Exact Neural Spike Train Inference via Mixed Integer Quadratic Programming and Graph Neural Network

Jongmin Mun
Department of Data Sciences and Operations
jongminm@usc.edu

Contents

1 One-page Proposal **1**

2 Detailed Background **3**

2.1 Autoregressive model for calcium dynamics 3

2.2 Algorithm for AR(1) Calcium Imaging 3

2.3 Algorithm for AR(2) Calcium Imaging 4

3 Details for MIQP derivation **4**

1 One-page Proposal

Application Background. Neural spike train inference aims to recover discrete action potentials from noisy, continuous neural signals. This reconstruction is a fundamental task for brain-computer interaction (BCI) technologies, such as Neuralink. I have participated in the development of a neural probe that records the neuron signal (Park et al., 2024) and conducted neural spike train inference. I want to focus specifically on calcium imaging, a video-based recording technology capable of monitoring thousands of neurons simultaneously. After pre-processing, the signal looks like Figure 1.

Preprint.

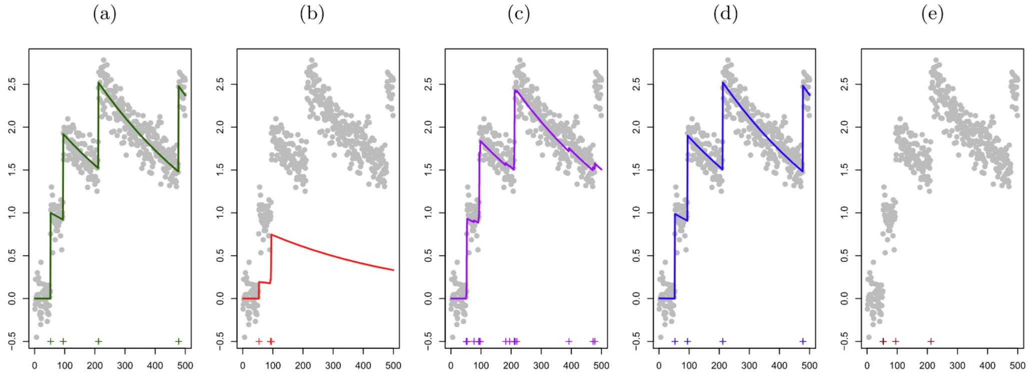

Figure 1: AR(1)-simulated calcium imaging data. (a) Ground-truth calcium concentration; (b)–(c) results obtained using ℓ_1-based methods; (d) ℓ_0 method proposed by Jewell and Witten (2018).

Optimization Perspective. Spike detection in calcium imaging is fundamentally an unsupervised peak detection problem on a one-dimensional time series. Because neuronal spiking is sparse, ℓ_0 regularization is theoretically desirable; however, the majority of the literature relies on ℓ_1 relaxation due to computational complexity. While Jewell and Witten (2018) devised an exact ℓ_0 algorithm using dynamic programming for first-order autoregressive (AR(1)) dynamics, modern high-speed calcium imaging requires AR(2) dynamics to accurately capture non-instantaneous rise times. Currently, there is no algorithm that guarantees global optimality for the ℓ_0 problem under AR(2) dynamics.

Database. Several datasets exist providing ground truth for these signals, and some methods have employed supervised deep learning to approximate spike detection (Rupprecht et al., 2021). However, there is currently no existing method utilizing Graph Neural Networks (GNNs) for "learning-to-optimize" in this domain.

Proposal. This project proposes a Mixed-Integer Quadratic Programming (MIQP) framework to solve the exact ℓ_0 optimization problem. Unlike previous heuristic methods, MIQP guarantees global optimality and allows for the seamless integration of biological constraints, such as refractory periods. The canonical problem can be formulated in two distinct ways. First, the Synthesis Formulation eliminates the latent calcium variable:

$$\begin{aligned} \underset{\mathbf{s},\mathbf{z}}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{s}\|_2^2 \\ \text{subject to} \quad & 0 \le s_t \le \mathcal{M}_U z_t, \quad t = 1, \ldots, T \\ & \sum_{t=1}^{T} z_t \le k \\ & z_t \in \{0,1\}, \qquad\quad t = 1, \ldots, T \end{aligned} \tag{1}$$

where $\mathbf{H}$ is a dense lower-triangular matrix representing the system impulse response. This effectively becomes a sparse regression problem with a specialized design matrix. Alternatively, we can retain the sparse structure of the inverse dynamics:

$$\begin{aligned} \underset{\mathbf{c},\mathbf{s},\mathbf{z}}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{y} - \mathbf{c}\|_2^2 \\ \text{subject to} \quad & 0 \le s_t \le \mathcal{M}_U z_t, \quad t = 1, \ldots, T \\ & \sum_{t=1}^{T} z_t \le k \\ & z_t \in \{0,1\}, \qquad\quad t = 1, \ldots, T \\ & \mathbf{s} = G\mathbf{c} \ge 0 \end{aligned} \tag{2}$$

where $\mathbf{G}$ is a p-banded lower-triangular matrix. (See Section 3 for matrix definitions). Without any tricks learnd in class, preliminary benchmarks on my MacBook Pro using Gurobi indicate that the solver can process a short time series ($T = 2,000$) in approximately one minute. However, practical applications require scaling to $T = 50,000$ within a similar timeframe. To bridge this gap, I propose the following:

1. **MIP for ML** (project idea 8)
   We implement an exact ℓ_0-penalized optimization framework for neural spike inference by formulating the AR(2) deconvolution problem as a Mixed-Integer Quadratic Program (MIQP). Leveraging state-of-the-art techniques from the mixed-integer optimization (MIO) literature, such as those pioneered by Bertsimas et al. (2016), we employ advanced branch-and-bound algorithms coupled with specialized problem reformulations to ensure computational efficiency.

2. **ML for MIP** (project idea 1)
   If time permits, as a secondary objective, we explore the acceleration of MIP solving tasks using **Graph Neural Networks (GNNs)**, leveraging large-scale neural signal and spike train databases Rupprecht et al. (2021).

2 Detailed Background

Neural spike train inference. Neural encoding is the scientific study of how neurons represent and transmit information through electrical activity (Kandel et al., 2021). The first step of this analysis, known as spike train inference, represents this information as a discrete, binary code (e.g., 010001...) rather than a continuous analog signal; analyzing the shape of the spike is considered a subsequent task (Wallisch et al., 2014). In this discrete framework, a '1' corresponds to the occurrence of an action potential—or 'spike'—recorded in the neural time series. Neural spike train inference is an unsupervised learning problem that learns this binary code of spike occurences from one-dimensional time series recording data.

Calcium imaging technology: scalable but noisy Calcium imaging is one of the most widely used techniques for simultaneously recording the activity of thousands of neurons. When a neuron spikes, calcium ions influx into the cell; to quantify this, the technique utilizes fluorescent indicator molecules that respond to intracellular calcium levels. Consequently, the resulting fluorescence trace serves as a continuous, albeit noisy, proxy for the underlying neural spiking activity (Ahrens et al., 2013; Dombeck et al., 2007; Prevedel et al., 2014). Although this technique offers scalability that far surpasses direct electrical recording via neural probes (which are often limited to tens of neurons (Park et al., 2024)), it introduces distinct analytical challenges: primarily, the slow decay of fluorescence relative to the rapid time course of actual spikes, and the limited signal-to-noise ratios inherent to large-scale, indirect optical recording.

2.1 Autoregressive model for calcium dynamics

Figure 1 illustrates a typical calcium recording. It is a one-dimensional, sawtooth-like time series where the slow decay of calcium transients frequently leads to signal superposition; That is, a new spike occurs before the previous one has fully dissipated.

For each time point $t = 1, \dots, T$, the calcium dynamics model is defined by three variables: the observed fluorescence y_t, the latent calcium concentration c_t, and the concentration increase caused by neuronal spikes s_t. We model y_t as a noisy linear proxy for c_t: $y_t = \beta_0 + \beta_1 c_t + \varepsilon_t$. The temporal evolution of c_t is governed by a stable autoregressive process of order p (AR(p)), typically where $p \in \{1, 2\}$ (Friedrich et al., 2017). Let $\gamma_1, \dots, \gamma_p$ represent the autoregressive coefficients, assumed known. In the absence of firing ($s_t = 0$), the concentration decays according to these AR dynamics, while a spike event ($s_t > 0$) triggers an instantaneous increase. The model is formulated as follows:

$$y_t = \beta_0 + \beta_1 c_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2), \quad t = 1, \dots, T. \tag{3}$$

$$c_t = \sum_{i=1}^{p} \gamma_i c_{t-i} + s_t, \quad t = p+1, \dots, T. \tag{4}$$

Crucially, since only the fluorescence trace y_t is observed, all other quantities are latent; consequently, the primary inferential goal is to recover the unobserved spike indicator, $\mathbb{1}(s_t > 0)$. For simplicity, we set $\beta_0 = 0$ and β_1 = 1.

2.2 Algorithm for AR(1) Calcium Imaging

An AR(1) process models the calcium response to a spike as an instantaneous increase followed by exponential decay. This provides a good approximation when the fluorescence rise time is small relative to the length of a time bin; for example, when using fluorescent indicator molecules such as GCaMP6f with a relatively slow imaging rate (Chen et al., 2013). Jewell and Witten (2018) and Jewell et al. (2020) formulated spike inference under the AR(1) model as the following ℓ_0-regularized optimization problem:

$$\min_{c_1,\dots,c_T,\, s_2,\dots,s_T} \left\{ \frac{1}{2} \sum_{t=1}^{T} (y_t - c_t)^2 + \lambda \sum_{t=2}^{T} \mathbb{1}(s_t \neq 0) \right\} \text{subject to} \quad s_t = c_t - \gamma c_{t-1} \geq 0.$$

The use of ℓ_0 regularization is natural in this setting because neural spike trains are inherently sparse: spikes occur infrequently in time, and we seek to penalize the number of spike events directly. In contrast, ℓ_1 regularization penalizes the total magnitude of the spike contributions to the calcium concentration, rather than the number of spikes themselves. As a result, it does not align as closely with the underlying nature of neural spike inference, where spike trains are more appropriately viewed as digital, binary code.

This formulation is closely related to the classical dynamic lot-sizing problem. In this analogy, c_t corresponds to the inventory level at time t, and s_t represents the order quantity in period t. The dynamics $c_t = s_t + \gamma c_{t-1}$ resemble those of a perishable good with carry-over rate γ, while the term $\frac{1}{2}(y_t - c_t)^2$ plays the role of a holding cost. Under this interpretation, the problem can be solved exactly via dynamic programming using the Wagner–Whitin algorithm. Although Jewell and Witten (2018) and Jewell et al. (2020) did not explicitly note this connection, as statisticians they reformulated the optimization problem as a changepoint detection problem and derived a dynamic programming algorithm that is closely related in structure to the Wagner–Whitin algorithm.

2.3 Algorithm for AR(2) Calcium Imaging

For fast imaging rates and slow indicators such as GCaMP6s, it is more accurate to use an AR(2) process (Friedrich et al., 2017). However, since the algorithms developed by Jewell and Witten (2018) and Jewell et al. (2020) are heavily dependent on the AR(1) structure, their dynamic programming approach cannot be directly applied to this setting.

Several methods have been developed for AR(2) calcium imaging, including the greedy ℓ_0 algorithm known as OASIS (Friedrich et al., 2017) and ℓ_1-penalized coordinate descent within the CNMF framework (Pnevmatikakis et al., 2016). Probabilistic approaches include Sequential Monte Carlo and Hidden Markov Models (Vogelstein et al., 2009), while recent advances rely on supervised deep learning trained on large ground-truth databases (Rupprecht et al., 2021).

However, none of these methods perform exact ℓ_0 optimization, nor do they offer the flexibility to incorporate specific biological constraints directly into the problem formulation. In spike train inference, ℓ_0 optimization outperforms ℓ_1 optimization in terms of inference quality, because the ℓ_1 penalty tends to overshrink the fitted estimates (Zou, 2006). This effect is evident in Figures 1(b) and 1(c), which show both under-selection and over-selection of spikes.

3 Details for MIQP derivation

As shown in equation (2.1) of Bertsimas et al. (2016), we formulate a problem similar to the ℓ_0 optimization problem as a big-M formulation by incorporating the x_t variables into a constraint. Due to nonconvexity, two formulations are not equivalent. This approach replaces the less interpretable penalty parameter λ with a more intuitive maximum number of spikes parameter, k.

$$\begin{array}{lll} \underset{\mathbf{c},\mathbf{s},\mathbf{z}}{\text{minimize}} & \frac{1}{2}\sum_{t=1}^{T}(y_t - c_t)^2 & \\ \text{subject to} & c_t = \gamma_1 c_{t-1} + \gamma_2 c_{t-2} + s_t, & t = 3, \ldots, T \\ & 0 \leq s_t \leq \mathcal{M}_U z_t, & t = 3, \ldots, T \\ & \sum_{t=3}^{T} z_t \leq k & \\ & z_t \in \{0, 1\}, & t = 3, \ldots, T \end{array} \tag{5}$$

Here, $\mathcal{M}_U$ is a constant such that if $\hat{\mathbf{s}}$ is a minimizer of the problem in equation (2.1), then $\mathcal{M}_U \geq \|\hat{\mathbf{s}}\|_\infty$. Specifically, if $z_t = 1$, then $0 \leq s_t \leq \mathcal{M}_U$, and if $z_t = 0$, then $s_t = 0$.

To improve computational efficiency, we can eliminate the latent calcium variables $\mathbf{c}$ by expressing the AR(2) dynamics as a linear system. For simplicity, assume $c_1 = c_2 = s_1 = s_2 = 0$. Let $\mathbf{H}$ be a lower-triangular Toeplitz matrix representing the discrete-time convolution kernel of the autoregressive process:

$$\mathbf{H} = \begin{bmatrix} h_0 & 0 & 0 & \ldots & 0 \\ h_1 & h_0 & 0 & \ldots & 0 \\ h_2 & h_1 & h_0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h_{T-1} & h_{T-2} & h_{T-3} & \ldots & h_0 \end{bmatrix} \tag{6}$$

The components h_t describe the impulse response of the calcium decay, where $h_0 = 1, h_1 = \gamma_1$, and $h_t = \gamma_1 h_{t-1} + \gamma_2 h_{t-2}$ for $t \geq 2$. Substituting $\mathbf{c} = \mathbf{Hs}$ into the objective function yields a Mixed-Integer Quadratic Programming (MIQP) problem. Since $\mathbf{H}^\top \mathbf{H}$ is a Gram matrix, it is positive semi-definite (PSD) by construction, ensuring the continuous relaxation of the objective is convex. The consolidated MIQP formulation is as follows:

$$
\begin{array}{lll}
\underset{\mathbf{c},\mathbf{s}}{\text{minimize}} & \frac{1}{2}\|\mathbf{y} - \mathbf{c}\|_2^2 & \\
\text{subject to} & 0 \leq s_t \leq \mathcal{M}_U z_t, & t = 1, \ldots, T \\
 & \sum_{t=1}^{T} z_t \leq k & \\
 & z_t \in \{0, 1\}, & t = 1, \ldots, T \\
 & \mathbf{s} = G\mathbf{c} \geq 0 &
\end{array}
\tag{7}
$$

In typical spike train inference, T is usually on the order of thousands or tens of thousands.

The MIO framework allows for the seamless integration of biologically meaningful constraints that are difficult, if not impossible, to incorporate within traditional convex relaxation or greedy frameworks. These include:

- **Refractory Periods:** To account for the biophysical limitations of neural firing, we can enforce a minimum time interval between consecutive events by requiring that $\sum_{j=0}^{d} z_{t+j} \leq 1$ for a window of size d.
- **Minimum Spike Magnitude:** To prevent the detection of non-physical, low-amplitude noise fluctuations—a common issue in ℓ_1 methods—we can impose the constraint $s_t \geq \delta z_t$. This ensures that any identified spike possesses a minimum physically significant magnitude δ.

Another equivalent formulation is:

$$
\begin{array}{lll}
\underset{\mathbf{c},\mathbf{s},\mathbf{z}}{\text{minimize}} & \frac{1}{2}\|\mathbf{y} - \mathbf{c}\|_2^2 & \\
\text{subject to} & 0 \leq s_t \leq \mathcal{M}_U z_t, & t = 1, \ldots, T \\
 & \sum_{t=1}^{T} z_t \leq k & \\
 & z_t \in \{0, 1\}, & t = 1, \ldots, T \\
 & \mathbf{s} = G\mathbf{c} \geq 0 &
\end{array}
\tag{8}
$$

where the ℓ_1 penalty on ŝ enforces sparsity of the neural activity. The lower triangular matrix G is defined as:

$$
G = \begin{bmatrix}
1 & 0 & \ldots & \ldots & \ldots & 0 \\
\gamma_1 & 1 & 0 & \ldots & \ldots & 0 \\
\vdots & \ddots & \ddots & \ddots & & \vdots \\
\gamma_p & \ldots & \gamma_1 & 1 & \ldots & 0 \\
0 & \ddots & & \ddots & \ddots & \vdots \\
0 & 0 & \gamma_p & \ldots & \gamma_1 & 1
\end{bmatrix}.
\tag{9}
$$

The deconvolution matrix G is banded lower triangular with bandwidth p for an AR(p) process.

References

Ahrens, M. B., Orger, M. B., Robson, D. N., Li, J. M., and Keller, P. J. (2013). Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature Methods*, 10(5):413–420.

Bertsimas, D., King, A., and Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2).

Chen, T.-W., Wardill, T. J., Sun, Y., Pulver, S. R., Renninger, S. L., Baohan, A., Schreiter, E. R., Kerr, R. A., Orger, M. B., Jayaraman, V., Looger, L. L., Svoboda, K., and Kim, D. S. (2013). Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature*, 499(7458):295–300.

Dombeck, D. A., Khabbaz, A. N., Collman, F., Adelman, T. L., and Tank, D. W. (2007). Imaging large-scale neural activity with cellular resolution in awake, mobile mice. *Neuron*, 56(1):43–57.

Friedrich, J., Zhou, P., and Paninski, L. (2017). Fast online deconvolution of calcium imaging data. *PLOS Computational Biology*, 13(3):e1005423.

Jewell, S. and Witten, D. (2018). Exact spike train inference via l0 optimization. *The Annals of Applied Statistics*, 12(4):2457–2482.

Jewell, S. W., Hocking, T. D., Fearnhead, P., and Witten, D. M. (2020). Fast nonconvex deconvolution of calcium imaging data. *Biostatistics*, 21(4):709–726.

Kandel, E. R., Koester, J. D., Mack, S. H., and Siegelbaum, S. A. (2021). *Principles of Neural Science, Sixth Edition*. McGraw Hill / Medical, New York.

Park, Y.-G., Kwon, Y. W., Koh, C. S., Kim, E., Lee, D. H., Kim, S., Mun, J., Hong, Y.-M., Lee, S., Kim, J.-Y., Lee, J.-H., Jung, H. H., Cheon, J., Chang, J. W., and Park, J.-U. (2024). In-vivo integration of soft neural probes through high-resolution printing of liquid electronics on the cranium. *Nature Communications*, 15(1):1772.

Pnevmatikakis, E. A., Soudry, D., Gao, Y., Machado, T. A., Merel, J., Pfau, D., Reardon, T., Mu, Y., Lacefield, C., Yang, W., Ahrens, M., Bruno, R., Jessell, T. M., Peterka, D. S., Yuste, R., and Paninski, L. (2016). Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron*, 89(2):285–299.

Prevedel, R., Yoon, Y.-G., Hoffmann, M., Pak, N., Wetzstein, G., Kato, S., Schrödel, T., Raskar, R., Zimmer, M., Boyden, E. S., and Vaziri, A. (2014). Simultaneous whole-animal 3D imaging of neuronal activity using light-field microscopy. *Nature Methods*, 11(7):727–730.

Rupprecht, P., Carta, S., Hoffmann, A., Echizen, M., Blot, A., Kwan, A. C., Dan, Y., Hofer, S. B., Kitamura, K., Helmchen, F., and Friedrich, R. W. (2021). A database and deep learning toolbox for noise-optimized, generalized spike inference from calcium imaging. *Nature Neuroscience*, 24(9):1324–1337.

Vogelstein, J. T., Watson, B. O., Packer, A. M., Yuste, R., Jedynak, B., and Paninski, L. (2009). Spike inference from calcium imaging using sequential monte carlo methods. *Biophysical Journal*, 97(2):636–655.

Wallisch, P., Lusignan, M. E., Benayoun, M. D., Baker, T. I., Dickey, A. S., and Hatsopoulos, N. G. (2014). *MATLAB for Neuroscientists: An Introduction to Scientific Computing in MATLAB*. Academic Press, Amsterdam.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.