient β. The maximizing value of parameter θ can be obtained by natural policy gradient ascent, which rotates and rescales the gradient $\nabla_\theta V_{\boldsymbol{\mu}_k}^{\pi,\beta}(\boldsymbol{\rho})$ by the inverse Fisher information matrix.

Mao et al. [12] and Cayci et al. [22] show that this equates to an update of the form $\theta \leftarrow \theta + \eta/(1-\gamma)w_\beta^\theta$, where

$$w_\beta^\theta = \underset{w:\|w\|_2 \le R}{\arg\min}\ \mathbb{E}_a\left[\left(w^\top \nabla_\theta \log\pi_\theta - A_{\mu_k}^{\pi_\theta,\beta}\right)^2\right], \tag{3}$$

for some learning rate η. Here, $A_{\mu_k}^{\pi_\theta,\beta}$ is the centered soft Q-function that essentially quantifies the unreqularized cumulative reward. The shifted Q-function, i.e. the entropy-regularized cumulative reward, defined as

$$q_{\mu_k}^{\pi_\theta,\beta} := r_k - \beta\big(a_k\log(\pi_\theta) + (1-a_k)\log(1-\pi_\theta)\big) + \gamma\mathbb{E}_{\boldsymbol{s}}\left[V_{\boldsymbol{\mu}_k}^{\pi,\beta}(\boldsymbol{s})\right], \tag{4}$$

allows $A_{\mu_k}^{\pi_\theta,\beta}$ to be expressed as

$$\begin{aligned}A_{\mu_k}^{\pi_\theta,\beta} = q_{\mu_k}^{\pi_\theta,\beta} &+ \beta\big(a_k\log(\pi_\theta) + (1-a_k)\log(1-\pi_\theta)\big) \\ &- \mathbb{E}_{a'}\left[q_{\mu_k}^{\pi_\theta,\beta} + \beta\big(a'\log(\pi_\theta) + (1-a')\log(1-\pi_\theta)\big)\right].\end{aligned} \tag{5}$$

The NAC algorithm proposed by Mao et al. [12] is used to optimize θ in two steps. First, in the policy evaluation step, an estimated of $q_{\mu_k}^{\pi_\theta,\beta}$, denoted by $\hat{q}_k^\beta$, is obtained using a temporal difference learning algorithm that approximates the shifted Q-function as a linear function of ϕ_k (Algorithm 2 in [12]; the "critic"). Then, an estimate of $A_{\mu_k}^{\pi_\theta,\beta}$ using $\hat{q}_k^\beta$, denoted by $\hat{A}_k^\beta$, is used in a stochastic gradient descent algorithm to obtain an estimate of the gradient w_β^θ, denoted $\hat{w}_k$ (Algorithm 3 in [12]; the "actor").

The NAC algorithm is applied to each neuron in a network after observing the network responses to a stimulation pattern. The probability of stimulating each neuron is updated after each pattern; averaging these updated policies for each neuron gives their optimal stimulation probability. The complete procedure is summarized in Algorithm 1; the learning rate η was selected so that $\eta = \mathcal{O}\big(K^{(-2/5)}/\beta\big)$ [12].

2.2. Simulating Neuronal Activity

To demonstrate the utility of Algorithm 1 and the advantage provided in Granger causal network inference [15, 16], we simulated neuronal activity of a network with 4 independent subnetworks of $N = 8$

Algorithm 1 Optimal Stimulation Policy Discovery

Input: $K, \beta, \gamma, \eta, N$
Output: $\{\pi^{(n)*}\}_{n=1}^{N}$

$\theta^{(n)} = 0$, $\pi_1^{(n)} = 0.5$, for $n = 1, \ldots, N$
for $k = 1$ to K **do**
  Sample a_k from PMF induced by $\{\pi_k^{(n)}\}_{n=1}^{N}$
  $a_k^{(n)} = \mathbb{1}\{a_k = n\}$ for $n = 1, \ldots, N$
  Observe spiking responses to stimulation pattern
  Compute $\boldsymbol{\mu}_k$, the network average spiking response
  for $n = 1$ to N **do**
    Estimate $\hat{q}_k^{\beta^{(n)}}$ (Algorithm 2 in [12])
    Calculate $\hat{w}_k^{(n)}$ using $\hat{q}_k^{\beta^{(n)}}$ (Algorithm 3 in [12])
    $\theta^{(n)} \leftarrow \theta^{(n)} + \eta(\hat{w}_k^{(n)} - \beta\theta^{(n)})$
  end for
  Compute $\phi_k^{(n)}$
  $\pi_{k+1}^{(n)} = \text{logit}^{-1}(\theta^{(n)}\phi_k^{(n)})$
end for
for $n = 1$ to N **do**
  $\pi^{(n)*} = \frac{1}{K}\sum_{k=2}^{K+1} \pi_k^{(n)}$
end for
return $\{\pi^{(n)*}\}_{n=1}^{N}$

neurons each (Fig. 1). The subnetworks had unique structure. The first subnetwork had one broadcast neuron with degree-out of 6 and degree-in of 0. The second subnetwork was a chain with source-target neuron pairs $\{(n, n+1)\}_{n=1:5}$. The third subnetwork had one sink neuron with degree-out of 0 and degree-in of 6. The number of links in each of these subnetworks is approximately $p = 0.1$ of the total number of possible links (56). Links between neurons in the fourth subnetwork were selected randomly with probability $p = 0.1$; two neurons had degree-out of 1, and one had degree-out of 3.

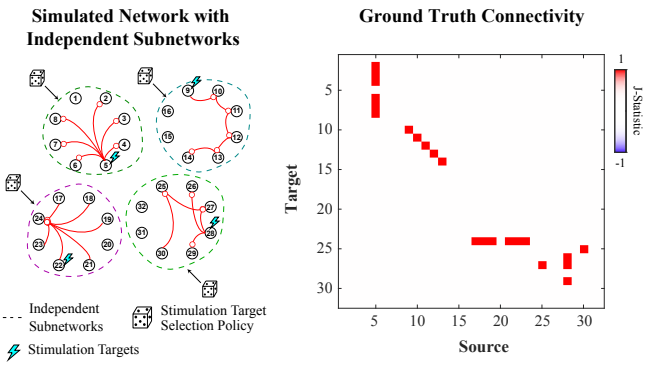

Fig. 1: Simulated Network. The network consisted of 4 independent subnetworks with varied structure, but the same level of connectivity ($\sim 10\%$ of all possible links). One stimulation target per subnetwork was chosen according to a learned policy.

For each single-cell stimulation pattern, simulated spiking activity of each neuron was generated for $L = 5$ repetitions of duration $T = 60$ samples. The spiking activity of neuron n was generated as a Bernoulli process whose conditionally independent success probabilities are given by the conditional intensity function (CIF) $\lambda_{l,t}^{(n)}$ at time t of trial l. The effect of optogenetic stimulation is to elicit a spike from the target neuron with high probability. Hence, it was modeled by an instantaneous increase in a neuron's CIF; namely,

$$\lambda_{l,t}^{(n)} = \text{logit}^{-1}\left(\nu + \sum_{n'=1}^{N} \boldsymbol{\omega}^{(n,n')^\top} \boldsymbol{\mathcal{H}}_{l,t}^{(n')} + U\mathbb{1}\{a^{(n)} = 1, t = 1\}\right), \quad (6)$$

where $\nu = -3$ denotes the base rate parameter of all neurons, $U\mathbb{1}\{a^{(n)} = 1, t = 1\}$ with $U = 6$ models the effect of the action to stimulate neuron n, $\boldsymbol{\mathcal{H}}_{l,t}^{(n')}$ the recent spiking history of neuron n', and $\boldsymbol{\omega}^{(n,n')}$ the cross-history coefficients that characterize the influence of neuron n' on neuron n. The cross-history coefficients were chosen to most likely induce spiking with a latency of 4–5 time bins. For all source-target pairs (n, n') for which a link existed in the network model, the cross-history coefficients were identical ($\boldsymbol{\omega}^{(n,n')} = \boldsymbol{\omega}$), and were $\mathbf{0}$ otherwise. Self-history coefficients $\boldsymbol{\omega}^{(n,n)} = \mathbf{0}$.

2.3. Granger Causality Analysis

Granger causality (GC) analysis evaluates the predictive influence of the recent spiking activity history of one neuron on the present spiking activity of another. Here, we adapted GC network inference procedure introduced in [15, 16] for point process models of neuronal spiking activity that accounts for sparse network interactions and controls the false discovery rate of GC links. The existence of a GC link from neuron $\tilde{n}$ to n was tested by comparing a full model of neuron n, parameterized by estimated cross-history coefficients $\left\{\hat{\boldsymbol{\omega}}^{(n,n')}\right\}_{n'=1}^{N}$ and base rate $\hat{\nu}^{(n)}$, to a reduced model that assumed $\hat{\boldsymbol{\omega}}^{(n,\tilde{n})} = \mathbf{0}$ using the de-biased deviance difference. The strength of GC links were characterized by Youden's J-statistic following false discovery rate control at a rate of 0.01.

3. RESULTS

In a simulation study outlined in Section 2.2, we demonstrate Algorithm 1 and the advantage of optimal stimulation policies for GC network inference [15, 16]. Algorithm 1 was applied to each independent subnetwork to simultaneously evaluate its performance given different latent network structures. The procedure for one subnetwork is visualized in Fig. 2**A**. A small discount factor of $\gamma = 0.1$ was used to prioritize the reward in the temporal difference learning algorithm used in Step 8 of Algorithm 1. Regularization coefficients of multiple orders of magnitude were considered, but the results presented here utilized $\beta = 1000$; smaller values were found to be no different than unregularized policy estimation, and larger values overregularized. The number of iterations was set to $K = 100$, and the learning rate was $\eta \approx 1.6 \times 10^{-4}$.

The optimal stimulation probabilities (Fig. 2**B**, top) reflected the structure of each subnetwork, showing in Fig. 3**A**. For instance, neuron 5 in subnetwork 1 had the highest probability of stimulation; as it is the source of 6 links, one could expect its activity to elicit a large subnetwork-wide response. In contrast, neurons in subnetwork 3 did not have remarkably higher probabilities than others because each neuron only linked to one other neuron. Probabilities significantly smaller than 0.5 (t-test, $p < 0.05$) were set to 0, and a probability mass function (PMF) was formed by normalizing these pruned probabilities (Fig. 2**B**, bottom).

This PMF was used to sample stimulation targets; specifically, 20 stimulation patterns were independently drawn (20 sets of 4 neurons, 1 per subnetwork). The first 30 samples of the network responses to stimulation were used to estimate GC networks [15, 16] with false discovery rate control [23] (Fig. 3**B**, top left). Spontaneous spiking

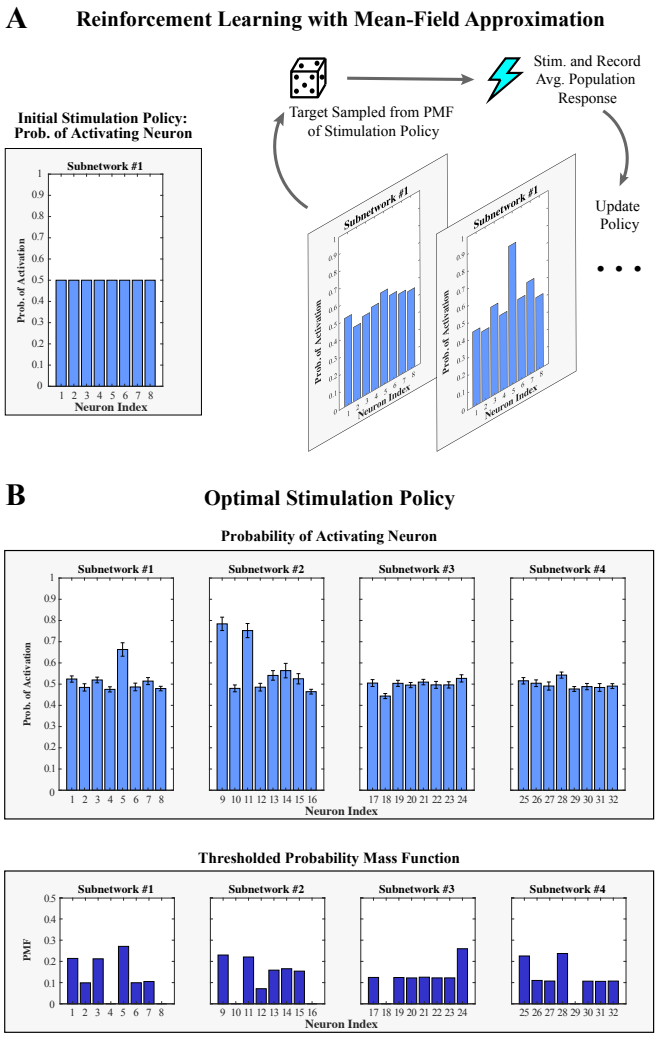

Fig. 2: Reinforcement Learning of Optimal Stimulation Policies. **A.** With probabilities of activating a neuron initialized to 0.5, Algorithm 1, applied iteratively, updated the probability of activating each neuron **B.** The optimal probabilities of activation, the average of all updates, are displayed with error bars corresponding to 2 SEM (top). Probabilities significantly less than 0.5 (t-test, $p < 0.05$) are pruned, and stimulation targets are sampled from the induced PMF (bottom).

of the same duration was similarly analyzed as a control (Fig. 3**B**, bottom left). The resulting networks were comparable. Both had similar hit rates and false discovery rates (Table 1), computed across the network.

However, when a subset of stimulation data corresponding to 10 patterns and spontaneous activity of the same length were used for GC analysis, a notable difference was evident (Fig. 3**B**, right column). While the network estimated from stimulated data had a hit rate of 100% and false discovery rate of 12%, the spontaneous activity-based estimate only had a hit rate of 54.5% due to low activity (Table 1), computed across the network. Subnetwork-specific hit rates differed notably in subnetwork 1 (100% vs. 16.67% in stimulated vs. spontaneous data) and subnetwork 4 (100% vs. 50%), where source neurons were linked to multiple targets. Thus, the simulated results demonstrate that stimulation enables recovery of latent network structure

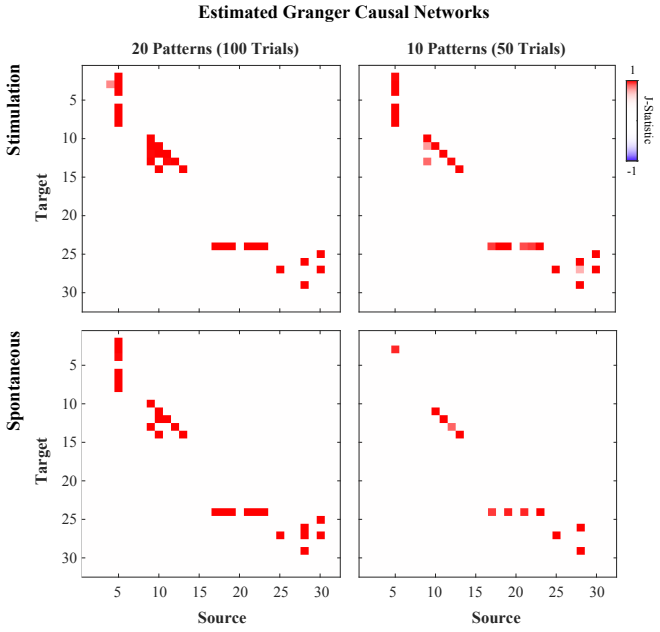

Fig. 3: Granger Causal Network analysis of network with sparse activity. GC networks estimated from stimulated spiking (top row) and spontaneous spiking (bottom row) were comparable when 20 patterns, i.e., 100 trials, were used (left column), but only half the links were detected when 10 patterns, i.e., 50 trials, were used (right column).

		20 Patterns	10 Patterns
Stimulation	Hit Rate	95.5%	100%
	FDR	25.0%	12.0%
Spontaneous	Hit Rate	100%	54.5%
	FDR	15.4%	0.0%

Table 1: Global hit rate and false discovery rate (FDR) of network inference results from Fig. 3.

from sparse neuronal activity based on parsimonious data acquisition.

4. DISCUSSION

By adapting a reinforcement learning algorithm for single-agent policy optimization in a mean-field game, we have introduced an adaptive approach to guide the selection of optogenetic stimulation targets while probing neuronal populations. In simulations, we showed that the algorithm can identify neurons that act as network hubs to elicit population-wide effects. Moreover, we have shown that stimulating such neurons indeed enables recovery of latent network structure with statistical confidence and less data than necessary when examining spontaneous activity.

In regards to experimental validation of the RL-MFG approach, implementing optogenetic stimulation is straightforward but requires real-time analysis, a technical challenge we aim to address in future work. Additionally, the optimality of stimulation policies warrants further investigation, though a closer examination of the formation of PMFs from learned probabilities is first needed. Future work would also include applications to networks with more complex dynamics, and extensions for optimal combinations of neuronal stimulation.

5. REFERENCES

[1] Valentina Emiliani, Adam E. Cohen, Karl Deisseroth, and Michael Häusser, "All-optical interrogation of neural circuits," *Journal of Neuroscience*, vol. 35, no. 41, pp. 13917–13926, 2015.

[2] Adam M. Packer, Lloyd E. Russell, Henry W. P. Dalgleish, and Michael Häusser, "Simultaneous all-optical manipulation and recording of neural cellular activity with cellular resolution *in vivo*," *Nature Methods*, vol. 12, pp. 140–146, 2015.

[3] Or A Shemesh, Dimitrii Tanese, Valeria Zampini, Changyang Linghu, Kiryl Piatkevich, Emiliano Ronzitti, Eirini Papagiakoumou, Edward S Boyden, and Valentina Emiliani, "Temporally precise single-cell-resolution optogenetics," *Nature Neuroscience*, vol. 20, pp. 1796–1806, 2017.

[4] Julian Day-Cooney, Jackson J. Cone, and John H. R. Maunsell, "Perceptual weighting of v1 spikes revealed by optogenetic white noise stimulation," *Journal of Neuroscience*, vol. 42, no. 15, pp. 3122–3132, 2022.

[5] Jonathan V. Gill, Gilad M. Lerman, Hetince Zhao, Benjamin J. Stetler, Dmitry Rinberg, and Shy Shoham, "Precise holographic manipulation of olfactory circuits reveals coding features determining perceptual detection," *Neuron*, vol. 108, no. 2, pp. 382–393.e5, 2020.

[6] Edmund Chong, Monica Moroni, Christopher Wilson, Shy Shoham, Stefano Panzeri, and Dmitry Rinberg, "Manipulating synthetic optogenetic odors reveals the coding logic of olfactory perception," *Science*, vol. 368, no. 6497, pp. eaba2357, 2020.

[7] Marco dal Maschio, Joseph C. Donovan, Thomas O. Helmbrecht, and Herwig Baier, "Linking neurons to network function and behavior by two-photon holographic optogenetics and volumetric imaging," *Neuron*, vol. 94, no. 4, pp. 774–789.e5, 2017.

[8] Luis Carrillo-Reid, Shuting Han, Weijian Yang, Alejandro Akrouh, and Rafael Yuste, "Controlling visually guided behavior by holographic recalling of cortical ensembles," *Cell*, vol. 178, no. 2, pp. 447–457.e5, 2019.

[9] James H. Marshel, Yoon Seok Kim, Timothy A. Machado, Sean Quirin, Brandon Benson, Jonathan Kadmon, Cephra Raja, Adelaida Chibukhchyan, Charu Ramakrishnan, Masatoshi Inoue, Janelle C. Shane, Douglas J. McKnight, Susumu Yoshizawa, Hideaki E. Kato, Surya Ganguli, and Karl Deisseroth, "Cortical layer–specific critical dynamics triggering perception," *Science*, vol. 365, no. 6453, pp. eaaw5202, 2019.

[10] Michael L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.

[11] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang, "Mean field multi-agent reinforcement learning," in *Proceedings of the 35th International Conference on Machine Learning*. 10–15 Jul 2018, vol. 80 of *Proceedings of Machine Learning Research*, pp. 5571–5580, PMLR.

[12] Weichao Mao, Haoran Qiu, Chen Wang, Hubertus Franke, Zbigniew Kalbarczyk, Ravi Iyer, and Tamer Basar, "A mean-field game approach to cloud resource management with function approximation," in *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, Eds., 2022.

[13] Richard S Sutton and Andrew G Barto, *Reinforcement learning: An introduction*, MIT press, 2018.

[14] Minyi Huang, Roland P. Malhamé, and Peter E. Caines, "Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle," *Communications in Information & Systems*, , no. 3, pp. 221–252, 2006.

[15] Sanggyun Kim, David Putrino, Soumya Ghosh, and Emery N. Brown, "A Granger causality measure for point process models of ensemble neural spiking activity," *PLOS Computational Biology*, vol. 7, no. 3, pp. e1001110, 03 2011.

[16] Alireza Sheikhattar, Sina Miran, Ji Liu, Jonathan B. Fritz, Shihab A. Shamma, Patrick O. Kanold, and Behtash Babadi, "Extracting neuronal functional network dynamics via adaptive Granger causality analysis," *Proceedings of the National Academy of Sciences*, vol. 115, no. 17, pp. E3869 – E3878, 2018.

[17] Qitong Gao, Michael Naumann, Ilija Jovanov, Vuk Lesi, Karthik Kamaravelu, Warren M. Grill, and Miroslav Pajic, "Model-based design of closed loop deep brain stimulation controller using reinforcement learning," in *2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems (ICCPS)*, 2020, pp. 108–118.

[18] Meili Lu, Xile Wei, Yanqiu Che, Jiang Wang, and Kenneth A. Loparo, "Application of reinforcement learning to deep brain stimulation in a computational model of parkinson's disease," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 1, pp. 339–349, 2020.

[19] Martin Spüler, Sebastian Nagel, and Wolfgang Rosenstiel, "A spiking neuronal model learning a motor control task by reinforcement learning and structural synaptic plasticity," in *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–8.

[20] Matthew Botvinick, Jane X. Wang, Will Dabney, Kevin J. Miller, and Zeb Kurth-Nelson, "Deep reinforcement learning and its neuroscientific implications," *Neuron*, vol. 107, no. 4, pp. 603–616, 2020.

[21] Sandeep Sathyanandan Nair, Vignayanandam Ravindernath Muddapu, C. Vigneswaran, Pragathi P. Balasubramani, Dhakshin S. Ramanathan, Jyoti Mishra, and V. Srinivasa Chakravarthy, "A generalized reinforcement learning based deep neural network agent model for diverse cognitive constructs," *Scientific Reports*, vol. 13, no. 5928, 2023.

[22] Semih Cayci and Niao He, "Linear convergence of entropy-regularized natural policy gradient with linear function approximation," *arXiv preprint arXiv:2106.04096*, 2021.

[23] Yoav Benjamini and Daniel Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *The Annals of Statistics*, vol. 29, no. 4, pp. 1165–1188, 2001.