Interpretable deep learning for deconvolution of multiplexed neural signals*

Bahareh Tolooshams[1], Sara Matias[2], 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

Summary One of the primary goals of neuroscience is to understand how features encoded in activity of single neurons support computations at the population level and ultimately the behavior of organisms. The ever-increasing amounts of neural data produced by new experimental techniques have led to the development of new unsupervised dimensionality reduction methods. These methods have leveraged advances in deep learning to build models that can capture the structure and dynamics of neural populations. Although these models can describe neural activity in complex tasks, they are based on "black-box" approaches and usually do not provide a link between neural activity and function. Here, we propose a novel method, Deconvolutional Unrolled Neural Learning (DUNL), using algorithm unrolling, an emerging technique in interpretable deep learning, to deconvolve single-trial neural activity into interpretable components. DUNL reframes dictionary learning as optimizing weights in a deep neural network to obtain a direct interpretation of network weights as parameters driving neural activity. DUNL can analyze single-trial neural data without the need for averaging over trials or animals and is applicable to naturalistic tasks with little or no trial structure. Moreover, DUNL is flexible with respect to the source signal, i.e., spike count data or on a proxy signal such as a fluorescent calcium indicator. We apply DUNL to disentangle two overlapping signals in the reward prediction errors of dopaminergic neurons in the midbrain: a first salience or surprise component and a second linked to the intrinsic relative value of the reward. We apply DUNL to the unsupervised deconvolution of these multiplexed signals and show that a) the learned parameters of DUNL can be attributed to salience and value, b) the inferred latent representations are more informative of the reward amount than neural activity estimated using ad-hoc windows and c) we can compare representations across recording modalities.

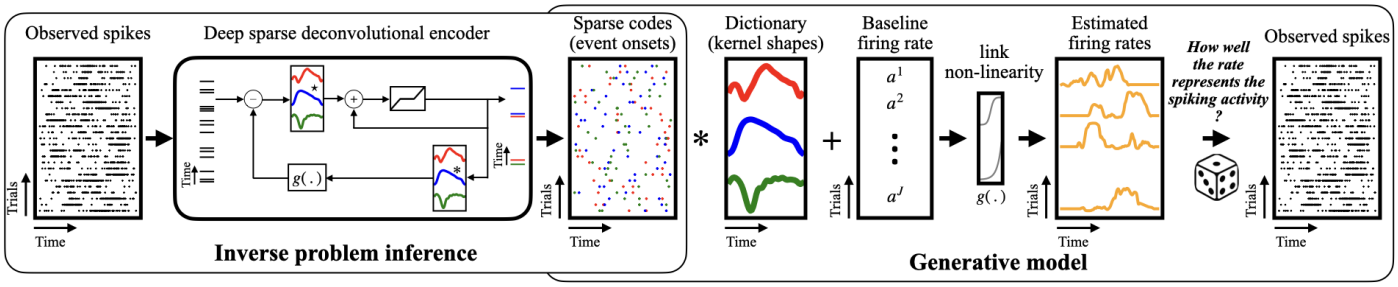

Fig. 1: Interpretable deep learning with deconvolutional unrolled neural learning (DUNL).

Additional Details For each neuron n, we impose a generative model on the neuron's activity on a trial-by-trial basis (i.e., the firing rate in the spiking setting) and model it as a function of a baseline activity $a_{n,j}$ and a set of localized kernels $\{\boldsymbol{h}_k^n\}_{k=1}^K$ characterizing the neuron's response to events that occur sparsely in time. The kernels capture characteristics that are shared among trials of a neuron or neural population. The events' onsets are modelled with a sparse code $\boldsymbol{x}_k^n$ whose amplitude encodes the strength of the contribution of the k-th kernel to the neuron's response, i.e., $\boldsymbol{\mu}^{n,j} = g(\sum_{k=1}^K \boldsymbol{h}_k^n \boldsymbol{x}_k^{n,j} + a^{n,j})$ where the distribution rules the link function g. Given $\{\boldsymbol{y}^{n,j}\}_{j=1}^J$ for each neuron n, we learn the kernels and codes by minimizing the negative log-likelihood with a sparse and structural connectivity prior on the codes, i.e.,

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

for $k = 1, \ldots, K$ where λ_k^n controls the sparsity of the codes (i.e., frequency of onsets in time) and $\boldsymbol{x}^{n,j} = [\boldsymbol{x}_1^{n,j\mathrm{T}}, \boldsymbol{x}_2^{n,j\mathrm{T}}, \ldots, \boldsymbol{x}_K^{n,j\mathrm{T}}]^\mathrm{T}$ for kernel k and neuron n. Moreover, $\boldsymbol{Q}$ is a symmetric matrix enforcing certain neural connectivity within the latent representations (e.g., discouraging simultaneous activation of two kernels). Based on algorithm unrolling [1, 2], we construct a deep neural architecture (Fig. 1) whose parameters and latent have one-to-one mapping to the above-mentioned optimization model (1) [3]. We construct a deep recurrent convolutional encoder based on proximal gradient descent to solve sparse coding in (1), i.e.,

*This work is accepted as a talk to *Computational and Systems Neuroscience, 2023*.

$\boldsymbol{x}^{n,j}_{k,r} = \boldsymbol{e}^{n,j}_{k} \cdot \mathcal{S}_{\alpha\lambda^{n,j}_{k}}\left(\boldsymbol{x}^{n,j}_{k,r-1} + \alpha \boldsymbol{h}^{n,j}_{k} \star (\boldsymbol{y}^{n,j} - g(\sum_{u=1}^{K} \boldsymbol{h}^{n}_{u} * \boldsymbol{x}^{n,j}_{u,r-1} + a^{n,j})) - 2\alpha \sum_{v=1}^{K} \boldsymbol{Q}_k \boldsymbol{x}^{n,j}_{v,r}\right)$ where $\mathcal{S}$ is shrinkage, $\star$ denotes correlation, and $\boldsymbol{e}^{n,j}_{k}$ is a known-event indicator. This inference network maps single neuron single trial observation $\boldsymbol{y}^{n,j}$, a vector in time, into an estimate of the sparse codes $\{\boldsymbol{x}^{n,j}_{k}\}_{k=1}$, encoding event onsets and their contribution to explain the data. The kernel is learned by training the network. For spiking data, the spikes at each trial are binned at B ms resolution and are modeled using natural exponential family, i.e., $\boldsymbol{y}^{n,j} \sim \text{Poisson}(\boldsymbol{\mu}^{n,j})$ and $\boldsymbol{y}^{n,j} \sim \text{Binomial}(B, \boldsymbol{\mu}^{n,j})$, where $\boldsymbol{\mu}^{n,j}$ models the mean of the distribution for neuron n at trial j. For continuous data, $\boldsymbol{y}^{n,j} \in \mathbb{R}^{T}$ is the raw measurement in time, modelled using Gaussian distribution.

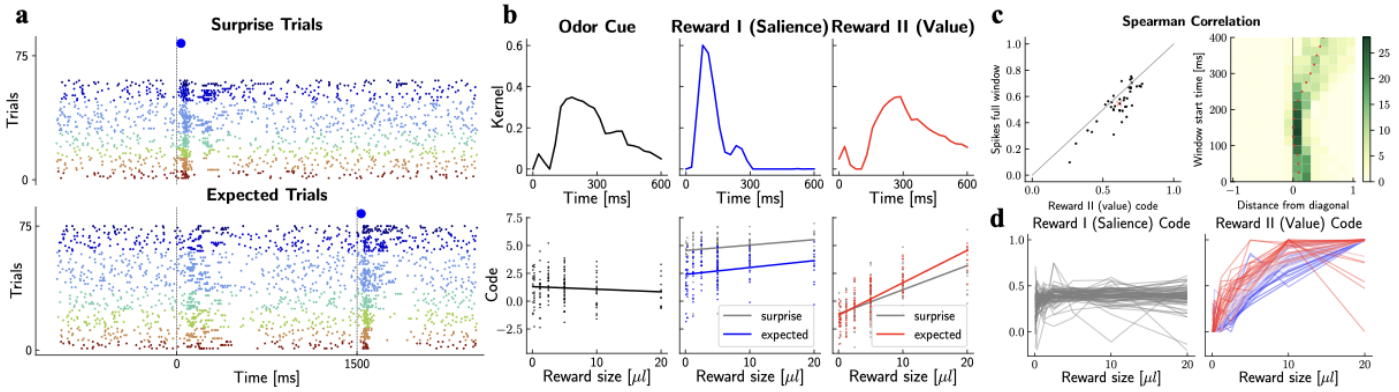

Fig. 2: Deconvolution of reward prediction error from spiking data.

Spiking data 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. 2a). Although the cue predicted the timing of the reward, it provided no information about its magnitude. We characterized the response of dopaminergic neurons to odor cue using one kernel/code (black) and model the reward response using two kernels/codes; as shown, the inferred salience (blue) is invariant to the reward amount, but the value code (red) is strongly modulated by reward amount (Fig. 2b) (no $\boldsymbol{Q}$-regularization is used for code inference in this experiment) [5]. Given this decomposition, as an alternative to spike counts from ad-hoc windows, we use the code amplitudes in single trials from the value kernel as a measure of the neurons' tuning to reward amount. We compute the Spearman rank correlation between reward amount and code amplitude or neural activity and show that the value code is more informative of the reward amount than the spike count within the full 600 ms window (Fig. 2c left, each dot represents one neuron and the red marker the average across all neurons, p=$2 \cdot 10^{-6}$, t-test). In addition, when we shrink the ad-hoc window to exclude the early activity attributed to salience, the value code stays more informative than all possible window choices (Fig. 2c right). Finally, Fig. 2d highlights that the inferred value code across dopaminergic neurons exhibit a diversity of tuning to reward amount (i.e., pessimistic vs optimistic neurons), supporting distributional reinforcement learning in dopaminergic neurons [6].

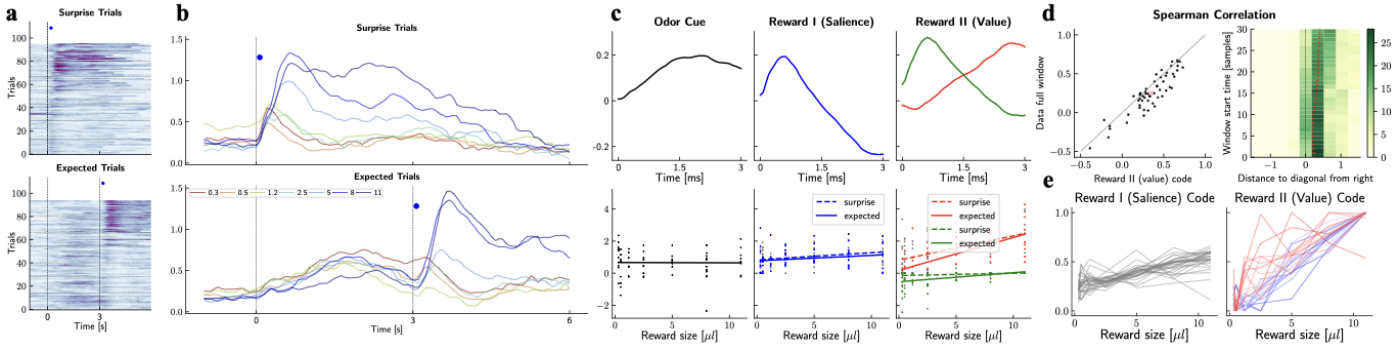

Fig. 3: Deconvolution of reward prediction error from two-photon calcium signals.

Two-photon calcium imaging We study dopaminergic neurons whose activities are recorded by two-photon imaging of *GCaMP7f* expressed in dopaminergic neurons in an adapted version of the task above; in *surprise* trials, a size-varying reward (i.e., 0.3 to 11 μl) was delivered without cue, and in *expected* trials, an odor cue preceded reward delivery by 3 s (Fig. 3a-b). We characterized (Fig. 3c) the neural responses to odor cue by one kernel/code (black), and the responses at the reward onset in both surprise and expected trials are modeled by three kernels; blue kernel resembles salience, and the other two kernels (green and red), which are discouraged to be active at the same time through $\boldsymbol{Q}$-regularization, model the value. Green and red value kernels can only

take negative and positive code, respectively; this is motivated by the calcium dynamics whose response to pauses in neural activity is slower to responses to increases in neural activity. Similar to the conclusions from spiking data, the inferred value code is highly correlated with the reward amount (Fig. 3c, bottom row) and it is more informative of reward amount than the ad-hoc windowing method (Fig. 3d with $p = 4 \cdot 10^{-8}$, t-test). Again, we observed a diversity of sensitivity to reward in the value code across neurons (Fig. 3e).

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.