Unsupervised sparse deconvolutional learning of features driving neural activity*

Bahareh Tolooshams[1], Hao Wu[2], Naoshige Uchida[2], Venkatesh N. Murthy[2], Paul Masset[2], and Demba Ba[1]

[1]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA
[2]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA

Understanding the activity of single neurons in relation to features in the environment is the first step in many neuroscience studies. We propose a method using algorithm unrolling, an emerging technique in interpretable deep learning, to deconvolve single-trial neuronal activity into interpretable components. Specifically, we model the firing rates of single neurons using a set of kernels characterizing neurons' responses to time-sensitive sparse events/stimuli. The kernels can be either unique or shared across the population and are weighted by codes whose amplitude and timing are trial-specific. Our inference results in a deep sparse deconvolutional encoder and, unlike sequential deep encoder approaches, is based on a generative model; hence, the learned parameters and encoding are directly interpretable. First, we characterize the performance regime of our method; this guides end users to understand the model's accuracy and limitations. Second, we apply our method to deconvolve overlapping signals in the response of dopaminergic neurons to rewards of varying size. Previous studies have suggested that reward prediction error responses of dopaminergic neurons are modulated by two components: salience and value. However, this multiplexing is often ignored or analyzed using ad-hoc windows to estimate the two contributions. Here, we deconvolve the two factors in an unsupervised manner; one kernel corresponds to salience whose code is common across reward sizes and another to value whose code changes as a function of reward amount. We show that the inferred codes are more informative than firing rates estimated using ad-hoc windows. Third, we study the response of piriform cortex neurons to brief odor pulses delivered at random time across trials. Based on the learned neural impulse responses, we uncover 3 clusters of response types across the population. Overall, we propose a novel method to deconvolve into interpretable components the factors driving neural activity in single trials.

Methods. Given a neuron's activity, the spikes at trial j are binned at B ms resolution and modeled using the

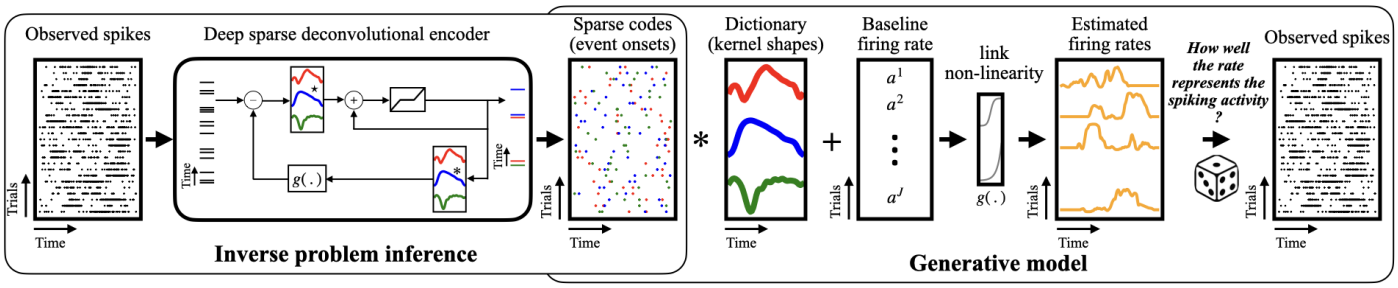

Fig. 1: Sparse deconvolutional learning (SDL) framework.

natural exponential family (i.e., $\boldsymbol{y}^j \sim \text{Poisson}(\boldsymbol{\mu}^j)$ or $\boldsymbol{y}^j \sim \text{Binomial}(B, \boldsymbol{\mu}^j)$). We impose a generative model (Fig. 1) on the neuron's firing rate and model it as a function of a baseline activity a^j and a set of localized kernels $\{\boldsymbol{h}_k\}_{k=1}^K$ characterizing the neuron's response to events that occur sparsely in time. The events' onsets are coded with a sparse code $\boldsymbol{x}^{j,k}$ whose amplitude encodes the strength of the contribution of the k^{th} kernel to the neuron's response (i.e., $\boldsymbol{\mu}^j = g(\sum_{k=1}^K \boldsymbol{h}_k * \boldsymbol{x}^{j,k} + a^j)$ where g is a non-linear function). Although the model results in an estimate of the neuron's firing rate on a trial basis, the kernels capture characteristics that are shared among trials of a neuron or neural population. Given the spike counts $\boldsymbol{y}^j$, we learn the kernels and codes by minimizing the negative log-likelihood with a sparse prior on the codes, i.e.,

$$\min_{\{\boldsymbol{h}_k\}_{k=1}^K, \{\boldsymbol{x}^{j,k}\}_{k=1}^K} -\sum_{j=1}^J \log p(\boldsymbol{y}^j \mid \{\boldsymbol{h}_k, \boldsymbol{x}^{j,k}\}_{k=1}^K) + \sum_{k=1}^K \lambda_k \|\boldsymbol{x}^{j,k}\|_1 \quad \text{s.t.} \quad \|\boldsymbol{h}_k\|_2 = 1 \text{ for } k = 1, \dots, K \tag{1}$$

which we call sparse deconvolutional learning (SDL). We use algorithm unrolling [1, 2] to map the problem into an interpretable deep neural network (Fig. 1) [3]. The inference is a deep sparse deconvolutional encoder performing proximal gradient descent (i.e., $\boldsymbol{x}_t^{j,k} = \boldsymbol{e}^{j,k} \cdot \mathcal{S}_{\alpha\lambda_k}\left(\boldsymbol{x}_{t-1}^{j,k} + \alpha \boldsymbol{h}_k \star (\boldsymbol{y}^j - g(\sum_{v=1}^K \boldsymbol{h}_v * \boldsymbol{x}^{j,v} + a^j))\right)$ where $\mathcal{S}$ is shrinkage, $\star$ the correlation operator, and $\boldsymbol{e}^{j,k}$ an indicator for known events). The combined generative and inference networks are trained to estimate the sparse events and learn the kernels such that the data likelihood is maximized.

*This work is accepted to *Computational and Systems Neuroscience, 2022*.

Model characterization. We generated 100 trials as we varied background firing rate and bin resolution. In each trial, 5 similar events happen uniformly at random with a minimum distance of 100 ms. We set the length of the neural response to 500 ms, and the response strength (i.e., code amplitude $\boldsymbol{x}^{j,k}$) follows a uniform distribution Unif(10, 40). We model the data using the Binomial distribution. First, when neither kernel nor event timings are known, kernel recovery error (i.e., $\sqrt{1-(\text{cosine similarity})^2}$, darker is better) improves as bin size B increases (Fig. 2a). When the kernel shape is known but events' timing are unknown, the event recovery error (i.e., $1-\frac{\#\text{ identified events}}{\#\text{ total events}}$) improves as bin size and neural activity increase (Fig. 2b). Finally, we show that when the events' timing are known, the model is robust to background firing rate and bin size, as measured in terms of fraction of variance unexplained for code amplitude recovery, regardless of whether the kernels are learned (Fig. 2c) or known (Fig. 2d). These simulations empower end users to assess the reliability of the recovered factors given their data statistics.

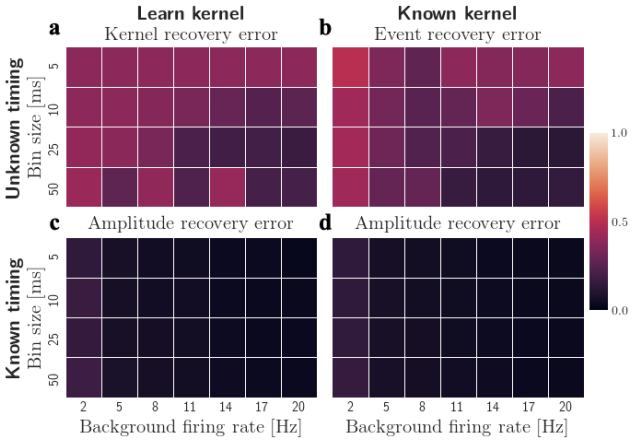

Fig. 2: Model characterization.

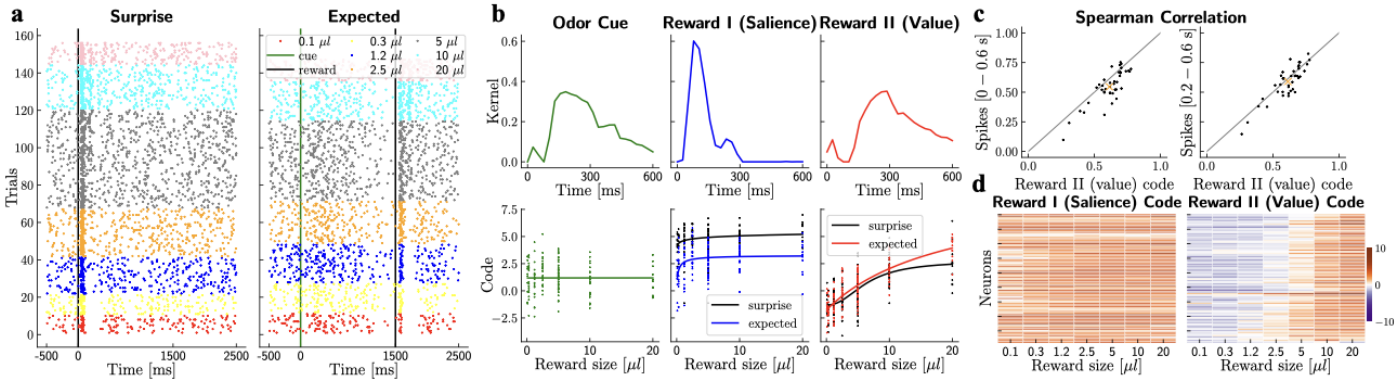

Fig. 3: Deconvolution of reward prediction error responses in dopaminergic neurons.

Dopamine. We study 40 optogenetically identified dopaminergic neurons recorded in a classical conditioning task [4]; in *surprise* trials, a size-varying reward (i.e., 0.1 to 20 μl) was delivered without cue, and in *expected* trials, an odor cue preceded reward delivery by 1.5 s (Fig. 3a). Prior work has shown that the reward prediction error of dopaminergic to a cue or a reward has two overlapping response components: a fast one but largely unselective in terms of value (salience), followed by a longer response strongly modulated by the value of the cue or reward [5]. Here, we show that SDL deconvolves these two overlapping kernels without supervision. We learn three non-negative kernels of length 600 ms using $B = 25$ ms; one characterizes the neural response to the odor cue in *expected* trials (green), and two reward-related kernels which strongly resemble salience (blue) and value (red, Fig. 3b). Given this decomposition, as an alternative to spike counts from ad-hoc windows, we can use the code amplitudes in single trials from each kernel as a measure of the neurons' tuning to reward amount (bottom row of Fig. 3b, fitted using a Hill function as in [4]). The codes for the odor cue and the salience kernel are essentially invariant to the reward amount (Fig. 3d left), but they are modulated by context (*surprise* vs. *expected*). The value code is strongly modulated by the reward amount. We demonstrate using Spearman's rank correlation that the value code carries more information about reward amount than the firing rates over ad-hoc windows (Fig. 3c, i.e., the average across all neurons (orange marker x) lies under the diagonal line, p=$2 \cdot 10^{-6}$ and p=0.037, t-test). This will allow experimenters to infer with higher accuracy than previous methods the parameters of single neurons for a given dataset. Furthermore, the value codes (Fig. 3d right) show a diverse sensitivity to reward size across the neural population, a potential signature of distributional reinforcement learning in dopaminergic neurons [6].

Olfaction. Next, we apply our method to an olfactory task in which stimuli occur at random times across trials. At each trial, 50 ms

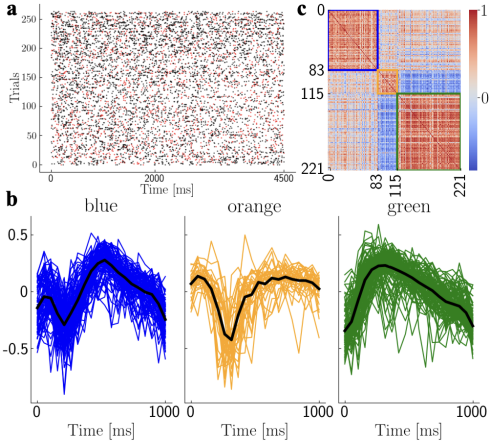

Fig. 4: Structure of piriform responses.

Poisson-distributed odor pulses (red dots in Fig. 4a) are delivered. 221 neurons were recorded and isolated from mice anterior piriform cortex. The data are downsampled to 1 ms resolution and the recorded spikes (black dots in Fig. 4a) were analyzed with $B = 50$ ms. We model the odor pulses by sparse codes with known timings and spike counts as a Poisson process. We learn one kernel for each neuron and identify 3 clusters in the population based on the kernel shapes (spectral clustering using cosine distances, Fig. 4b-c) highlighting how our method can be used for exploratory data analysis.

References

[1] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proc. International Conference on Machine Learning*, 2010, pp. 399–406.

[2] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *arXiv:1912.10557*, pp. 1–27, 2019.

[3] B. Tolooshams, A. Song, S. Temereanca, and D. Ba, "Convolutional dictionary learning based auto-encoders for natural exponential-family distributions," in *Proc. the 37th International Conference on Machine Learning*, vol. 119, 2020, pp. 9493–9503.

[4] N. Eshel, J. Tian, M. Bukwich, and N. Uchida, "Dopamine neurons share common response function for reward prediction error," *Nature Neuroscience*, vol. 19, no. 3, pp. 479–486, 2016.

[5] W. Schultz, "Dopamine reward prediction-error signalling: a two-component response," *Nature Reviews Neuroscience*, vol. 17, no. 3, pp. 183–195, 2016.

[6] W. Dabney, Z. Kurth-Nelson, N. Uchida, C. K. Starkweather, D. Hassabis, R. Munos, and M. Botvinick, "A distributional code for value in dopamine-based reinforcement learning," *Nature*, vol. 577, no. 7792, pp. 671–675, 2020.