Spike inference from calcium imaging using sequential Monte Carlo methods

Joshua T. Vogelstein*, Brendon O. Watson$^\#$, Adam M. Packer$^\#$,
Rafael Yuste$^{\#\S}$, Bruno Jedynak$^{||}$ and Liam Paninski¶
* Department of Neuroscience, Johns Hopkins School of Medicine
$^\#$ Department of Biological Sciences, Columbia University
§ Howard Hughes Medical Institute,
$^{||}$ Department of Applied Mathematics and Statistics, Johns Hopkins University
¶Department of Statistics and Center for Theoretical Neuroscience, Columbia University

December 2, 2008

Abstract

As recent advances in calcium sensing technologies facilitate simultaneously imaging many neurons, complementary analytical tools must also be developed to maximize the utility of this experimental paradigm. While the observations here are fluorescence images, the signals of interest — spike trains and/or time varying intracellular calcium concentrations — are hidden. Inferring these hidden signals is often problematic due to noise, nonlinearities, slow imaging rate, and unknown biophysical parameters. We overcome these difficulties by developing sequential Monte Carlo methods (using particle filters) based on biophysical models of spiking, calcium dynamics, and fluorescence. We show that even in simple cases, the particle filters outperform the optimal linear (i.e., Wiener) filter, both by obtaining better estimates and by providing errorbars. We then relax a number of our model assumptions to incorporate nonlinear saturation of the fluorescence signal, as well external stimulus and spike history dependence (e.g., refractoriness) of the spike trains. Using both simulations and in vitro fluorescence observations, we demonstrate superresolution by inferring when within a frame each spike occurs. Furthermore, the model parameters may be estimated using expectation-maximization with only a very limited amount of data (e.g. $\sim 5-10$ s or $5-40$ spikes), without the requirement of any simultaneous electrophysiology or imaging experiments.

Key words: two photon; nonlinear deconvolution; fluorescent protein; calcium dye; fluorescent protein; generalized linear model

Corresponding author: Joshua Vogelstein, joshuav@jhu.edu

Introduction

Recently, great advances in the development of calcium indicators, delivery techniques, and microscopy technologies have facilitated imaging a wide array of neural substrates (1). Calcium sensitive organic dyes (2, 3) have been targeted to populations of neurons both in vivo and in vitro using bulk loading (4, 5) and electroporation (6, 7). Similarly, viral infection, transgenics, and knock-ins have been used to genetically target neurons with fluorescent proteins (8–10) . In conjunction with the development of improved calcium indicators and loading techniques, the advent of 2-photon microscopy now enables the visualization of neurons deep within scattering tissue (11–14).

Thus, using calcium sensitive fluorescence to study neural dynamics is becoming increasingly popular in a wide variety of neural substrates, including individual spines (15–18), dendrites (19–21), boutons (22, 23), neurons (24–26), and populations of neurons (6, 27–35). While the data collected from these experiments are fluorescence movies, the signals of interest are the precise spike times and/or the intracellular calcium concentrations, $[\mathrm{Ca}^{2+}]$, of the observable neurons.

Inferring the spike trains and calcium concentrations from a fluorescence signal, however, is a difficult problem for a number of reasons. First, observations are noisy. This is a problem unlikely to be solved in the near future, as a major noise source is photon shot noise (36), which reflects the quantal nature of light emission and detection. Second, observations may have poor temporal resolution. While this problem may be partially mitigated by faster cameras and scanning systems (14, 37–39), faster imaging tends to exacerbate the noise problem, as fewer photons can be collected per image frame (36). Third, the relationship between fluorescence observations and $[\mathrm{Ca}^{2+}]$ is nonlinear, especially for fluorescent proteins (40, 41). This has placed undesirable and unnecessary restrictions on the calcium indicators used for analysis, as the standard analytical tools assume a linear relationship between $[\mathrm{Ca}^{2+}]$ and fluorescence (36, 42–44) (though see (45) for an exception). Fourth, the parameters governing the calcium and fluorescence dynamics are typically unknown a priori, and must be inferred from the data.

Nevertheless, there has been some significant recent progress. For instance, Smetters et al. (28) demonstrated reliable detection of single action potentials and spike trains by imaging bulk loaded fluorescent calcium dyes in vitro. Kerr et al. (46) — motivated by the observation that neurons in the rat motor and somatosensory cortices exhibit sparse spiking — developed a custom template-matching algorithm to detect the presence of single spikes in vivo using only fluorescence signals (and more recently further refined this approach (47)). The following year, Yaksi and Friedrich (44) — motivated by the observation that neurons in the intact zebrafish olfactory bulb tend to respond to different odors with different time-varying firing rates — developed a linear smoothing convolution kernel that effectively inferred the time varying firing rate for an explant of an intact zebrafish brain. More recently, Sato et al. (34) designed a clustering algorithm using only in vivo calcium sensitive fluorescence signals to determine whether whisker stimulation successfully induced a spike. Earlier this year, Holekamp et al. (48) applied the optimal linear filter for deconvolving a fluorescence signal from anesthetized mice. Finally, Sasaki et al. developed a nonparametric approach to infer spikes from somatic calcium fluctuations (49).

The present work differs from previous efforts in several key aspects. We start by constructing a well defined probabilistic “forward model” of the signals of interest and the imaging process. Then, utilizing a sequential Monte Carlo expectation maximization framework, we design a particle filter smoother (PFS) to optimally infer the spike times and calcium transients, given the observed fluorescence signals and the model. Even for relatively simple scenarios, the PFS outperforms optimal linear deconvolution by providing both a better inference and errorbars. The forward model may be generalized to account for a number of features present in typical data sets. Specifically, by incorporating saturation and signal dependent noise sources, we can perform inference on typical in vitro data sets. Furthermore, by allowing for intermittent observations (typical of 2-photon scanning experiments), we can perform superresolution inference, i.e., detect not just whether a spike occurs within a particular image frame, but also when within that frame the spike occurred. By also introducing stimulus and spike history dependence into the model, we can further refine our estimate. Moreover, estimating the parameters requires only a few seconds of fluorescence observations and a small number of spikes (e.g., $5 - 40$), and does not require tedious simultaneous electrophysiology and imaging experiments. We close by discussing further generalizations of the model that may be required to apply a PFS to other experimental preparations, such as in vivo imaging. All code is available from the corresponding author upon request.

Model

The data sets of interest are sequences of images corresponding to the calcium sensitive fluorescence signals of some neural activity. We aim here to construct the simplest forward model that permits one to satisfactorily infer the spike trains and calcium transients underlying these images. By forward model, we mean a complete characterization of the probability distributions governing the hidden dynamics and noisy observations, going "forward" from the spike train to the images. To infer the spike trains from the observations, we then invert our model. Below, we introduce a very simple model used to explain the mathematical formalism developed to infer the spike trains. Many of the simplifying assumptions are then relaxed in the Results section to improve our estimates when using in vitro data.

First, we assume a single-compartmental, equipotential model of the imaged neuron, over which the fluorescence signal may be spatially averaged — justified by the observation that the calcium dynamics within the neuron are relatively fast (19, 50) — yielding a one-dimensional time varying fluorescence signal for each image frame, F_t. Next, we assume that the fluorescence at any time is a noisy linear function of [Ca^{2+}] at that time:

$$F_t = \alpha[\text{Ca}^{2+}]_t + \beta + \sigma_F \varepsilon_{F,t}, \qquad \varepsilon_{F,t} \sim \mathcal{N}(\varepsilon_t; 0, 1), \tag{1}$$

where α and β set the scale and offset for the fluorescence signal, respectively, σ_F is the standard deviation of the noise, and $\varepsilon_{\cdot,t}$ denotes a standard normal Gaussian throughout this text (and $x \sim \mathcal{N}(x; \mu, \sigma^2)$ indicates that x has a Gaussian distribution with mean μ and variance σ^2).

Modeling $[\text{Ca}^{2+}]_t$ requires some additional assumptions. First, after each spike, $[\text{Ca}^{2+}]_t$ jumps instantaneously. This approximation is justified by the observation that calcium rise time is quick relative to the decay time (42, 51). Second, each jump is the same size, A; that is, for now we neglect $[\text{Ca}^{2+}]_t$ saturation effects due to channel inactivation and buffering (52). Third, $[\text{Ca}^{2+}]_t$ decays exponentially with time constant τ, to a baseline calcium concentration, $[\text{Ca}^{2+}]_b$; i.e., we lump the myriad calcium extrusion and endogenous buffering mechanisms and assume a single average time constant. Fourth, the $[\text{Ca}^{2+}]_t$ dynamics themselves have some Gaussian noise source, scaled by σ_c. Taken together, these assumptions imply the following model:

$$[\mathrm{Ca}^{2+}]_t - [\mathrm{Ca}^{2+}]_{t-1} = -\frac{\Delta}{\tau}([\mathrm{Ca}^{2+}]_{t-1} - [\mathrm{Ca}^{2+}]_b) + An_t + \sigma_c\sqrt{\Delta}\varepsilon_t, \tag{2}$$

where Δ =1 /(frame rate) is the time step size (the variance is scaled by Δ to ensure that the noise statistics are independent of the frame rate), n_t is the number of spikes that occurred in the t-th frame, and σ_c scales the noise. Note that because we have assumed here a linear observation model (i.e., Eq. 1 states that F_t is a linear function of $[\mathrm{Ca}^{2+}]_t$), our model is overparameterized. More precisely, both A and α set the scale, and $[\mathrm{Ca}^{2+}]_b$ and β set the offset. Furthermore, because the noise is not signal dependent, both σ_F^2 and α set the effective signal-to-noise ratio (SNR). Therefore, in the following, we let $\alpha = 1$, $\beta = 0$, and $\sigma_F^2 = 1$, without loss of generality (later, we deal with this overparameterization by introducing a nonlinear observation model).

To model the spike train, we let n_t be a Bernoulli (binary) random variable, which spikes in each time step with probability $p\Delta$:

$$n_t \sim \mathcal{B}(n_t; p\Delta), \tag{3}$$

where $\mathcal{B}(n_t; p\Delta)$ indicates that $n_t = 1$ with probability $p\Delta$, and $n_t = 0$ with probability $1 - p\Delta$ (where $0 < p\Delta < 1$). Eq. 3 therefore implies that spiking at time t is independent of other spikes and the intracellular calcium concentration. Fig. 1 depicts a spike train (top panel), the resulting calcium transients (second panel), and the fluorescence observations (third panel), simulated according to this model.

Mathematical Methods

Given the above model, our goal is to take the entire sequence of fluorescence observations, $F_{1:T} = [F_1, \ldots, F_T]$ (where T indexes the final observation in the sequence), and infer the underlying spike train, $n_{1:T}$. More formally, we want to find $P_{\boldsymbol{\theta}}(n_t|F_{1:T})$, the probability of the neuron spiking in each frame (which depends on the parameters, $\boldsymbol{\theta} = \{\tau, [\mathrm{Ca}^{2+}]_b, A, \sigma_c, p\}$), given all the fluorescence observations. We use a framework referred to as sequential Monte Carlo (using a PFS) to find these probabilities (53), embedded within an expectation maximization algorithm (54) to estimate the parameters. As this approach is becoming relatively common within neuroscience (55–61) — and it may be thought of as a generalization of either (i) the Baum-Welch algorithm for Hidden Markov Models (62), or (ii) the Kalman filter smoother for state-space models (63) — we relegate the details to the Appendices, and simply state the general procedure here.

We must first define a number of terms. Our model consists of a number of time-varying states, each governed by a set of parameters (which are constant). The states may be subdivided into *observation states*, denoted by $\boldsymbol{O}_t$, and *hidden states*, denoted by $\boldsymbol{H}_t$. Together, the states comprise the complete likelihood, which may be simplified, given our model assumptions, as follows (62):

$$P_{\boldsymbol{\theta}}(\boldsymbol{O}_{1:T}, \boldsymbol{H}_{1:T}) = P_{\boldsymbol{\theta}}(\boldsymbol{H}_0)\prod_{t=1}^{T} P_{\boldsymbol{\theta}}(\boldsymbol{H}_t|\boldsymbol{H}_{t-1})P_{\boldsymbol{\theta}}(\boldsymbol{O}_t|\boldsymbol{H}_t), \tag{4}$$

where $P_{\boldsymbol{\theta}}(\boldsymbol{H}_0)$ is the initial distribution distribution of hidden states, $P_{\boldsymbol{\theta}}(\boldsymbol{O}_t|\boldsymbol{H}_t)$ is the *observation distribution* and $P_{\boldsymbol{\theta}}(\boldsymbol{H}_t|\boldsymbol{H}_{t-1})$ is the *transition distribution.* For this model, the observation state is the fluorescence measurement, $\boldsymbol{O}_t = F_t$; and the hidden states are whether or not the neuron spiked, and the magnitude of the intracellular calcium concentration, $\boldsymbol{H}_t = \{n_t, [\text{Ca}^{2+}]_t\}$. We typically take the initial distribution to be baseline values, i.e., the initial calcium is $[\text{Ca}^{2+}]_b$ and initial value for the spike train is 0. The observation distribution is defined for the above model as:

$$\begin{aligned} P_{\boldsymbol{\theta}}(\boldsymbol{O}_t|\boldsymbol{H}_t) = P_{\boldsymbol{\theta}}(F_t \mid [\text{Ca}^{2+}]_t, n_t) = \\ P_{\boldsymbol{\theta}}(F_t \mid [\text{Ca}^{2+}]_t) \overset{def}{=} \mathcal{N}(F_t; \alpha[\text{Ca}^{2+}]_t + \beta, \sigma_F^2) = \mathcal{N}(F_t; [\text{Ca}^{2+}]_t, 1), \end{aligned} \tag{5}$$

which follows from Eq. 1 and the discussion following (where $\overset{def}{=}$ indicates that $P_{\boldsymbol{\theta}}(F_t| \mid [\text{Ca}^{2+}]_t)$ is defined as above). Similarly, the transition distribution for the above model is defined as:

$$\begin{aligned} P_{\boldsymbol{\theta}}(\boldsymbol{H}_t|\boldsymbol{H}_{t-1}) &= P_{\boldsymbol{\theta}}([\text{Ca}^{2+}]_t, n_t \mid [\text{Ca}^{2+}]_{t-1}, n_{t-1}) \overset{def}{=} P_{\boldsymbol{\theta}}([\text{Ca}^{2+}]_t \mid [\text{Ca}^{2+}]_{t-1}, n_t) P_{\boldsymbol{\theta}}(n_t) \\ &= \begin{cases} \mathcal{N}\big([\text{Ca}^{2+}]_t; [\text{Ca}^{2+}]_t - \Delta/\tau([\text{Ca}^{2+}]_{t-1} - [\text{Ca}^{2+}]_b) + An_t, \sigma_c^2\Delta\big)(p\Delta) & \text{if } n_t = 1 \\ \mathcal{N}\big([\text{Ca}^{2+}]_t; [\text{Ca}^{2+}]_t - \Delta/\tau([\text{Ca}^{2+}]_{t-1} - [\text{Ca}^{2+}]_b), \sigma_c^2\Delta\big)(1 - p\Delta) & \text{otherwise} \end{cases} \end{aligned} \tag{6}$$

which follows from Eqs. 2 and 3.

Now the goal is to efficiently estimate $P_{\boldsymbol{\theta}}(\boldsymbol{H}_t|\boldsymbol{O}_{1:T}) = P_{\boldsymbol{\theta}}(n_t, [\text{Ca}^{2+}]_t \mid F_{1:T})$ for all t. Estimating this distribution is problematic, because spike trains are inherently nonlinear. Therefore, linear filters (such as the Wiener filter), are inadequate, so nonlinear filters (such as particle filters), must be employed. We proceed by taking a particle filter-smoother (PFS) approach, which breaks this problem down into two recursions. In the forward recursion, we recursively estimate $P_{\boldsymbol{\theta}}(n_t, [\text{Ca}^{2+}]_t \mid F_{1:t})$, the probability of spiking and $[\text{Ca}^{2+}]$ at time t, given the fluorescence observations from time 1 up to and including t. Upon reaching time T, we recurse backward until $t = 1$, to get $P_{\boldsymbol{\theta}}(n_t, [\text{Ca}^{2+}]_t \mid F_{1:T})$, the probability of spiking and $[\text{Ca}^{2+}]$ at time t given *all* the fluorescence observations (i.e., both before and after t).

We use a particle filter to approximate the forward recursion. The key is that $P_{\boldsymbol{\theta}}(\boldsymbol{H}_t|\boldsymbol{O}_{1:t})$ may be well approximated by generating a number of weighted samples (or "particles") (53):

$$P_{\boldsymbol{\theta}}(\boldsymbol{H}_t|\boldsymbol{O}_{1:t}) \approx \sum_{i=1}^{N} w_t^{(i)} \delta\big(\boldsymbol{H}_t - \boldsymbol{H}_t^{(i)}\big), \tag{7}$$

where $w_t^{(i)}$ is the relative likelihood of the state at time t taking value $\boldsymbol{H}_t^{(i)}$, and $\delta(\cdot)$ is the Dirac delta function (i.e., $\delta(x) = 1$ when $x = 0$ and $\delta(x) = 0$ otherwise). Thus, at each time step, one samples N particles, and then computes the weight of each. It can be shown that the weights may be recursively computed by using (53):

$$w_t^{(i)} \approx \frac{P_{\boldsymbol{\theta}}\big(\boldsymbol{O}_t|\boldsymbol{H}_t^{(i)}\big) P_{\boldsymbol{\theta}}\big(\boldsymbol{H}_t^{(i)}|\boldsymbol{H}_{t-1}^{(i)}\big) w_{t-1}^{(i)}}{q\big(\boldsymbol{H}_t^{(i)}\big)}, \tag{8}$$

where $q\big(\boldsymbol{H}_t^{(i)}\big)$, the *sampling distribution* (or sampler) — which in general may depend on all the particle history and any observations (both past and future) — is chosen to make the approximation in Eq. 7 as accurate as possible. The most common choice is the "prior sampler", $q\big(\boldsymbol{H}_t^{(i)}\big) = P_{\boldsymbol{\theta}}(\boldsymbol{H}_t^{(i)}|\boldsymbol{H}_{t-1}^{(i)})$, in which we sample directly from the transition distribution. The prior sampler is very simple to use, because we know how to sample from each of the distributions comprising the transition distribution for this model (given by Eq. 6). The next most common choice is the "one-observation-ahead sampler" (53), $q\big(\boldsymbol{H}_t^{(i)}\big) = P_{\boldsymbol{\theta}}(\boldsymbol{H}_t^{(i)}|\boldsymbol{H}_{t-1}^{(i)}, \boldsymbol{O}_t)$, which may be written explicitly in terms of our model:

$$\begin{aligned} q\big(\boldsymbol{H}_t^{(i)}\big) = P_{\boldsymbol{\theta}}(\boldsymbol{H}_t^{(i)}|\boldsymbol{H}_{t-1}^{(i)}, \boldsymbol{O}_t) &= P_{\boldsymbol{\theta}}(n_t^{(i)}, [\text{Ca}^{2+}]_t^{(i)}|n_{t-1}^{(i)}, [\text{Ca}^{2+}]_{t-1}^{(i)}, F_t) \\ &= P_{\boldsymbol{\theta}}\big(F_t \mid [\text{Ca}^{2+}]_t^{(i)}\big) P_{\boldsymbol{\theta}}\big([\text{Ca}^{2+}]_t^{(i)} \mid [\text{Ca}^{2+}]_{t-1}^{(i)}, n_t^{(i)}\big) P_{\boldsymbol{\theta}}\big(n_t^{(i)}\big)/Z, \end{aligned} \tag{9}$$

where the equalities follow from our model assumptions, and Z acts as a normalizing constant that does not depend on $\{n_t, [\text{Ca}^{2+}]_t\}$. The one-observation-ahead sampler conditions directly on the next fluorescence observation, and therefore "anticipates" where to best place the next hidden samples (see Appendix A for details). In practice, the one-observation-ahead sampler is more efficient than the prior sampler, meaning that we can use fewer particles to obtain the same accuracy for the approximation in Eq. 7 (53). Thus, all the particle filters developed here implement the one-observation-ahead sampler (or a close approximation to it).

When implementing either sampler, after iterating several time steps, the weights of some of the particles approach zero, making the representation in Eq. 7 degenerate, and therefore hurting the quality of the particle approximation. To remedy this situation, whenever the approximate effective number of particles drops below some threshold (typically taken to be $N/2$), the particles may be "resampled", by sampling (with replacement) from the population of particles. The probability of resampling each particle is related to its weight (64) (see Appendix A for details of how to weight and resample from this distribution).

One recursively repeats these three steps (sampling, computing weights, and resampling if necessary) for each time step, starting at $t = 1$, and continuing through $t = T$, thus completing the forward recursion (i.e., the particle filter), and yielding an approximation to $P_{\boldsymbol{\theta}}(\boldsymbol{H}_t|\boldsymbol{O}_{1:t})$ for each time step. Upon reaching $t = T$, one initializes $P_{\boldsymbol{\theta}}(\boldsymbol{H}_T^{(i)}|\boldsymbol{O}_{1:T}) = w_T^{(i)}$, and then uses the following backward recursion, going from $t = T$ to $t = 1$, to approximate $P_{\boldsymbol{\theta}}(\boldsymbol{H}_t|\boldsymbol{O}_{1:T})$ for each time step:

$$P_{\boldsymbol{\theta}}(\boldsymbol{H}_t^{(i)}, \boldsymbol{H}_{t-1}^{(j)}|\boldsymbol{O}_{1:T}) = P_{\boldsymbol{\theta}}\big(\boldsymbol{H}_t^{(i)}|\boldsymbol{O}_{1:T}\big) \frac{P_{\boldsymbol{\theta}}\big(\boldsymbol{H}_t^{(i)}|\boldsymbol{H}_{t-1}^{(j)}\big) w_{t-1}^{(j)}}{\sum_j P_{\boldsymbol{\theta}}\big(\boldsymbol{H}_t^{(i)}|\boldsymbol{H}_{t-1}^{(j)}\big) w_{t-1}^{(j)}} \tag{10a}$$

$$P_{\boldsymbol{\theta}}(\boldsymbol{H}_{t-1}^{(j)}|\boldsymbol{O}_{1:T}) = \sum_{i=1}^{N} P_{\boldsymbol{\theta}}(\boldsymbol{H}_t^{(i)}, \boldsymbol{H}_{t-1}^{(j)}|\boldsymbol{O}_{1:T}). \tag{10b}$$

This backward recursion is often referred to as a "particle smoother", and comprises the backward component of our PFS approach. Thus, our PFS provides the distributions in Eq. 10 (for a particular model). For instance, the *linear observation* particle filter provides the distributions in Eq. 10, when modeling the spiking, calcium, and fluorescence dynamics according to Eqs. 1 – 3 (cf. Fig. 1, bottom panel). Given the distributions in Eq. 10, we can perform various inferences. For example, the expected number of spikes at each time step, given all the observations, may be computed by:

$$E[n_t|F_{1:T}] = \sum_{i=1}^{N} n_t^{(i)} P_{\boldsymbol{\theta}}(n_t^{(i)}|F_{1:T}) = \sum_{i=1}^{N} n_t^{(i)} P_{\boldsymbol{\theta}}(\boldsymbol{H}_t^{(i)}|\boldsymbol{O}_{1:T}). \tag{11}$$

Other quantities of interest (such as the posterior variance, median, etc.) may be computed in a similar fashion, since we have computed the full posterior distribution, $P_{\boldsymbol{\theta}}(n_t|F_{1:T})$ (which, hereafter, is referred to as the posterior mean of the spike train, or simply inferred spike train). All these computations require reasonable estimates of the parameters. By using an expectation-maximization approach (54), we can iterate inferring the distributions of interest (e.g., $P_{\boldsymbol{\theta}}(n_t|F_{1:T})$), and learning the parameters. More precisely, we optimize the following expected loglikelihood (65):

$$\begin{aligned}\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{t=1}^{T} \Bigg(&\sum_{i,j=1}^{N} P_{\boldsymbol{\theta}'}\big(\boldsymbol{H}_t^{(i)}, \boldsymbol{H}_{t-1}^{(j)}|\boldsymbol{O}_{1:T}\big) \times \ln P_{\boldsymbol{\theta}}\big(\boldsymbol{H}_t^{(i)}|\boldsymbol{H}_{t-1}^{(j)}\big) \\ &+ \sum_{i=1}^{N} P_{\boldsymbol{\theta}'}\big(\boldsymbol{H}_t^{(i)}|\boldsymbol{O}_{1:T}\big) \times \ln P_{\boldsymbol{\theta}}\big(\boldsymbol{O}_t|\boldsymbol{H}_t^{(i)}\big)\Bigg),\end{aligned} \tag{12}$$

where $\boldsymbol{\theta}'$ is the estimate of the parameters from the previous iteration, i.e. those used to obtain the distributions in Eq. 10, which may be thought of as weights on the transition and observation log-densities. Importantly, the above loglikelihood for this model was constructed to ensure that all the parameters may be quickly estimated using standard gradient ascent techniques. Details may be found in Appendix B.

Experimental Methods

Slice Preparation and Imaging All animal handling and experimentation was done according to the National Institutes of Health and local Institutional Animal Care and Use Committee guidelines. Somatosensory thalamocortical slices 400 μm thick were prepared from C57BL/6 mice at age P14 as described (66). Neurons were filled with 50 μM Fura 2 pentapotassium salt (Invitrogen, Carlsbad, CA) through the recording pipette. Pipette solution contained 130 K-methylsulfate, 2 $MgCl_2$, 0.6 EGTA, 10 HEPES, 4 ATP-Mg, and 0.3 GTP-Tris, pH 7.2 (295 mOsm). After cells were fully loaded with dye, imaging was done by using a modified BX50-WI upright confocal microscope (Olympus, Melville, NY). Image acquisition was performed with the C9100-12 CCD camera from Hamamatsu Photonics (Shizuoka, Japan) with arclamp illumination at 385 nm and $510/60$ nm collection filters (Chroma, Fuerstenfeldbruck, Germany). Images were saved and analyzed using custom software written in Matlab (Mathworks, Natick, MA).

Electrophysiology All recordings were made using the Multiclamp 700B amplifier (Molecular Probes, Sunnyvale, CA), digitized with National Instruments 6259 multichannel cards and recorded using custom software written using the LabView platform (National Instruments, Austin, TX) . Waveforms were generated using Matlab and were given as current commands to the amplifier using the LabView and National Instruments system. The shape of the waveforms mimicked excitatory (inhibitory) synaptic inputs, with a maximal amplitude of $+70$ pA (-70 pA).

Results

Main Result The main result of this work is depicted in Fig. 1, which shows a spike train, calcium concentration, and resulting fluorescence observations (first through third panels, respectively) when simulated according to the simple *linear observation* model, Eqs. 1 – 3 (where linear observation refers to the relationship between $[\mathrm{Ca}^{2+}]_t$ and F_t). For this model, we developed a linear observation particle filter smoother (PFS), to perform optimal inference of the spike train (see Appendix A.1 for details). While the optimal linear deconvolution (i.e., the Wiener filter; see (48) for a detailed discussion on using the Wiener filter to infer spikes from calcium imaging) performs reasonably well (fourth panel), even in this relatively simple example, the linear observation PFS (bottom panel) provides several advantages. First, the spike train inferred by the linear observation PFS (dark blue, bottom panel) is a better estimate of the actual spike train than the estimate using the Wiener filter (red and blue, fourth panel). This follows because the Wiener filter assumes that the spike train has a Gaussian distribution, and therefore admits both partial and negative spikes, neither of which are possible in our model. Second, the PFS provides not only the probability of a spike occurring in each time bin, but also the entire distribution (from which we may compute errorbars; light blue in bottom panel). An even more fundamental advantage of the PFS framework is its generalizability. Below, we address a number of important generalizations to the model, each of which requires just a minor modification of the dynamics equations, sampling distribution, and particle filter (but the smoother remains the same). We then apply each generalization to in vitro data to demonstrate its utility.

Saturation The relationship between the fluorescence signal and $[\mathrm{Ca}^{2+}]_t$ is often characterized by a nonlinear saturating function, $S([\mathrm{Ca}^{2+}]_t)$:

$$F_t = \alpha S([\mathrm{Ca}^{2+}]_t) + \beta + \eta_t. \tag{13}$$

The above equation states that at any time, the *expected value* of fluorescence is a nonlinear saturating function of the calcium signal. The gain (or slope), α, accounts for all the factors contributing to signal amplification, including the number of fluorophores in the neuron, the brightness of each fluorophore, the gain of the image acquisition system, etc. The offset, β, accounts for any factor leading to a constant background signal, such as baseline fluorescence. The nonlinear saturation function, $S([\mathrm{Ca}^{2+}]_t)$, is often taken to be the Hill equation, i.e. $S(x) = x^n/(x^n + k_d)$ (42). The variance, η_t, may be generalized similarly. Assuming the primary noise source is photon shot noise, it would be appropriate to model noise as a Poisson process, which could be well approximated by a Gaussian distribution for large photon counts (36):

$$\eta_t = \big(S([\mathrm{Ca}^{2+}]_t) + \sigma_F\big)\varepsilon_t, \tag{14}$$

where σ_F scales the signal-to-noise ratio (SNR). Note that there is no scale term for the saturated fluorescence component of the noise, as it would not be identifiable (α, β, and σ_F could be normalized by it without loss of generality). These assumptions change the observation distribution from Eq. 5 to:

$$P_{\boldsymbol{\theta}}(\boldsymbol{O}_t|\boldsymbol{H}_t) = P_{\boldsymbol{\theta}}^{NL}(F_t \mid [\text{Ca}^{2+}]_t) \overset{def}{=} \mathcal{N}\big(F_t; \alpha S([\text{Ca}^{2+}]_t) + \beta, (S([\text{Ca}^{2+}]_t) + \sigma_F)^2\big). \tag{15}$$

To perform optimal inference on this model (i.e., Eqs. 2, 3, 13, and 14), we construct a *nonlinear observation* PFS (where nonlinear observation refers to the relationship between F_t and $[\text{Ca}^{2+}]_t$ given by Eq. 15). The nonlinear observation PFS is different from the linear observation PFS because the observation distributions for which the two filters were designed differ, thus the one-observation-ahead sampler ($q\big(\boldsymbol{H}_t^{(i)}\big) = P_{\boldsymbol{\theta}}(\boldsymbol{H}_t^{(i)}|\boldsymbol{H}_{t-1}^{(i)}, \boldsymbol{O}_t)$) changes (See Appendix A.2 for details).

Fig. 2 shows an example of data simulated using the above model (Eqs. 2, 3, 13, and 14; top three panels). Two important differences between this model and the linear model are apparent. First, the nonlinear saturating function, $S([\text{Ca}^{2+}]_t)$, causes the fluorescence to decay more slowly than the calcium. Thus, if one were to simply deconvolve the spike trains from the raw fluorescence observations, the estimate of the spike train $(n_{1:T})$ and time constant (τ) would be biased. Second, as $[\text{Ca}^{2+}]$ accumulates, the fluorescence transients due to a spike become smaller. This reduces the effective SNR, obfuscating estimating the jump size, A. The Wiener filter (fourth panel), which cannot incorporate a nonlinearity, performs less well in this scenario than in the linear scenario. This may be evident from the observation that peaks in the Wiener filter output become smaller and closer to the noise when the signal approaches saturation. The nonlinear observation PFS, however, explicitly models this nonlinearity, and therefore can infer spikes very accurately even in the saturating regime (fifth panel). Furthermore, using the nonlinear observation PFS, we can reconstruct the unsaturated $[\text{Ca}^{2+}]_t$ (bottom panel) in addition to the spike train (when assuming Eq. 15 accurately describes the relationship between calcium and fluorescence). This is an absolute estimate of $[\text{Ca}^{2+}]_t$, meaning that we infer the baseline calcium concentration and jump size in real units (as opposed to only relative units), which follows because relative changes in fluorescence correspond with absolute changes in the *unsaturated* calcium concentration, due to the assumed nonlinear relationship between F_t and $[\text{Ca}^{2+}]_t$.

Fig. 3 shows an example of saturating fluorescence observations recorded in vitro (top panel). Within a burst, later spikes cause fluorescent transients that are smaller than the first few spikes. This is evident from the Wiener filter, in which the inferred spike size becomes much smaller in large bursts (second panel). The nonlinear observation PFS, however, accurately infers exactly one spike for each frame in which a spike occurred (third panel). Furthermore, we infer the underlying and non-saturating calcium transients (bottom panel), which is not possible using linear methods. Fig. 4 shows another example of a spike train recorded in vitro, but with far noisier observations and a more “naturalistic” spike train. As in Fig. 3, even though the effective SNR of the Wiener filter output deteriorates as the fluorescence signal saturates, the nonlinear observation PFS can accurately infer precise spike times.

Superresolution Technological limitations often impose an undesirable upper bound on the imaging frame rate. In this context, superresolution denotes the ability to infer spike trains with more precision than the frame rate. Our assumptions may be generalized for superresolution inference by modifying the observation model. First, we reduce the time step size by a factor, d, such that $\Delta = 1/(d \times \text{frame rate})$. Now we have two cases for the observation distribution: the case described by Eq. 15 (which now occurs every d time steps), and the "null" case, where no observation occurs (and therefore, $P_{\boldsymbol{\theta}}(\boldsymbol{O}_t|\boldsymbol{H}_t) = 1$). To perform optimal inference given this more sophisticated observation distribution, we develop a *superresolution* PFS (see Appendix A.3 for details). Fig. 5 shows how the superresolution PFS inference precision scales with both imaging frame rate and observation noise. Importantly, the probability of spiking in each time step within an image is not uniform, but rather, tends to be higher around the actual spike time. As the noise is increased, the probabilities further spread and flatten, but still yield an accurate estimate of the total number of spikes per frame (assuming one tends to collect a large enough number of photons per pixel to be detected by the imaging system).

One interesting result of this analysis is that imaging faster, while increasing noise and *decreasing* SNR per frame (36), can actually increase fidelity (i.e., effective SNR). This may be seen by comparing panels arranged diagonally ascending to the right, which show how the inference performs upon increasing frame rate and noise proportionally. Although the SNR per frame decreases, because more information is available about the decay, superior inference precision may be achieved. This suggests that given the option, it is always advantageous to image as quickly as possible, even at the expense of reduced SNR per frame.

Spike History and Stimulus Dependence So far, we have assumed that our neuron generates spikes independent of both external stimuli and its own spike history (cf. Eq. 3). These two inputs (stimuli and spike histories) may be incorporated into this framework by replacing p of Eq. 3 with a Generalized Linear Model (GLM) (67). GLMs have recently been used extensively to model spike trains from a variety of different preparations and modalities (see, for example, (68)). While many GLMs could be applied here, to fit within the sequential Monte Carlo expectation maximization framework, we require that: (i) the loglikelihood is concave in the parameters of the GLM, and (ii) the dynamics are Markovian. To satisfy our first constraint (concavity), we propose to allow the probability of spiking, p_t, to be a time-varying nonlinear function of the input to the neuron, y_t:

$$p_t = 1 - e^{f(y_t)\Delta}, \tag{16}$$

where $f(\cdot)$ is some convex function (see (69) for more details on Eq. 16). In general, the input to the neuron, y_t, may be subdivided into a multidimensional stimulus, $\boldsymbol{x}_t$, and a set of spike history terms, $\boldsymbol{h}_t = \{h_{1,t}, \ldots, h_{L,t}\}$, yielding:

$$y_t = \boldsymbol{k}'\boldsymbol{x}_t + \boldsymbol{w}'\boldsymbol{h}_t, \tag{17}$$

where $\boldsymbol{k}$ is a linear filter operating on the stimulus (which is closely related to the spike-triggered-average of the neuron (70)), $\boldsymbol{w}$ weights the spike history terms (71), and $'$ denotes the transpose operation. To satisfy the second constraint above, the spike history terms must be Markovian. We therefore elect to use a set of exponentially decaying terms, each with a unique time constant, which is sufficiently general to account for most spike history effects, including refractoriness, burstiness, facilitation, adaptation, and oscillations (72):

$$h_{l,t} - h_{l,t-1} = -\frac{\Delta}{\tau_{h_l}} h_{l,t-1} + n_{t-1} + \sigma_{h_l}\sqrt{\Delta}\varepsilon_{l,t}, \tag{18}$$

which implies that after each spike, each spike history term jumps, and then decays back to zero with time constant τ_{h_l} (and each process has noise with variance $\sigma^2_{h_l}\Delta$). To optimally infer spikes given this more sophisticated model (i.e., Eqs. 2, 13, 14, and 16 – 18), we modify our superresolution PFS to incorporate the above GLM, yielding a GLM PFS (see Appendix A.4 for details).

Fig. 6 shows a simulation using a model that incorporates saturation and signal dependent noise, as well as stimulus and spike history dependent spiking, with an unsatisfactorily slow frame rate (top six panels). Although the superresolution PFS accurately infers in which frame spikes occur (seventh panel), its superresolution abilities are limited due to saturation and low SNR. By contrast, the GLM PFS accurately infers spike times with superresolution precision by utilizing the input and spike history dependence (bottom panel). Note that even when multiple spikes occur within a single image frame, the GLM PFS correctly infers the number of spikes, and provides a good estimate for the precise timing of each spike (see image frames between 0.5 and 1 sec).

Fig. 7 uses in vitro data to compare the Wiener filter, superresolution PFS, and GLM PFS. Here, a neuron under patch clamp (current clamp mode) was stimulated with a time-varying current (top panel). The exact spike times were recorded electrophysiologically (second panel), while simultaneously imaging the fluorescence signal (third panel). The Wiener filter (fourth panel) generates "bumps" near the frames in which spikes arrived, but generally fails to identify individual spike times.

The superresolution PFS succeeds in identifying the spikes, but with limited temporal resolution (fifth panel). By including stimulus information and spike history dependence, the GLM PFS further refines the temporal estimates beyond that of our sampling interval. From this data set, we could achieve a temporal precision of approximately 25 msec, even though observations were only obtained once per 100 msec (bottom panel).

Learning the parameters All of the above results depend on our ability to estimate the parameters. Although the above models have several hidden states (i.e., n_t, $[\text{Ca}^{2+}]_t$, and potentially $\boldsymbol{h}_t$) and even more parameters, learning the parameters is still fast. The models were constructed to ensure that the loglikelihood functions were concave jointly in all the parameters, facilitating using standard gradient ascent techniques to find their maximum likelihood estimators. Table 1 shows the parameter estimates using only noisy fluorescence observations including very few spikes. As the number of spikes underlying the observations increases, our parameter estimates improve both in accuracy and precision. This suggests that upon learning the parameters from the in vitro data, our absolute calcium concentration estimates reflect the true values (which could be confirmed using ratiometric dyes or calibration experiments (42)). Importantly, these computations may be performed relatively quickly. More specifically, the number of computations scales linearly with T and quadratically with N (due to Eq. 10). In practice, for all the above examples (both simulated and real), a single iteration ran in approximately real time on a standard laptop computer (i.e., 5 s of data required 5 s of computation; requiring only ~ 100 particles to obtain sufficiently accurate approximations for all examples). Moreover, parameters typically converged in < 50 iterations, so inference on data collected during the day can be completed overnight.

Discussion

We started by constructing a very simple model relating spiking, calcium, and fluorescence observations, and showed that our linear observation PFS both improves inference accuracy over the optimal linear method, and provides errorbars (cf. Fig. 1). Then, we relaxed a number of the assumptions, to show how our method can be generalized. First, we postulated a more realistic observation model, by incorporating both saturation and signal dependent noise, and showed that a nonlinear observation PFS outperforms the Wiener filter (cf. simulated data in Fig. 2 and real data in Figs. 3 and 4). Then, we demonstrated superresolution capabilities, by inferring when within an image frame spikes occur, using our superresolution PFS (cf. Fig. 5). By incorporating a GLM to govern spiking activity in our model, we could also account for spike history and stimulus dependencies, utilizing our GLM PFS (cf. Fig. 6), and further enhance the inference precision using in vitro data (cf. Fig. 7). These results all depend on an ability to accurately estimate the model parameters, even when given only short ($\sim 5-10$ s and $5-50$ spikes) and noisy fluorescence observations. Importantly, estimating these parameters did not require any additional simultaneous electrophysiology or imaging experiments; rather, all inferences and parameter estimations were performed using only the fluorescence observations. Simultaneous imaging and electrophysiological experiments, however, for confirmation, would be desirable in novel preparations. Finally, as each iteration may be performed in real time, and the parameters converged in < 50 iterations, this analysis does not impose severe computational restrictions, and may be performed between experimental sessions, for instance (though see (73) for a complementary "online" algorithm). These examples demonstrate the power of the proposed particle filtering methods.

While the above generalizations were sufficient to infer the spikes in this data set, further generalizations may be necessary for other preparations. Perhaps most importantly, we ignored several prominent noise sources. For instance, the point spread function of a 2-photon microscope in vivo often spans ~ 10 μm in the axial dimension, which is sufficiently large to capture activity in the surrounding neuropil (74). Furthermore, tissue movement is often a problem, especially for awake and/or behaving preparations (75). While both axial resolution and movement artifacts are currently being addressed experimentally, we could incorporate these additional noise sources into our model as well (by modifying our noise assumptions, Eq. 14).

The dynamics of each of the states could also be generalized in a number of ways. First, bleaching is often a problem, especially for in vivo settings. This could easily be incorporated in our framework by allowing the observation parameters, $\{\alpha, \beta, \sigma_F\}$, to decay with time constants that could be inferred directly. Second, while we implicitly assumed that fluorescence achieves steady-state instantaneously, we could instead include more realistic fluorescence dynamics, which may be necessary for slower indicators, such as the genetically encoded probes (41). Third, the proposed model for calcium dynamics, Eq. 2, could be generalized in a number of ways. For instance, we could (i) enable the transient influx in $[\mathrm{Ca}^{2+}]_t$ due to a spike be variable, or (ii) incorporate additional time constants, to facilitate a non-instantaneous rise time, adaptation, extrusion, or other more sophisticated calcium dynamics (76).

Finally, one of the major goals of large-scale calcium fluorescence imaging experiments is to understand the dynamics of neural populations (29). The proposed methodology could readily be implemented while imaging a heterogeneous population of neurons, by estimating the observation, calcium, and spiking dynamics parameters independently for each observable neuron. Alternately, an important aspect of our proposed model is the spike history terms, which here only cause effects in a single neuron. This model may easily be generalized to include not only the “self-coupling” spike history effects discussed here (cf. Fig. 6), but also “cross-coupling” terms, which model the effects that one neuron’s activity has upon other “target” neurons in the observed population (70, 71, 77). Then, estimating these interneuronal spike history weights $\boldsymbol{\omega}$ corresponds to estimating a functional connectivity matrix of the network. We will address the practical limitations of inference quality and parameter estimation accuracy for large populations of neurons in future work.