Fast active set methods for online deconvolution of calcium imaging data

Johannes Friedrich, Pengcheng Zhou, Liam Paninski

Abstract

Fluorescent calcium indicators are a popular means for observing the spiking activity of large neuronal populations, but extracting the activity of each neuron from raw fluorescence calcium imaging data is a nontrivial problem. We present a fast online active set method to solve this sparse non-negative deconvolution problem. Importantly, the algorithm progresses through each time series sequentially from beginning to end, thus enabling real-time online estimation of neural activity during the imaging session. Our algorithm is a generalization of the pool adjacent violators algorithm (PAVA) for isotonic regression and inherits its linear-time computational complexity. We gain remarkable increases in processing speed: more than one order of magnitude compared to currently employed state of the art convex solvers relying on interior point methods. Our method can exploit warm starts; therefore optimizing model hyperparameters only requires a handful of passes through the data. A minor modification can further improve the quality of activity inference by imposing a constraint on the minimum spike size. The algorithm enables real-time simultaneous deconvolution of $O(10^5)$ traces of whole-brain larval zebrafish imaging data on a laptop.

Problem formulation. Calcium imaging methods enable simultaneous measurement of the activity of thousands of neighboring neurons, but come with major caveats: the slow decay of the fluorescence signal compared to the time course of the underlying neural activity, limitations in signal quality, and the large scale of the data all complicate the goal of efficiently extracting accurate estimates of neural activity from the observed video data. Further, current activity extraction methods are typically applied to imaging data after the experiment is complete. However, in many cases we would prefer to run closed-loop experiments - analyzing data on-the-fly to guide the next experimental steps or to control feedback - and this requires new methods for accurate real-time processing.

Approach. Here we address the pressing need for scalable online spike inference methods. Building on previous work, we approximate the calcium dynamics $\boldsymbol{c}$ as autoregressive (AR) process of order p and frame the estimation problem as a sparse non-negative deconvolution [1] that can be interpreted as MAP estimate of a generative model.

$$\begin{aligned} &\underset{\boldsymbol{c}}{\text{minimize}} && \tfrac{1}{2}\|\boldsymbol{c}-\boldsymbol{y}\|^2 + \lambda\|\boldsymbol{s}\|_1 \\ &\text{subject to} && \boldsymbol{s} = G\boldsymbol{c} \geq 0 \end{aligned} \tag{1}$$

$$\text{with } G = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ -\gamma_1 & 1 & 0 & \dots & 0 \\ -\gamma_2 & -\gamma_1 & 1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & -\gamma_2 & -\gamma_1 & 1 \end{pmatrix}$$

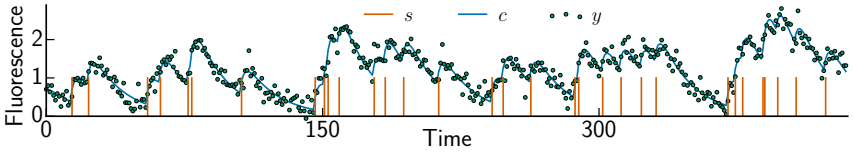

Figure 1: Generative AR model for calcium dynamics. Spike train $\boldsymbol{s}$ gets filtered to produce calcium trace $\boldsymbol{c}$; here we used an AR(2) process. Added noise yields the observed fluorescence $\boldsymbol{y}$.

This approach has already been taken up widely in various experimental labs and has thus been well validated in practice. However, current algorithms employ interior point methods to solve the ensuing optimization problem that can not handle larger data sets such as whole-brain zebrafish imaging in real time. Furthermore, these interior point methods do not run online.

We noted a close connection between the MAP problem and isotonic regression, which fits data by a monotone piecewise constant function ($c_{t+1} \geq c_t$), whereas our constraint $c_{t+1} \geq \gamma c_t$ (considering AR1 for briefness) bounds the rate of decay instead of enforcing monotonicity. A classic algorithm for isotonic regression is the pool adjacent violators algorithm (PAVA) [2], which can be understood as an online active-set optimization method. We generalized PAVA to derive an Online Active Set method to Infer Spikes (OASIS); it sweeps through the data looking for violations of the constraint $c_{t+1} \geq \gamma c_t$, cf. Fig. 2. When it finds one, it backtracks to the most recent spike and adjusts the estimate to the best possible fit with constraints, which amounts to pooling the data points where no spike happened. During the sweep adjacent pools that violate the constraints are merged. Importantly, OASIS operates in an online fashion, progressing through the fluorescence time series sequentially from beginning to end. Further, OASIS can be warm-started, which is useful in the inner loop of CNMF [3], and also when adjusting model hyperparameters. The hyperparamter λ can be set by inclusion of the residual sum of squares (RSS) as a hard constraint [3].

$$\underset{\boldsymbol{c}}{\text{minimize}} \quad \|\boldsymbol{s}\|_1 \quad \text{subject to} \quad \boldsymbol{s} = G\boldsymbol{c} \geq 0 \quad \text{and} \quad \|\boldsymbol{c}-\boldsymbol{y}\|^2 \leq \sigma^2 T. \tag{2}$$

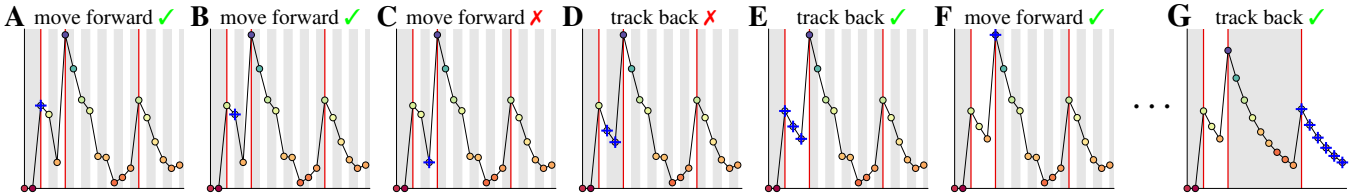

Figure 2: Illustration of OASIS for an AR(1) process (see supplementary video). Red lines depict true spike times. The shaded background shows how the time points are gathered in pools. The pool currently under consideration is indicated by the blue crosses. The algorithm proceeds moving forward **(A-C)** until the next violation occurs **(C)** and triggers backtracking and merging **(D-E)** as long as constraints are violated. When the most recent spike time has been reached **(E)** the algorithm proceeds forward again **(F)**. The process continues until the end of the series has been reached **(G)**. The solution is obtained and pools span the inter-spike-intervals.

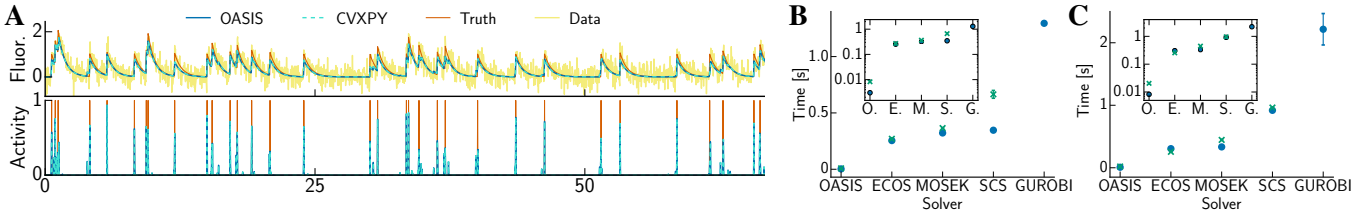

Figure 3: OASIS produces the same results as convex solvers at least an order of magnitude faster. **(A)** Raw and inferred traces. **(B)** Computation time for AR(1) data with given λ (blue circles, Eq. 1) or inference with hard noise constraint (green x, Eq. 2). GUROBI failed on the noise constrained problem. **(C)** Computation time for AR(2) data.

Results. Our suggested (dual) active set method yields the same results as other convex solvers and extracts spikes well (Fig. 3A). We compared the run time of our algorithm on a standard laptop to a variety of state of the art convex solvers that can all be conveniently called from the convex optimization toolbox CVXPY [4]: embedded conic solver (ECOS), MOSEK, splitting conic solver (SCS) and GUROBI. OASIS is about one to two magnitudes faster than any other method (Fig. 3B,C). We also ran the algorithms on longer traces up to $T = 300{,}000$ frames, confirming that OASIS scales linearly with T. Running OASIS on a total of 91,478 neurons from a whole-brain zebrafish imaging dataset from the M. Ahrens lab took 745 s for OASIS, less than the 1,500 s recording duration, and over 25,780 s for ECOS and other candidates.

We solved the noise constraint problem (Eq. 2) by increasing λ in the dual formulation until the noise constraint is tight. Thus far the AR coeffient γ was either known or estimated based on the autocorrelation in the above analyses, which often yields a crude estimate. We improved upon these results by not only optimizing the sparsity parameter λ, but also the AR coeffient γ. Further, we can include and optimize an explicit fluorescence baseline b to increase the accuracy of spike inference, always exploiting warm starts.

OASIS solves a LASSO problem resulting in soft shrinkage. We ran a slightly modified version of the algorithm that replaces the sparsity penalty by a constraint on the minimal spike size $s_{\min}$, thus performing hard thresholding and yielding sparser solutions but rendering the problem non-convex. Although we are not guaranteed to find the global minimum, we obtained improved results.

We were interested in how the method performs if backtracking is limited to just a few frames. We varied the lag in the online estimator, i.e. the number of future samples observed before assigning a spike at time zero, for different signal-to-noise ratios (SNR). For realistic SNR and sample rates, lags of merely 2-5 yielded virtually the same results as offline estimation. The exact number depends on the noise; however, the main effect of noise was to reduce the optimal performance attainable even with batch processing.

Conclusion. OASIS' advances in speed paired with its online fashion enable true real-time online spike inference during the imaging session, with the potential to significantly impact experimental paradigms.

[1] Vogelstein et al. (2010) *J Neurophysiol* 104(6):3691
[2] Ayer et al. (1955) *Ann Math Stat* 26(4):641
[3] Pnevmatikakis et al. (2016) *Neuron* 89(2):285
[4] Diamond & Boyd (2016) *J Mach Learn Res* 17(83):1