Fast Nonnegative Deconvolution for Spike Train Inference From PopulationCalcium Imaging

Joshua T. Vogelstein,1 Adam M. Packer,2,3 Timothy A. Machado,2,3 Tanya Sippy,2,3 Baktash Babadi,4 Rafael Yuste,2,3 and Liam Paninski4,5

1Department of Neuroscience, Johns Hopkins University, Baltimore, Maryland; 2Howard Hughes Medical Institute, Chevy Chase, Maryland; 3Department of Biological Sciences, 4Center for Theoretical Neuroscience, and 5Department of Statistics, Columbia University, New York, New York

Submitted 9 December 2009; accepted in final form 3 June 2010

Vogelstein JT, Packer AM, Machado TA, Sippy T, Babadi B, Vuste R, Paninski L. Fast nonnegative deconvolution for spike train inference from population calcium imaging. J Neurophysiol 104: 3691–3704, 2010. First published June 16, 2010; doi:10.1152/jn.01073.2009. Fluorescent calcium indicators are becoming increasingly popular as a means for observing the spiking activity of large neuronal populations. Unfortunately, extracting the spike train of each neuron from a raw fluorescence movie is a nontrivial problem. This work presents a fast nonnegative deconvolution filter to infer the approximately most likely spike train of each neuron, given the fluorescence observations. This algorithm outperforms optimal linear deconvolution (Wiener filtering) on both simulated and biological data. The performance gains come from restricting the inferred spike trains to be positive (using an interior-point method), unlike the Wiener filter. The algorithm runs in linear time, and is fast enough that even when simultaneously imaging >100 neurons, inference can be performed on the set of all observed traces faster than real time. Performing optimal spatial filtering on the images further refines the inferred spike train estimates. Importantly, all the parameters required to perform the inference can be estimated using only the fluorescence data, obviating the need to perform joint electrophysiological and imaging calibration experiments.

INTRODUCTION

Simultaneously imaging large populations of neurons using calcium sensors is becoming increasingly popular (Yuste and Katz 1991; Yuste and Konnerth 2005), both in vitro (Ikegaya et al. 2004; Smetters et al. 1999) and in vivo (Göbel and Helmchen 2007; Luo et al. 2008; Nagayama et al. 2007), and will likely continue to improve as the signal-to-noise ratio (SNR) of genetic sensors continues to improve (Garaschuk et al. 2007; Mank et al. 2008; Wallace et al. 2008). Whereas the data from these experiments are movies of time-varying fluorescence intensities, the desired signal consists of spike trains of the observable neurons. Unfortunately, finding the most likely spike train is a challenging computational task, due to limitations on the SNR and temporal resolution, unknown parameters, and analytical intractability.

A number of groups have therefore proposed algorithms to infer spike trains from calcium fluorescence data using very different approaches. Early approaches simply thresholded $dF/F$ [typically defined as $(F - F_b)/F_b$, where $F_b$ is baseline fluorescence; e.g., Mao et al. 2001; Schwartz et al. 1998] to obtain “event onset times.” More recently, Greenberg et al. (2008) developed a dynamic programming algorithm to identify individual spikes. Holekamp et al. (2008) then applied an optimal linear deconvolution (i.e., the Wiener filter) to the fluorescence data. This approach is natural from a signal processing standpoint, but does not realize the knowledge that spikes are always positive. Sasaki et al. (2008) proposed using machine learning techniques to build a nonlinear supervised classifier, requiring many hundreds of examples of joint electrophysiological and imaging data to “train” the algorithm to learn what effect spikes have on fluorescence. Vogelstein and colleagues (2009) proposed a biophysical model-based sequential Monte Carlo (SMC) method to efficiently estimate the probability of a spike in each image frame, given the entire fluorescence time series. Although effective, that approach is not suitable for on-line analyses of populations of neurons because the computations run in about real time per neuron (i.e., analyzing 1 min of data requires about 1 min of computational time on a standard laptop computer).

In the present work, a simple model is proposed relating spiking activity to fluorescence traces. Unfortunately, inferring the most likely spike train, given this model, is computationally intractable. Making some reasonable approximations leads to an algorithm that infers the approximately most likely spike train, given the fluorescence data. This algorithm has a few particularly noteworthy features, relative to other approaches. First, spikes are assumed to be positive. This assumption often improves filtering results when the underlying signal has this property (Cunningham et al. 2008; Huys et al. 2006; Lee and Seung 1999; Lin et al. 2004; Markham and Conchello 1999; O’Grady and Pearlmutter 2006; Paninski et al. 2009; Portugal et al. 1994). Second, the algorithm is fast: it can process a calcium trace from 50,000 images in about 1 s on a standard laptop computer. In fact, filtering the signals for an entire population of $>100$ neurons runs faster than real time. This speed facilitates using this filter on-line, as observations are being collected. In addition to these two features, the model may be generalized in a number of ways, including incorporating spatial filtering of the raw movie, which can improve effective SNR. The utility of the proposed filter is demonstrated on several biological data sets, suggesting that this algorithm is a powerful and robust tool for on-line spike train inference. The code (which is a simple Matlab script) is available for free download from http://www.optophysiology.org.

METHODS

Data-driven generative model

Figure 1 shows data from a typical in vitro epifluorescence experiment (for data collection details see Experimental methods...
The top panel shows the mean frame of this movie, including four neurons, three of which are patched. To build the model, the pixels within a region of interest (ROI) are selected (white circle). Given the ROI, all the pixel intensities of each frame can be averaged, to get a one-dimensional fluorescence time series, as shown in the bottom left panel (black line). By patching onto this neuron, the spike train can also be directly observed (black bars; bottom left). Previous work suggests that this fluorescence signal might be well characterized by convolving the spike train with an exponential and adding noise (Yuste and Konnerth 2005). This model is confirmed by convolving the true spike train with an exponential (gray line; bottom left) and then looking at the distribution of the residuals. The bottom right panel shows a histogram of the residuals (dashed line) and the best-fit Gaussian distribution (solid line).

The preceding observations may be formalized as follows. Assume there is a one-dimensional fluorescence trace \( F \) from a neuron [throughout this text \( X \) indicates the vector \( (X_1, \ldots, X_T) \), where \( T \) is the index of the final frame]. At time \( t \), the fluorescence measurement \( F_t \) is a linear-Gaussian function of the intracellular calcium concentration at that time \( [Ca^{2+}]_n \):

\[
F_t = \alpha [Ca^{2+}]_n + \beta + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2). \tag{1}
\]

The parameter \( \alpha \) absorbs all experimental variables influencing the scale of the signal, including the number of sensors within the cell, photons per calcium ion, amplification of the imaging system, and so on. Similarly, the offset \( \beta \) absorbs, for example, the baseline calcium concentration of the cell, background fluorescence of the fluorophore, and imaging system offset. The noise at each time \( \varepsilon_t \) is independently and identically distributed according to a normal distribution with zero mean and \( \sigma^2 \) variance, as indicated by the notation \( \varepsilon_t \sim N(0, 1) \). This noise results from calcium fluctuations independent of spiking activity, fluorescence fluctuations independent of calcium, and other sources of imaging noise.

Then, assuming that the intracellular calcium concentration \( [Ca^{2+}]_n \), jumps by \( \Delta \mu \text{M} \) after each spike and subsequently decays back down to baseline \( C_B \mu \text{M} \), with time constant \( \tau \) s, one can write:

\[
[Ca^{2+}]_{n+1} = (1 - \Delta/\tau)[Ca^{2+}]_n + (\Delta/\tau)C_B + \Delta n_t. \tag{2}
\]

where \( \Delta \) is the time step size—which is the frame duration, or \( 1/\text{frame rate} \)—and \( n_t \) indicates the number of times the neuron spiked in frame \( t \). Note that because \([Ca^{2+}]_n\) and \( F_t \) are linearly related to one another, the fluorescence scale \( \alpha \) and calcium scale \( \Delta \) are not identifiable. In other words, either can be set to zero without loss of generality because the other can absorb the scale entirely. Similarly, the fluorescence offset \( \beta \) and calcium baseline \( C_B \) are not identifiable, so either can be set to zero without loss of generality. Finally, letting \( \gamma = (1 - \Delta/\tau), \) Eq. 2 can be rewritten by replacing \([Ca^{2+}]_n\) with its nondimensionalized counterpart \( C_t \):

\[
C_t = \gamma C_{t-1} + n_t. \tag{3}
\]

Note that \( C_t \) does not refer to absolute intracellular concentration of calcium, but rather, a relative measure (for a more general model see Vogelstein et al. 2009). The gray line in the bottom left panel of Fig. 1 corresponds to the putative \( C \) of the observed neuron.

To complete the “generative model” (i.e., a model from which simulations can be generated), the distribution from which spikes are sampled must be defined. Perhaps the simplest first-order description of spike trains is that at each time, spikes are sampled according to a Poisson distribution with some rate:

\[
n_t \sim \text{Poisson}(\lambda \Delta) \tag{4}
\]

where \( \lambda \Delta \) is the expected firing rate per bin and \( \Delta \) is included to ensure that the expected firing rate is independent of the frame rate. Thus Eqs. 1, 3, and 4 complete the generative model.

**Goal**

Given the above model, the goal is to find the maximum posteriori (MAP) spike train, i.e., the most likely spike train \( \hat{n} \), given the fluorescence measurements, \( F \):

\[
\hat{n} = \arg \max_{n \in \mathbb{N}_0^T} P[n|F], \tag{5}
\]

where \( P[n|F] \) is the posterior probability of a spike train \( n \), given the fluorescent trace \( F \), and \( n_t \) is constrained to be an integer \( n_t \in \{0, 1, \ldots, T\} \).
because of the above assumed Poisson distribution. From Bayes’ rule, the posterior can be rewritten:

\[ P[n|F] = \frac{P[F|n]P[n]}{P[F]} = \frac{1}{P[F]} P[F|n]P[n], \quad (6) \]

where \( P[F] \) is the evidence of the data, \( P[F|n] \) is the likelihood of observing a particular fluorescence trace \( F \), given the spike train \( n \), and \( P[n] \) is the prior probability of a spike train. Plugging the far right-hand side of Eq. 6 into Eq. 5, yields:

\[ \hat{n} = \arg\max_{n \in N_0^T} \frac{1}{P[F]} P[F|n]P[n] = \arg\max_{n \in N_0^T} P[F|n]P[n], \quad (7) \]

where the second equality follows because \( P[F] \) merely scales the results, but does not change the relative quality of any particular spike train. Note that the prior \( P[n] \) acts as a regularizing term, potentially imposing sparseness or smoothness, depending on the assumed distribution (Seeger 2008; Wu et al. 2006). Both \( P[F|n] \) and \( P[n] \) are available from the preceding model:

\[ P[F|n] = P[F|C] = \prod_{t=1}^{T} P[F_t|C_t], \quad (8a) \]
\[ P[n] = \prod_{t=1}^{T} P[n_t], \quad (8b) \]

where the first equality in Eq. 8a follows because \( C \) is deterministic given \( n \), and the second equality follows from Eq. 1. Further, Eq. 8b follows from the Poisson process assumption, Eq. 4. Both \( P[F|C] \) and \( P[n] \) can be written explicitly:

\[ P[F_t|C_t] = \mathcal{N}(\alpha C_t + \beta, \sigma^2), \quad (9a) \]
\[ P[n_t] = \text{Poisson}(\lambda \Delta), \quad (9b) \]

where both equations follow from the preceding model and the Poisson distribution acts as a sparse prior. Now, plugging Eq. 9 back into Eq. 8, and plugging that result into Eq. 7, yields:

\[ \hat{n} = \arg\max_{n \in N_0^T} \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} (F_t - \alpha C_t - \beta)^2 \right\} \exp\left\{ -\lambda \Delta \right\} n_t! \]

\[ = \arg\max_{n \in N_0^T} \sum_{t=1}^{T} \left\{ -\frac{1}{2\sigma^2} (F_t - \alpha C_t - \beta)^2 + n_t \ln \lambda \Delta - n_t \right\}, \quad (10a) \]

where the second equality follows from taking the logarithm of the right-hand side and dropping terms that do not depend on \( n \). Unfortunately, solving Eq. 10b exactly is analytically intractable because it requires a nonlinear search over an infinite number of possible spike trains. The search space could be restricted by imposing an upper bound \( k \) on the number of spikes within a frame. However, in that case, the computational complexity scales exponentially with the number of image frames—i.e., the number of computations required would scale with \( k^T \)—which for pragmatic reasons is intractable.

**Inferring the approximately most likely spike train, given a fluorescence trace**

The goal here is to develop an algorithm to efficiently approximate \( \hat{n} \), the most likely spike train given the fluorescence trace. Because of the intractability described earlier, one can approximate Eq. 4 by replacing the Poisson distribution with an exponential distribution of the same mean (note that potentially more accurate approximations are possible, as described in the discussion). Modifying Eq. 10 to incorporate this approximation yields:

\[ \hat{n} = \arg\max_{n \geq 0} \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} (F_t - \alpha C_t - \beta)^2 \right\} \exp\left\{ -\lambda \Delta \right\} \exp\left\{ -n_t \lambda \Delta \right\} \]

\[ = \arg\max_{n \geq 0} \sum_{t=1}^{T} \left\{ -\frac{1}{2\sigma^2} (F_t - \alpha C_t - \beta)^2 - n_t \lambda \Delta \right\}, \quad (11a) \]

which is a sum of terms that are concave in \( C \), so the whole right-hand side is concave in \( C \). Unfortunately, the integer constraint has been lost, i.e., the answer could include “partial” spikes. This disadvantage can be remedied by thresholding (i.e., setting \( n_t = 1 \) for all \( n_t \) greater than some threshold and the rest setting to zero) or by considering the magnitude of a partial spike at time \( t \) as a rough indication of the probability of a spike occurring during frame \( t \). Note the relaxation of a difficult discrete optimization problem into an easier continuous problem is a common approximation technique in the machine learning literature (Boyd and Vandenberghe 2004; Paninski et al. 2009).

In particular, the exponential distribution is a convenient nonnegative log-concave approximation of the Poisson (see the discussion for more details).

Although this convex relaxation makes the problem tractable, the “sharp” threshold imposed by the nonnegativity constraint prohibits the use of standard gradient ascent techniques. This may be rectified by using an “interior-point” method (Boyd and Vandenberghe 2004). Interior-point methods solve nondifferentiable problems indirectly by instead solving a series of differentiable subproblems that converge to the solution of the original nondifferentiable problem. In particular, each subproblem within the series drops the sharp threshold and adds a weighted barrier term that approaches \(-\infty\) as \( n_t \) approaches zero. Iteratively reducing the weight of the barrier term guarantees convergence to the correct solution. Thus the goal is to efficiently solve:

\[ \hat{C} = \arg\max_{C} \sum_{t=1}^{T} \left\{ -\frac{1}{2\sigma^2} (F_t - \alpha C_t - \beta)^2 - (C_t - \gamma C_{t-1}) \lambda \Delta \right\} \quad (12) \]

where \( \ln (\cdot) \) is the “barrier term” and \( z \) is the weight of the barrier term (note that the constraint has been dropped). Iteratively solving for \( \hat{C} \) for \( z \) going down to nearly zero guarantees convergence to \( \hat{C} \) (Boyd and Vandenberghe 2004). The concavity of Eq. 13 facilitates using any number of techniques guaranteed to find the global maximum. Because the argument of Eq. 13 is twice analytically differentiable, one can use the Newton–Raphson technique (Press et al. 1992). The special tridiagonal structure of the Hessian enables each Newton–Raphson step to be very efficient (as described below). To proceed, Eq. 13 is first rewritten in more compact matrix notation. Note that:
Hessian for this model, with respect to zero, \(\ln MC\) tions (where \(k\) standard scales). Where the exponents on the vector \(\mathbf{MC}\) acquire inverting the Hessian, i.e., solving \(0 \rightarrow 1\) (Press et al. 1992). Given this normalization, \(\alpha\) is set to one. Because spiking is sparse in many experimental settings, \(F\) tends to be around baseline, so \(\beta\) is initialized to be the median of \(F\) and \(\sigma\) is initialized as the median absolute deviation of \(F\), where \(\sigma\) indicates the median of \(X\) with respect to index \(i\) and \(K = 1.4785\) is the correction factor when using the median absolute deviation as a robust estimator of the SD of a normal distribution. Because in these data the posterior tends to be relatively flat along the \(y\) dimension (i.e., large changes in \(y\) result in relatively small changes in the posterior), estimating \(y\) is difficult. Further, previous work has shown that results are somewhat robust to minor variations in the time constant (Yaksi and Friedrich 2006); therefore \(\gamma\) is initialized at \(1 - \Delta(1)\), which is fairly standard (Pologruto et al. 2004). Finally, \(\lambda\) is initialized at 1 Hz, which is between average baseline and evoked spike rate for data of interest.

ESTIMATING THE PARAMETERS GIVEN \(\hat{\theta}\). Ideally, one could integrate out the hidden variables, to find the most likely parameters:

\[\hat{\theta} = \arg\max_\theta P(F, C|\theta) dC = \arg\max_\theta P(F|C; \theta)P(C|\theta) dC.\]  

(17)

However, evaluating those integrals is not currently tractable. Therefore Eq. 17 is approximated by simply maximizing the parameters given the MAP estimate of the hidden variables:

\[\hat{\theta} = \arg\max_\theta P(F, \hat{C}|\theta) = \arg\max_\theta P(F|\hat{C}; \theta)P(\hat{C}|\theta)\]  

(18)

where \(\hat{\theta}\) and \(\hat{\theta}\) are determined using the above-described inference algorithm. The approximation in Eq. 18 is good whenever most of the mass in the integral in Eq. 18 is around the MAP sequence \(\hat{C}\). The argument from the right-hand side of Eq. 18 may be expanded:

\[\ln P(F|\hat{C}; \theta) + \ln P(\hat{C}|\theta) = \sum_{i=1}^T \ln P(F_i|\hat{C}_i; \alpha, \beta, \sigma) + \sum_{i=1}^T \ln P(\hat{C}_i|\lambda).\]  

(19)

Note that the right-hand side of Eq. 19 decouples \(\lambda\) from the other parameters. The maximum likelihood estimate (MLE) for the observation parameters \(\{\alpha, \beta, \sigma\}\) is therefore given by:

\[\{\hat{\alpha}, \hat{\beta}, \hat{\sigma}\} = \arg\max_{\alpha, \beta, \sigma} \sum_{i=1}^T \ln P(F_i|\hat{C}_i; \alpha, \beta, \sigma)\]  

(20)

Note that a rescaling of \(\alpha\) may be offset by a complementary rescaling of \(\hat{C}\). Therefore because the scale of \(\hat{C}\) is arbitrary (see Eqs. 2 and 3), \(\alpha\) can be set to one without loss of generality. Plugging \(\alpha = 1\) into Eq. 20 and maximizing with respect to \(\beta\) yields:

\[\hat{\beta} = \arg\max_{\hat{\beta}} \sum_{i=1}^T (F_i - \hat{C}_i - \beta)^2.\]  

(21)

Computing the gradient with respect to \(\beta\), setting the answer to zero, and solving for \(\beta\) yields \(\hat{\beta} = (1/T) \sum_{i=1}^T (F_i - \hat{C}_i)\). Similarly, computing the gradient of Eq. 20 with respect to \(\sigma\), setting it to zero, and solving for \(\sigma\) yields:

\[^1\text{Equation 18 may be considered a crude Laplace approximation (Kass and Raftery 1995).}\]
\[
\hat{\sigma} = \sqrt{\frac{1}{T} \sum (F_t - \hat{C}_t - \hat{\beta})^2}, \tag{22}
\]
which is simply the root-mean-square of the residual error. Finally, the MLE of \(\hat{\lambda}\) is given by solving:
\[
\hat{\lambda} = \arg\max_{\lambda > 0} \sum_t [\ln (\lambda \Delta) - \hat{\sigma}_t \lambda \Delta], \tag{23}
\]
which, again, computing the gradient with respect to \(\lambda\), setting it to zero, and solving for \(\hat{\lambda}\), yields \(\hat{\lambda} = \theta(\Delta, \hat{\sigma}_t)\), which is the inverse of the inferred average firing rate.

Iterations stop whenever 1) the iteration number exceeds some upper bound or 2) the relative change in likelihood does not exceed some lower bound. In practice, parameter estimates tend to converge after several iterations, given the above initializations.

**Spatial filtering**

In the preceding text, we assumed that the raw movie of fluorescence measurements collected by the experimenter had undergone two stages of preprocessing before filtering. First, the movie was segmented, to determine ROIs, yielding a vector \(F_t\), which corresponded to the fluorescence intensity for each of the \(N_p\) pixels in the ROI (note that we use the \(X\) throughout to indicate row vectors in space vs. \(X\) to indicate column vectors in time). Second, at each time \(t\), that vector was projected into a scalar, yielding \(F_s\), the assumed input to the filter. In this section, the optimal projection is determined by considering a more general model:
\[
F_{x,t} = \alpha_i C_t + \beta_i + \sigma e_x + \varepsilon_{x,t}, \quad e_x \sim \text{N}(0,1), \tag{24}
\]
where \(\alpha_i\) corresponds to the number of photons that are contributed due to calcium fluctuations \(C_t\), and \(\beta_i\) corresponds to the static photon emission at pixel \(x\). Further, the noise is assumed to be both spatially and temporally white, with standard deviation (SD) \(\sigma\), in each pixel (this assumption can always be approximately accurate by prewhitening; alternately, one could relax the spatial independence by representing joint noise over all pixels with a covariance matrix \(\Sigma_x\), with arbitrary structure). Performing inference in this more general model proceeds in a nearly identical manner as before. In particular, the maximization, gradient, and Hessian become:
\[
\hat{C}_t = \arg\max_{MC \geq \theta} -\frac{1}{2\sigma^2} \| \tilde{F} - C\hat{a} - 1_i \hat{\beta} \|^2_{\Sigma} - (MC)^T \Lambda + z \ln z (MC)^T \mathbf{1}, \tag{25}
\]
\[
g = (\tilde{F} - C\hat{a} - 1_i \hat{\beta})^T \hat{a} - M^T \lambda + z M^T (MC)^T \mathbf{1}, \tag{26}
\]
\[
H = -\frac{\hat{a}}{\sigma^2} g^T - \frac{M^T \lambda}{\sigma^2} - \frac{z M^T (MC)^T}{\sigma^2} M, \tag{27}
\]
where \(\tilde{F}\) is an \(N_o \times T\) element matrix, \(1_i\) is a column vector of ones with length \(T\), \(I\) is an \(N_o \times N_o\) identity matrix, and \(||\cdot||_{\Sigma}\) indicates the Frobenius norm, i.e., \(||X||_{\Sigma} = \sum_{ij} x_{ij}^2\), and the exponents and log operator on the vector \(MC\) again indicate element-wise operations. Note that to speed up computation, one can first project the background subtracted \((N_o \times T)\)-dimensional movie onto the spatial filter \(\hat{a}\), yielding a one-dimensional time series \(F_s\), reducing the problem to evaluating a \(T \times 1\) vector norm, as in Eq. 15.

The parameters \(\hat{a}\) and \(\hat{\beta}\) tend to be unknown and thus must be estimated from the data. Following the strategy developed in the previous section, we first initialize the parameters. Because each voxel contains some number of fluorophores, which sets both the baseline fluorescence and the fluorescence due to calcium fluctuations, let both the initial spatial filter and initial background be the median image frame, i.e., \(\hat{a}_s = \hat{\beta}_s = \text{median}(F_{x,t})\). Given these robust initializations, the maximum likelihood estimator for each \(\alpha_i\) and \(\beta_i\) is given by:
\[
\{\hat{\alpha}_i, \hat{\beta}_i\} = \arg\max_{\alpha,\beta} P[F|\hat{C}] \tag{28a}
\]
\[
= \arg\max_{\alpha,\beta} \sum_t \ln P[F_{x,t}|\hat{C}_t] \tag{28b}
\]
\[
= \arg\max_{\alpha,\beta} \sum_t \left\{ -\frac{1}{2} \ln (2\pi\sigma^2) - \frac{1}{2\sigma^2} (F_{x,t} - \alpha_i \hat{C}_i - \beta_i)^2 \right\} \tag{28c}
\]
\[
= \arg\max_{\alpha,\beta} \sum_t (F_{x,t} - \alpha_i \hat{C}_i - \beta_i)^2, \tag{28d}
\]
where the first equalities follow from Eq. 1 and the last equality follows from dropping irrelevant constants. Because this is a standard linear regression problem, let \(A = [\hat{C}_1, \ldots, \hat{C}_T]^T\) be a \(2 \times T\) element matrix and \(Y_i = [\alpha_i, \beta_i]^T\) be a \(2 \times 1\) element column vector. Substituting \(A\) and \(Y_i\) into Eq. 28d yields:
\[
\hat{Y}_i = \arg\max_{Y_i} \| F_i - A^T Y_i \|^2, \tag{29}
\]
which can be solved by computing the derivative of Eq. 29 with respect to \(Y_i\) and setting to zero, or using Matlab notation: \(\hat{Y}_i = \text{A}^T \text{F}_i\). Note that solving \(N_o\) two-dimensional quadratic problems is more efficient than solving a single \((2 \times N_o)\)-dimensional quadratic problem. Also note that this approach does not regularize the parameters at all, by smoothing or sparsening, for instance. In the DISCUSSION we propose several avenues for further development, including the elastic net (Zou and Hastie 2005) and simple parameter models of the neuron. As in the scalar case, we iterate estimating the parameters of this model \(\theta = (\hat{a}, \hat{\beta}, \sigma, \gamma, \lambda)\) and the spike train \(n\). Because of the free scale term discussed earlier, the absolute magnitude of \(\hat{a}\) is not identifiable. Thus convergence is defined here by the “shape” of the spike train converging, i.e., the norm of the difference between the inferred spike trains from subsequent iterations, both normalized such that \(\max(\hat{\sigma}_t) = 1\). In practice, this procedure converged after several iterations.

**Overlapping spatial filters**

It is not always possible to segment the movie into pixels containing only fluorescence from a single neuron. Therefore the above-cited model can be generalized to incorporate multiple neurons within an ROI. Specifically, letting the superscript \(i\) index the \(N_c\) neurons in this ROI yields:
\[
\hat{F}_i = \sum_{i=1}^{N_c} \hat{\alpha}_i C_{i,t} + \hat{\beta}_{i,t} + \hat{\varepsilon}_{i,t} \sim \text{N}(0,\sigma^2 I) \tag{30}
\]
\[
C_{i,t} = \gamma C_{i,t-1} + n_{i,t} \sim \text{Poisson}(n_{i,t};\lambda,\Delta) \tag{31}
\]
where each neuron is implicitly assumed to be independent and each pixel is conditionally independent and identically distributed with variance \(\sigma^2\), given the underlying calcium signals. To perform inference in this more general model, let \(n_s = [n_s^1, \ldots, n_s^{N_o}]\) and \(C_s = [C_s^1, \ldots, C_s^{N_o}]\) be \(N_o\)-dimensional column vectors. Then, let \(\Gamma = \text{diag}(\gamma^1, \ldots, \gamma^{N_o})\) be an \(N_o \times N_o\) diagonal matrix and let \(I\) and \(0\) be an identity and zero matrix of the same size, respectively, yielding:

\[J\text{ Neurphysiol} \hspace{1em} \text{VOL 104} \hspace{1em} \text{DECEMBER 2010} \hspace{1em} \text{www.jn.org} \]
If the parameters are unknown, they must be estimated. Initialize bulk loaded with an acetoxymethyl ester of Fura-2 (Fura-2 AM; letting for this block-tridiagonal structure (Press et al. 1992). Performing efficiency gained from using the tridiagonal structure is maintained ear equations with block-tridiagonal matrices in linear time, so the simplified form of Gaussian elimination, finds the solution to lin-

\[
MC = \begin{bmatrix}
-\Gamma & 1 & 0 & 0 & \cdots & 0 \\
0 & -\Gamma & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & \cdots & 0 & -\Gamma & 1 \\
0 & \cdots & 0 & 0 & -\Gamma & 1
\end{bmatrix}
\begin{bmatrix}
C_1 \\
C_2 \\
\vdots \\
C_{T-1} \\
C_T
\end{bmatrix}
= \begin{bmatrix}
n_1 \\
n_2 \\
\vdots \\
n_{T-1}
\end{bmatrix}
\tag{32}
\]

and proceed as before. Note that Eq. 32 is very similar to Eq. 14, except that \( M \) is no longer bidiagonal, but rather, block bidiagonal (and \( C_i \) and \( n_i \) are vectors instead of scalars), making the Hessian block-tridiagonal. Importantly, the Thomas algorithm, which is a simplified form of Gaussian elimination, finds the solution to linear equations with block-tridiagonal matrices in linear time, so the efficiency gained from using the tridiagonal structure is maintained.

\[
\hat{C}_t = \arg\max_{MC \succeq 0} -\frac{1}{2\sigma^2} \| \hat{F} - C\hat{\alpha} - 1_{T}\hat{\beta}^2_{T} - (MC)^{T}\lambda + z \ln (MC)^{T}1, \tag{33}
\]

\[
g = (\hat{F} - C\hat{\alpha} - 1_{T}\hat{\beta}^2_{T}) \frac{\hat{\alpha}^T}{\sigma^2} - M^T\lambda + zM^T(MC)^{-1} \tag{34}
\]

\[
H = -\frac{\hat{\alpha}\hat{\alpha}^T}{\sigma^2} - zM^T(MC)^{-2}M. \tag{35}
\]

If the parameters are unknown, they must be estimated. Initialize \( \beta \) as above. Then, define \( \alpha_k = [\alpha_1^k, \ldots, \alpha_n^k]^T \) and initialize manually by assigning some pixels to each neuron (of course, more sophisticated algorithms could be used, as described in the DISCUSSION). Given this initialization, iterations and stopping criteria proceed as before, with the minor modification of incorporating multiple spatial filters, yielding:

\[
\{\hat{\alpha}, \hat{\beta}\} = \arg\max_{\alpha_k, \beta_i} -\frac{1}{2} \sum_{t=1}^{T} (F_{t,j} - \sum_{i=1}^{N} \alpha_i^k \hat{C}_i - \beta_i)^2, \tag{36}
\]

Now, generalizing the above single spatial filter case, let \( A = [\hat{C}, 1_j]^T \) be an \((N_c + 1) \times T\) element matrix and \( Y_j = [\alpha_c, \beta]^T \) be an \((N_c + 1)\)-dimensional column vector. Then, one can again use Eq. 29 to solve for \( \hat{\alpha} \) and \( \hat{\beta} \) for all \( x \).

**Experimental methods**

**SLICE PREPARATION AND IMAGING.** All animal handling and experimentation were done according to the National Institutes of Health and local Institutional Animal Care and Use Committee guidelines. Somatosensory thalamocortical or coronal slices 350–400 \( \mu \)m thick were prepared from C57BL/6 mice at age P14 as described (MacLean et al. 2005). Pyramidal neurons from layer V somatosensory cortex were filled with 50 \( \mu \)M Oregon Green BAPTA 1 hexapotassium salt (OGB-1; Invitrogen, Carlsbad, CA) through the recording pipette or bulk loaded with an acetoxymethyl ester of Fura-2 (Fura-2 AM; Invitrogen). The pipette solution contained 130 mM K-methylsulfate, 2 mM MgCl\(_2\), 0.6 mM EGTA, 10 mM HEPES, 4 mM ATP-Mg, and 0.3 mM GTP-Tris (pH 7.2, 295 mOsm). After cells were fully loaded with dye, imaging was performed in one of two ways. First, when using Fura-2, images were collected using a modified BX50-WI upright microscope (Olympus, Melville, NY) with a confocal spinning disk (Solamere Technology Group, Salt Lake City, UT) and an Orca charge-coupled device (CCD) camera from Hamamatsu Photonics (Shizuoka, Japan), at 33 Hz. Second, when using Oregon Green, images were collected with epifluorescence with the C9100-12 CCD camera from Hamamatsu Photonics, with arc-lamp illumination with excitation and emission band-pass filters at 480–500 and 510–550 nm, respectively (Chroma, Rockingham, VT). Images were saved and analyzed using custom software written in Matlab (The MathWorks, Natick, MA).

**ELECTROPHYSIOLOGY.** All recordings were made using the Multi-clamp 700B amplifier (Molecular Devices, Sunnyvale, CA), digitized with National Instruments 6259 multichannel cards and recorded using custom software written using the LabVIEW platform (National Instruments, Austin, TX). Square pulses of sufficient amplitude to yield the desired number of action potentials were given as current commands to the amplifier using the LabVIEW and National Instruments system.

**FLUORESCENCE PREPROCESSING.** Traces were extracted using custom Matlab scripts to segment the mean image into ROIs. The Fura-2 fluorescence traces were inverted. Because some slow drift was sometimes present in the traces, each trace was Fourier transformed, and all frequencies <0.5 Hz were set to zero (0.5 Hz was chosen by eye); the resulting fluorescence trace was then normalized to be between zero and one.

**RESULTS**

**Main result**

The main result of this study is that the fast filter can find the approximately most likely spike train \( \hat{\mathbf{n}} \), very efficiently, and that this approach yields more accurate spike train estimates than optimal linear deconvolution. Figure 2 depicts a simulation showing this result. Clearly, the fast filter’s inferred “spike train” (third panel) more closely resembles the true spike train (second panel) than the optimal linear deconvolution’s inferred spike train (bottom panel; Wiener filter). Note that neither filter results in an integer sequence, but rather, each infers a real number at each time.
The Wiener filter implicitly approximates the Poisson spike rate with a Gaussian spike rate (see APPENDIX B for details). A Poisson spike rate indicates that in each frame, the number of possible spikes is an integer, e.g., 0, 1, 2,…. The Gaussian approximation, however, allows any real number of spikes in each frame, including both partial spikes (e.g., 1.4) and negative spikes (e.g., −0.8). Although a Gaussian well approximates a Poisson distribution when rates are about 10 spikes per frame, this example is very far from that regime, so the Gaussian approximation performs relatively poorly. Further, the Wiener filter exhibits a “ringing” effect. Whenever fluorescence drops rapidly, the most likely underlying spiking signal is a proportional drop. Because the Wiener filter does not impose a nonnegative constraint on the underlying spiking signal, it infers such a drop, even when it causes it to go below zero. After such a drop has been inferred, since no corresponding drop occurred in the true underlying signal here, a complementary jump is often then inferred, to realign the inferred signal with the observations. This oscillatory behavior results in poor inference quality. The nonnegative constraint imposed by the fast filter prevents this because the underlying signal never drops below zero, so the complementary jump never occurs either.

The inferred “spikes,” however, are still not binary events when using the fast filter. This is a by-product of approximating the Poisson distribution on spikes with an exponential (cf. Eq. 11a) because the exponential is a continuous distribution, versus the Poisson, which is discrete. The height of each spike is therefore proportional to the inferred calcium jump size and can be thought of as a proxy for the confidence with which the algorithm believes a spike occurred. Importantly, by using the Gaussian elimination and interior-point methods, as described in METHODS, the computational complexity of the fast filter is the same as an efficient implementation of the Wiener filter. Note that whereas the Gaussian approximation imposes a shrinkage prior on the inferred spike trains (Wu et al. 2006), the exponential approximation imposes a sparse prior on the inferred spike trains (Seeger 2008).

Figure 3 quantifies the relative performance of the fast and Wiener filters. The top left panel shows a typical simulated spike train (bottom), a corresponding relatively low SNR fluorescence trace (middle), and a relatively high SNR fluorescence trace (top), as examples. The top right panel compares the mean-squared-error (MSE) of the inferred spike trains using the fast (solid) and Wiener (dashed) filters, as a function of expected firing rate. Clearly, the fast filter has a better (lower) MSE for all expected firing rates. Error bars show SD over 10 repeats. Simulation parameters: $\sigma = 0.2$, $T = 1,000$ time steps. Bottom left: receiver-operator-characteristic (ROC) curve comparing the fast (solid line) and Wiener (dashed-dotted line) filter. Note that for any given threshold, the Wiener filter has a better (higher) ratio of true positive rate to false positive rate. Simulation parameters as in top right panel, except $\sigma = 0.35$ and $T = 10,000$ time steps. Bottom right: area under the curve (AUC) for fast (solid line) and Wiener (dashed-dotted line) filter as a function of SD ($\sigma$). Note that the fast filter has a better (higher) AUC for all $\sigma$ values until noise gets very high. The 2 simulated fluorescence traces in the top left panel show the bounds for SD here. Error bars show SD over 10 repeats.

Although in Fig. 2 the model parameters were provided, in the general case, the parameters are unknown and must therefore be estimated from the observations (as described in Learning the parameters in METHODS). Importantly, this algorithm does not require labeled training data, i.e., there is no need for joint imaging and electrophysiological experiments to estimate the parameters governing the relationship between the two. Figure 4 shows another simulated example; in this example, however, the parameters are estimated from the observed fluorescence trace (Seeger 2008).
fluorescence trace. Again, it is clear that the fast filter far outperforms the Wiener filter.

Given the preceding two results, the fast filter was applied to biological data. More specifically, by jointly recording electrophysiologically and imaging, the true spike times are known and the accuracy of the two filters can be compared. Figure 5 shows a result typical of the 12 joint electrophysiological and imaging experiments conducted (see METHODS for details). As in the simulated data, the fast filter output is much “cleaner” than the Wiener filter: spikes are more well defined, and not spread out, due to the sparse prior imposed by the exponential approximation. Note that this trace is typical of epifluorescence techniques, which makes resolving individual spikes quite difficult, as evidenced by a few false positives in the fast filter. Regardless, the fast filter output is still more accurate than the Wiener filter, both as determined qualitatively by eye and as quantified (described in the following text). Furthermore, although it is difficult to see in this figure, the first four events are actually pairs of spikes, reflected by the width and height of the corresponding inferred spikes when using the fast filter. This suggests that although the scale of $n$ is arbitrary, the fast filter can correctly ascertain the number of spikes within spike events.

Figure 6 further evaluates this claim. While recording and imaging, the cell was forced to spike once, twice, or thrice for each spiking event. The fast filter infers the correct number of spikes in each event. On the contrary, there is no obvious way to count the number of spikes within each event when using the Wiener filter. We confirm this impression by computing the correlation coefficient, $r^2$, between the sum of each filter’s output and the true number of spikes, for all 12 joint electrophysiological and imaging traces. Indeed, whereas the fast filter’s $r^2$ was 0.47, the Wiener filter’s $r^2$ was $-0.01$ (after thresholding all negative spikes), confirming that the Wiener filter output cannot reliably convey the number of spikes in a fluorescence trace, whereas the fast filter can. Furthermore, varying the magnitude of the threshold for the Wiener filter to discard more “low-amplitude noise” could increase the magnitude of $r^2$ ($<0.24$), still significantly lower than the fast filter’s $r^2$ value. On the other hand, no amount of thresholding the fast filter yielded an improved $r^2$, indicating that thresholding the output of the fast filter is unlikely to improve spike inference quality.

**On-line analysis of spike trains using the fast filter**

A central aim for this work was the development of an algorithm that infers spikes fast enough to use on-line while imaging a large population of neurons (e.g., $>100$). Figure 7 shows a segment of the results of running the fast filter on 136 neurons, recorded simultaneously, as described earlier in Experimental methods. Note that the filtered fluorescence signals show fluctuations in spiking much more clearly than the unfiltered fluorescence trace. These spike trains were inferred in less than imaging time, meaning that one could infer spike trains for the past experiment while conducting the subsequent experiment. More specifically, a movie with 5,000 frames of 100 neurons can be analyzed in about 10 s on a standard desktop computer. Thus if that movie was recorded at 50 Hz, whereas collecting the data would require 100 s, inferring spikes would require only 10 s, a 10-fold improvement over real time.

**Extensions**

Earlier in METHODS, Data-driven generative model describes a simple principled first-order model relating the spike trains to the fluorescence trace. A number of the simplifying assumptions can be straightforwardly relaxed, as described next.

Replacing Gaussian observations with poisson. In the preceding text, observations were assumed to have a Gaussian distribution. The statistics of photon emission and counting, however, suggest that a Poisson distribution would be more natural in some conditions, especially for two-photon data (Sjulson and Miesenböck 2007), yielding:

$$F_i \sim \text{Poisson} \left( \alpha C_i + \beta \right),$$

where $\alpha C_i + \beta \geq 0$. One additional advantage to this model over the Gaussian model is that the variance parameter $\sigma^2$ no
The fast filter infers spike trains from a large population of neurons imaged simultaneously in vitro, faster than real time. Specifically, inferring the spike trains from this 400-s-long movie including 136 neurons required only about 40 s on a standard laptop computer. The inferred spike trains much more clearly convey neural activity than the raw fluorescence traces. Although no intracellular “ground truth” is available from these population data, the noise seems to be reduced, consistent with the other examples with ground truth. ROIs), each containing a single neuron using custom software. Middle: example fluorescence traces. Right: fast filter output corresponding to each associated trace. Note that neuron identity is indicated by color across the 3 panels. Data were collected using a confocal microscope and Fura-2, as described in METHODS.

longer exists, which might make learning the parameters simpler. Importantly, the log-posterior is still concave in $C$, as the prior remains unchanged, and the new log-likelihood term is a sum of terms concave in $C$:

$$\ln P(F\mid C) = \sum_{t=1}^{T} \ln P(F_t\mid C_t) = \sum_{t=1}^{T} \left\{ F_t \ln (\alpha C_t + \beta) - (\alpha C_t + \beta) - \ln (F_t) \right\}. \quad (38)$$

The gradient and Hessian of the log-posterior can therefore be computed analytically by substituting the above likelihood terms for those implied by Eq. 1. In practice, however, modifying the filter for this model extension did not seem to significantly improve inference results in any simulations or data available at this time (not shown).

Allowing for a time-varying prior. In Eq. 4, the rate of spiking is a constant. Often, additional knowledge about the experiment, including external stimuli or other neurons spiking, can provide strong time-varying prior information (Vogelstein et al. 2009). A simple model modification can incorporate that feature:

$$n_{t,i} \sim \text{Poisson} (\lambda_{t,i}), \quad (39)$$

where $\lambda_t$ is now a function of time. Approximating this time-varying Poisson with a time-varying exponential with the same time-varying mean (similar to Eq. 11a) and letting $\lambda = [\lambda_1, \ldots, \lambda_T]^T \Delta$, yields an objective function very similar to Eq. 15, so log-concavity is maintained and the same techniques may be applied. However, as before, this model extension did not yield any significantly improved filtering results (not shown).

Saturating fluorescence. Although all the abovementioned models assumed a linear relationship between $F_t$ and $C_t$, the relationship between fluorescence and calcium is often better approximated by the nonlinear Hill equation (Pologruto et al. 2004). Modifying Eq. 1 to reflect this change yields:

$$F_t = \alpha \frac{C_t}{C_t + k_d} + \beta + \epsilon_t \sim \mathcal{N}(0, \sigma^2). \quad (40)$$

Importantly, log-concavity of the prior is no longer guaranteed in this nonlinear model, meaning that converging to the global maximum is no longer guaranteed. Assuming a good initialization can be found, however, and Eq. 40 is more accurate than Eq. 1, then ascending the gradient for this model is likely to yield improved inference results. In practice, initializing with the inference from the fast filter assuming a linear model (e.g., Eq. 30) often resulted in nearly equally accurate inference, but inference assuming the above nonlinearity was far less robust than the inference assuming the linear model (not shown).

Using the fast filter to initialize the SMC filter. A sequential Monte Carlo (SMC) method to infer spike trains can incorporate this saturating nonlinearity, as well as other model extensions discussed earlier (Vogelstein et al. 2009). However, this SMC filter is not nearly as computationally efficient as the fast filter proposed here. Like the fast filter, the SMC filter estimates the model parameters in a completely unsupervised fashion, i.e., from the fluorescence observations, using an expectation-maximization algorithm (which requires iterating between computing the expected value of the hidden variables—$C$ and $n$—and updating the parameters). In Vogelstein and colleagues (2009), parameters for the SMC filter were initialized based on other data. Although effective, this initialization was often far from the final estimates and thus required a relatively large number of iterations (e.g., 20–25) before converging. Thus it seemed that the fast filter could be used to obtain an improvement to the initial parameter estimates, given an appropriate rescaling to account for the nonlinearity, thereby reducing the required number of iterations to convergence. Indeed, Fig. 8 shows how the SMC filter outperforms the fast filter on biological data and required only three to five iterations to converge on these data, given the initialization from the fast filter (which was typical). Note that the first few events of the spike train are individual spikes, resulting in relatively small fluorescence fluctuations, whereas the next events are actually spike doublets or triplets, causing a much larger fluorescence fluctuation. Only the SMC filter correctly infers the presence of isolated spikes in this trace, a frequently occurring result when the SNR is poor. Thus these two inference algorithms are complementary: the fast filter can be used for rapid, on-line inference, and for initializing the SMC filter, which can then be used to further refine the spike train. 
Spatial filter

In the preceding text, the filters operated on one-dimensional fluorescence traces. The raw data are in fact a time series of images that are first segmented into regions of interest (ROIs) and then (usually) spatially averaged to obtain a one-dimensional time series $F$. In theory, one could improve the effective SNR of the fluorescence trace by scaling each pixel according to its SNR. In particular, pixels not containing any information about calcium fluctuations can be ignored and pixels that are partially anticorrelated with one another could have weights with opposing signs.

Figure 9 demonstrates the potential utility of this approach. The top row shows different depictions of an ROI containing a single neuron. On the far left panel is the true spatial filter for this neuron. This particular spatial filter was chosen based on experience analyzing both in vitro and in vivo movies; often, it seems that the pixels immediately around the soma are antici-correlated with those in the soma (MacLean et al. 2005; Watson et al. 2008). This effect is possibly due to the influx of calcium from the extracellular space immediately around the soma. The standard approach, given such a noisy movie, would be to first segment the movie to find an ROI corresponding to the soma of this cell and then spatially average all the pixels found to be within this ROI. The second panel shows this standard “boxcar spatial filter.” The third panel shows the mean frame. The fourth panel shows the learned filter, using Eq. 29 to estimate the spatial filter and background. Clearly, the learned filter is very similar to the mean filter and the true filter.

The middle panels of Figure 9 show the fluorescence traces obtained by background subtracting and then projecting each frame onto the corresponding spatial filter (black line) and true spike train (gray + symbols). The bottom panels show the inferred spike trains (black bars) using these various spatial filters and, again, the true spike train (gray + symbols). Although the

FIG. 8. In vitro data with SNR of only about 3 (estimated by dividing the fluorescent jump size by the SD of the baseline fluorescence) for single action potentials depicting the fast filter, effectively initializing the parameters for the sequential Monte Carlo (SMC) filter, significantly reducing the number of expectation-maximization iterations to convergence, using OGB-1. Note that whereas the fast filter clearly infers the spiking events in the end of the trace, those in the beginning of the trace are less clear. On the other hand, the SMC filter more clearly separates nonspiking activity from true spikes. Also note that the ordinate on the third panel corresponds to the inferred probability of a spike having occurred in each frame.

estimate. Importantly, although the SMC filter often outperforms the fast filter, the fast filter is more robust, meaning that it more often works “out of the box.” This follows because the SMC filter operates on a highly nonlinear model that is not log-concave. Thus although the expectation-maximization algorithm used often converges to reasonable local maxima, it is not guaranteed to converge to global maxima and its performance in general will depend on the quality of the initial parameter estimates.
performance is very similar for all of them, the boxcar filter’s inferred spike train is not as clean.

Overlapping spatial filters

The preceding text shows that if the ROI contains only a single neuron, the effective SNR can be enhanced by spatially filtering. However, this analysis assumes that only a single neuron is in the ROI. Often, ROIs are overlapping, or nearly overlapping, making the segmentation problem more difficult. Therefore it is desirable to have an ability to crudely segment, yielding only a few neurons in each ROI, and then spatially filter within each ROI to pick out the spike trains of each neuron. This may be achieved in a principled manner by generalizing the model as described in Overlapping spatial filters in METHODS. The true spatial filters of the neurons in the ROI are often unknown and thus must be estimated from the data. This problem may be considered a special case of blind source separation (Bell and Sejnowski 1995; Mukamel et al. 2009). Figure 10 shows that given reasonable assumptions of spiking correlations and SNR, multiple signals can be separated. Note that separation occurs even though the signal is significantly overlapping (top panels). To estimate the spatial filters, they are initialized using the boxcar filters (middle panels). After a few iterations, the spatial filters converge to very close approximation to the true spatial filters [compare true (left) and learned (right) spatial filters for the two neurons]. Note that both the true and learned spatial filters yield much improved spike inference relative to the boxcar filter. This suggests that even when spatial filters of multiple neurons are significantly overlapping, each spike train is potentially independently recoverable.

DISCUSSION

Summary

This work describes an algorithm that finds the approximate maximum a posteriori (MAP) spike train, given a calcium fluorescence movie. The approximation is required because finding the actual MAP estimate is not currently computationally tractable. Replacing the assumed Poisson distribution on spikes with an exponential distribution yields a log-concave optimization problem, which can be solved using standard gradient ascent techniques (such as Newton–Raphson). This exponential distribution has an advantage over a Gaussian distribution by restricting spikes to be positive, which improves inference quality (cf. Fig. 2), is a better approximation to a Poisson distribution with low rate, and imposes a sparse constraint on spiking. Furthermore, all the parameters can be estimated from only the fluorescence observations, obviating the need for joint electrophysiology and imaging (cf. Fig. 4).

This approach is robust, in that it works “out of the box” on all the in vitro data analyzed (cf. Figs. 5 and 6). By using the special banded structure of the Hessian matrix of the log-posterior, this approximate MAP spike train can be inferred fast enough on standard computers to use it for on-line analyses of over 100 neurons simultaneously (cf. Fig. 7).

Finally, the fast filter is based on a biophysical model capturing key features of the data and may therefore be straightforwardly generalized in several ways to improve ac-

![Fig. 10. Simulation showing that when 2 neurons’ spatial filters are largely overlapping, learning the optimal spatial filters using Eq. 36 can yield improved inference of the standard boxcar type filters. The 3 columns show the effect of the true (left), boxcar (center), and learned (right) spatial filters. A: the sum of the 2 spatial filters for each approach, clearly depicting overlap. B: the spatial filters (top row), one-dimensional fluorescence projection, and inferred spike train (bottom row) for one of the neurons. C: same as B for the other neuron. Note that the inferred spike trains when using the learned filter are close to optimal, unlike the boxcar filter. Simulation parameters: $\alpha^2 = \mathcal{N}([-1, 0], 2I) - 0.5\mathcal{N}([-1, 0], 2.5I)$, $\alpha^2 = \mathcal{N}([1, 0], 2I) - 0.5\mathcal{N}([1, 0], 2.5I)$, $\beta = 0$, $\sigma = 0.02$, $\tau = 0.5$ s, $\lambda = 5$ Hz, $\Delta = 5$ ms, $T = 1,200$ time steps (not all time steps are shown).](./image)
Innovative Methodology

Q

log-density

\ln Q

not search over all possible densities, but rather, simply vary

will be banded (because enforcing concavity of \( Q \) is a local

constraint). This approximation could potentially be more

accurate than our exponential approximation. Further, this

approximation encourages integer solutions for \( n \), and is there-

fore of interest for future work.

The abovementioned three approaches may be thought of as

complementary because each has unique advantages relative to

the others. Both the greedy methods and the analytic approx-

imations could potentially be used to initialize a Bayesian

approach, possibly limiting the burn-in period, which can be

computationally prohibitive in certain contexts. A greedy ap-

proach has the advantage of providing actual spike trains (i.e.,

binary sequences), unlike the analytic approximations. How-

ever, the actual spike trains could be quite far from the MAP

spike train because greedy approaches, in general, have no

guarantee of consistency. The analytic approximations, on the

other hand, are guaranteed to converge to solutions close to the

MAP spike train, where closeness is determined by the accu-

racy of the above approximation. Thus developing these dis-

tinct approaches and combining them is a potential avenue for

further research.

Alternate algorithms

This work describes but one specific approach to solving a

problem that does not admit an exact solution that is comput-

ationally feasible. Several other approaches warrant consider-

ation, including 1) a Bayesian approach, 2) a greedy approach,

and 3) different analytical approximations.

First, a Bayesian approach could use Markov Chain Monte

Carlo methods to recursively sample spikes to estimate the full

joint posterior distribution of the entire spike train, conditioned

on the fluorescence data (Andrieu et al. 2001; Joucla et al.

2010; Mishchenko et al. 2010). Although enjoying several

desirable statistical properties that are lacking in the current

approach (such as consistency), the computational complexity

of such an approach renders it inappropriate for the aims of this

work.

Second, a common relatively expedient approximation to

Bayesian sampling is a so-called greedy approach. Greedy

algorithms are iterative, with each iteration adding another

spike to the putative spike train. Each spike that is added is the

most likely spike (thus the greedy term) or the one that most

increases the likelihood of the fluorescence trace. Template

matching, projection pursuit regression (Friedman and Stuetzle

1981), and matching pursuit (Mallat and Zhang 1993) are

eamples of such a greedy approach (the algorithm proposed by

Grewe et al. (2010) could also be considered a special case of

such a greedy approach).

Third, approximations other than the exponential distribu-

tion are possible. For instance, the Gaussian approximation is

more appropriate for high firing rates, although in simulations,

this more accurate approximation did not improve the Wiener

filter output relative to the fast filter output (cf. Fig. 3). Perhaps

the best approximation would use the closest log-concave

relaxation to the Poisson model (Koenker and Mizera 2010).

More formally, let \( P(i) \) represent the Poisson mass at \( i \) and let

\( \ln Q \) be some concave density. Then, one could find the

log-density \( Q \) such that \( Q \) maximizes \( \sum_i P(i)Q(i) - \lambda \int \exp(Q(x))dx \) over the space of all concave \( Q \). The first term

corresponds to the log-likelihood, equivalent to the Kullback–Leibler divergence (Cover and Thomas 1991), and the second is a Lagrange multiplier to ensure that the density \( \exp(Q(x)) \) integrates to unity. This is a convex problem be-

cause the space of all concave \( Q \) is convex and the objective function is concave in \( Q \). In addition, it is easy to show that the optimal \( Q \) has to be piecewise linear; this means that one need not search over all possible densities, but rather, simply vary \( Q(i) \) at the integers. Note that \( \int \exp(Q(x))dx \) can be computed explicitly for any piecewise linear \( Q \). This optimization prob-

lem can be solved using simple interior point methods and, in

fact, the Hessian of the inner loop of the interior point method

will be banded (because enforcing concavity of \( Q \) is a local

Spatial filtering

Spatial filtering could be improved in a number of ways. For

instance, pairing this approach with a crude but automatic

segmentation tool to obtain ROIs would create a completely

automatic algorithm that converts raw movies of populations of

neurons into populations of spike trains. Furthermore, this filter

could be coupled with more sophisticated algorithms to initial-

ize the spatial filters when they are overlapping (for instance,

principal component analysis (Horn and Johnson 1990) or

independent component analysis (Mukamel et al. 2009)). One

could also use a more sophisticated model to estimate the

spatial filters. One option would be to assume a simple para-

metric form of the spatial filter for each neuron (e.g., a basis

set) and then merely estimate the parameters of that model.

Alternatively, one could regularize the spatial filters, using an

elastic net type approach (Grosenick et al. 2009; Zou and

Hastie 2005), to enforce both sparseness and smoothness.

Model generalizations

In this work, we made two simplifying assumptions that can

easily be relaxed: 1) instantaneous rise time of the fluorescence

transient after a spike and 2) constant background. In practice,

often either or both of these assumptions are inaccurate. Spe-

cifically, genetic sensors tend to have a much slower rise time

than that of organic dyes (Reiff et al. 2005). Further, the

background often exhibits slow baseline drift due to move-

ment, temperature fluctuations, laser power, and so forth, not to

mention bleaching, which is ubiquitous for long imaging ex-

periments. Both slow rise and baseline drift can be incorpo-

rated into our forward model using a straightforward general-

ization.

Consider the following illustrative example: the fluorescein

rise time in a particular data set is quite slow, much slower

than that of a single image frame. Thus fluorescence might be

well modeled as the difference of two different calcium extrusion

mechanisms, with different time constants. To model this

scenario, one might proceed as follows: posit the existence of

a two-dimensional time-varying signal, each like the calcium
APPENDIX A: PSEUDOCODE

Algorithm 1 Pseudocode for inferring the approximately most likely spike train, given fluorescence data. Note that the algorithm is robust to small variations $\xi_i$, $\xi_j$. The equations listed below refer to the most general equations in the text (simpler equations could be substituted when appropriate). Curly brackets, { }, indicate comments.

1: initialize parameters, $\theta$ (see Initializing the parameters in METHODS)
2: while convergence criteria not met do
3: for $z = 1, 0.1, 0.01, \ldots$, $\xi$ do {interior point method to find $\hat{C}$}
4: Initialize $n_t = \xi$, for all $t = 1, \ldots, T$, $C_i = 0$ and $C_i = \gamma C_{i-1} + n_t$ for all $t = 2, \ldots, T$
5: let $C_t$ be the initialized calcium, and $P_{C}$ be the posterior given this initialization
6: while $P_C < \hat{P}_C$ do {Newton–Raphson with backtracking line searches}
7: compute $g$ using Eq. 34
8: compute $H$ using Eq. 35
9: compute $d$ using $B_d$ {block-tridiagonal Gaussian elimination}
10: let $C_t = C_t + sd$, where $s$ is between 0 and 1, and $P_C > \hat{P}_C$ {backtracking line search}
11: end while
12: end for
13: check convergence criteria
14: update $\alpha$ and $\beta$ using Eq. 36 {only if spatial filtering}
15: let $s$ be the root-mean square of the residual
16: let $\lambda = T(\Delta \Sigma p)$
17: end while

signal assumed in the simpler models described earlier. Therefore each signal has a time constant and each signal is dependent on spiking. Finally, the fluorescence could be a weighted signal. Given these conventions, we have the following generalized model:

$$F_t = \alpha^T X_t + \beta + e, \quad e \sim \mathcal{N}(0, \sigma^2) \quad (41)$$

$$X_t = \Gamma X_{t-1} + An_t, \quad n_t \sim \text{Poisson} (\lambda \Delta) \quad (42)$$

Note that this model simplifies to the model proposed earlier when $d = 1$. Because $X$ is still Markov, all the theory developed above still applies directly for this model. There are, however, additional complexities with regard to identifiability. Specifically, the parameters $\alpha$ and $A$ are closely related. Thus we enforce that $A$ is a known binary vector, simply encoding whether a particular element responds to spiking. The matrix $\Gamma$ will not be uniquely identifiable, for the same reason that $\gamma$ was not identifiable, as described in Learning the parameters in METHODS. Thus we would assume $\Gamma$ was known, a priori. Note that other approaches to dealing with baseline drift are also possible, such as letting $\beta$ be a time-varying state: $\beta_t = \beta_{t-1} + e_{t}$, where $e_t$ is a normal random variable with variance $\sigma^2_{\beta}$ that sets the effective drift rate. Both these models are the subjects of further development.

Concluding thoughts

In summary, the model and algorithm proposed in this work potentially provide a useful tool to aid in the analysis of calcium-dependent fluorescence imaging and establish the groundwork for significant further development.

APPENDIX B: WIENER FILTER

The Poisson distribution in Eq. 4 can be replaced with a Gaussian instead of an exponential distribution, i.e., $n_t \sim \mathcal{N}(\lambda \Delta, \lambda \Delta)$ that, when plugged into Eq. 7, yields:

$$\hat{\alpha} = \argmax_{\alpha} \sum_{t=1}^{T} \left[ \frac{1}{2\sigma_t^2} (F_t - \alpha \beta_t - \beta_1)^2 + \frac{1}{2\lambda \Delta} (n_t - \lambda \Delta)^2 \right]$$

Note that since fluorescence integrates over $\Delta$, it makes sense that the mean scales with $\Delta$. Further, since the Gaussian here is approximating a Poisson with high rate (Sjulson and Miesenböck 2007), the variance should scale with the mean. Using the same tridiagonal trick as before, Eq. 11b can be solved using Newton–Raphson once (because this expression is quadratic in $n$). Writing the above in matrix notation, substituting $C_t - \gamma C_{t-1}$ for $n_t$, and letting $\alpha = 1$ yields:

$$\hat{C} = \argmax_{C} \sum_{t=1}^{T} \left[ \frac{1}{2\sigma_t^2} \| F_t - \beta_t - \beta_1 \| - \frac{1}{2\lambda \Delta} \|MC_t - \lambda \Delta C_1\|^2 \right], \quad (B2)$$

which is quadratic in $C$. The gradient and Hessian are given by:

$$g = \frac{1}{\sigma_t^2} (C - F - \beta_1) + \frac{1}{\lambda \Delta} (MC) M + \lambda \Delta M^2 1 \quad (B3)$$

$$H = \frac{1}{\sigma_t^2} I + \frac{1}{\lambda \Delta} M^2 M. \quad (B4)$$

Note that this solution is the optimal linear solution, under the assumption that spikes follow a Gaussian distribution, and is often referred to as the Wiener filter, regression with a smoothing prior, or ridge regression (Boyd and Vandenberghe 2004). Estimating the parameters for this model follows a pattern similar to that described in Learning the parameters in METHODS.

Acknowledgments

We thank V. Bonin for helpful discussions.

Grants

This work was supported by National Institute on Deafness and Other Communication Disorders Grant DC-00109 to J. T. Vogelstein; National Science Foundation (NSF) Faculty Early Career Development award, an Alfred P. Sloan Research Fellowship, and a McKnight Scholar Award to L.