Communicated by Scott Kirkpatrick

Bayesian Modeling and Classification of Neural Signals

Michael S. Lewicki
Computation and Neural Systems Program,
California Institute of Technology 216–76, Pasadena, CA 91125 USA

Identifying and classifying action potential shapes in extracellular neural waveforms have long been the subject of research, and although several algorithms for this purpose have been successfully applied, their use has been limited by some outstanding problems. The first is how to determine shapes of the action potentials in the waveform and, second, how to decide how many shapes are distinct. A harder problem is that action potentials frequently overlap making difficult both the determination of the shapes and the classification of the spikes. In this report, a solution to each of these problems is obtained by applying Bayesian probability theory. By defining a probabilistic model of the waveform, the probability of both the form and number of spike shapes can be quantified. In addition, this framework is used to obtain an efficient algorithm for the decomposition of arbitrarily complex overlap sequences. This algorithm can extract many times more information than previous methods and facilitates the extracellular investigation of neuronal classes and of interactions within neuronal circuits.

1 Introduction

Waveforms of extracellular neural recordings often contain action potentials (APs) from several different neurons. Each voltage spike in the waveform shown in Figure 1 is the result of APs from one or more neurons. An individual AP typically has a fast positive component and a fast negative component and may have additional slower components depending on the type of neuron and where the electrode is positioned with respect to the cell. Determining what cell fired when is a difficult, ill-posed problem and is compounded by the fact that cells frequently spike simultaneously, which results in large variations in the observed shapes.

Identifying and classifying the APs in a waveform, which is commonly referred to as "spike sorting," have three major difficulties. The first is determining the AP shapes, the second is deciding the number of distinct shapes, and the third is decomposing overlapping spikes into their component parts. In general, these cannot be solved independently since the solution of one will affect the solution of the others. Algorithms

 © 1994 Massachusetts Institute of Technology

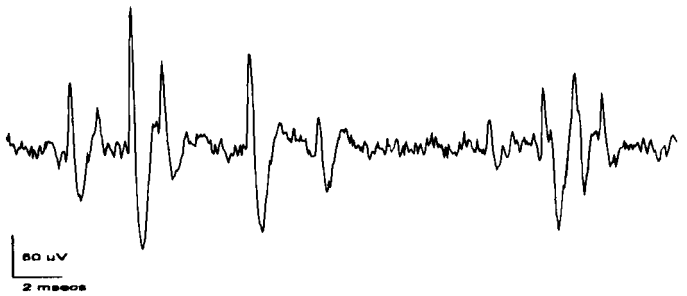

Figure 1: The extracellular waveform shows several different action potentials (APs) generated by an unknown number of neurons. Note the frequent presence of overlapping APs, which can, in the case of the right-most group, completely obscure individual spikes. The waveform was recorded with a glass-coated platinum iridium electrode in zebra finch nucleus lMAN (courtesy of Allison Doupe, Caltech).

for identifying and classifying APs (see Schmidt 1984 for a review) fall into two main categories: feature clustering and template matching.

Feature clustering involves describing features of APs, such as the peak value, spike width, slope, etc., and using a clustering algorithm to determine distinct classes in the set of features. Using a small set of features, although computationally efficient, is often sufficient only to discriminate the cells with the largest APs. Increasing the number of features in the clustering often yields better discrimination, but there still remains the problem of how to choose the features, and it is difficult with such techniques to handle overlapping spikes.

In template matching algorithms, typical action potential shapes are determined, either by an automatic process or by the user. The waveform is then scanned and each event classified according to how well it fits each template. Template matching algorithms are better suited for classifying overlaps since some underlying APs can be correctly classified if the template is subtracted from the waveform each time a fit is found. The main difficulty in template matching algorithms is in choosing the templates and in decomposing complex overlap sequences.

The approach demonstrated in this paper is to model the waveform directly, obtaining a probabilistic description of each action potential and, in turn, of the whole waveform. This method allows us to compute the class conditional probabilities of each AP, which quantifies the certainty with which an AP is assigned to a given class. In addition, it will be possible to quantify the certainty of both the form and number of spike

shapes. Finally, we can use this description to decompose overlapping APs efficiently and to assign probabilities to alternative spike model sequences.

2 Modeling Action Potentials

First we consider the problem of fitting a model to events from a single cell. Let us assume that the data from the event we observe (at time zero) is a result of a fixed underlying spike function, $s(t)$, plus noise:

$$d_i = s(t_i) + \eta_i \tag{2.1}$$

A computationally convenient form for $s(t)$ is a continuous piece-wise linear function:

$$s(t) = y_j + \frac{v_j}{h}(t - x_j), \qquad x_j \le t < x_{j+1} \tag{2.2}$$

where $h = x_{j+1} - x_j$, $j = 1 \ldots R$, and $v_j = y_{j+1} - y_j$. We will treat R and the x_js as known. The noise, η, is modeled as gaussian with zero mean and standard deviation σ_η.

2.1 The Posterior for the Model Parameters. From the Bayesian perspective, the task is to infer the posterior distribution of the parameters, $\mathbf{v} = \{v_1, \ldots, v_R\}$, given the data from the observed events, D, and our prior assumptions of the spike model, M. Applying Bayes' rule we have

$$P(\mathbf{v} \mid D, \sigma_\eta, \sigma_w, M) = \frac{P(D \mid \mathbf{v}, \sigma_\eta, M)\, P(\mathbf{v} \mid \sigma_w, M)}{P(D \mid \sigma_\eta, \sigma_w, M)} \tag{2.3}$$

$P(D \mid \mathbf{v}, \sigma_\eta, M)$ is the probability of the data for the model given in (2.2) and is assumed to be gaussian:

$$P(D \mid \mathbf{v}, \sigma_\eta, M) = \frac{1}{Z_D(\sigma_\eta)} \exp\left\{ -\frac{1}{2\sigma_\eta^2} \sum_{i=1}^{I} [d_i - s(t_i)]^2 \right\} \tag{2.4}$$

where $Z_D(\sigma_\eta) = 1/(2\pi\sigma_\eta^2)^{I/2}$. The time of the ith data point, d_i, is taken to be relative to the corresponding event, i.e., $t_i = t_i^{(n)} - \tau^{(n)}$. By convention, $\tau^{(n)}$ is the time of the inferred AP peak. The data range over the predetermined extent of the action potential.[1]

$P(\mathbf{v} \mid \sigma_w, M)$ specifies prior assumptions of the structure of $s(t)$. Ideally, we want a distribution over $\mathbf{v}$ from which typical samples result only in shapes that are plausible APs. Conversely, this space should not be so

[1]For the examples shown here, this range is from 1 msec before the spike peak to 4 msec after the peak.

restrictive that legitimate AP shapes are excluded. We adopt a simple approach and use a prior of the form

$$P(s(t) \mid \sigma_w, M) \propto \exp\left[-\int du\, s^{(m)}(u)^2/\sigma_w^2\right] \tag{2.5}$$

where the superscript (m) denotes differentiation. $m = 1$ corresponds to linear splines, $m = 2$ corresponds to cubic splines, etc. The smoothness of $s(t)$ is controlled through the parameter σ_w with small values of σ_w penalizing large fluctuations. A prior simply favoring smoothness ensures minimal restrictions on the kinds of functions we can interpolate, but it does not buy us anything either. If we had a more informative prior, we would require less data to reach the same conclusions about the form of $s(t)$. Any reasonable prior should have little effect on the shape of the final spike function if there are abundant data. Even though the prior may have little effect on the shape, it still plays an important role in model comparison which will be discussed in Section 4.

The components of the posterior distribution for $\mathbf{v}$ are now defined. There still remains, however, the problem of determining σ_η and σ_w. An exact Bayesian analysis requires that we eliminate the dependence of the posterior on σ_η and σ_w by integrating them out:

$$P(\mathbf{v} \mid D, M) = \int d\sigma_\eta\, d\sigma_w\, P(\mathbf{v} \mid D, \sigma_\eta, \sigma_w, M)\, P(\sigma_w, \sigma_\eta \mid M) \tag{2.6}$$

In this paper, we use the approximation $P(\mathbf{v} \mid D, M) \approx P(\mathbf{v} \mid D, \sigma_w^{\text{MP}}, \sigma_\eta^{\text{MP}}, M)$. The most probable values of $\mathbf{v}$, σ_w, and σ_η were obtained using the methods of MacKay (1992), which we briefly summarize here. First, we transform $\mathbf{v}$ to a basis in which the Hessian of $\log P(\mathbf{v} \mid \sigma_w, M)$ is the identity. For splines, this is the Fourier representation:

$$\begin{aligned} s(t) &= a_0 + \sum_{j=1}^{(R/2)-1} \left(a_j \frac{\sqrt{2}\cos 2\pi jt}{(2\pi j)^m} + b_j \frac{\sqrt{2}\sin 2\pi jt}{(2\pi j)^m} \right) \\ &\quad + a_{R/2} \frac{\cos 2\pi Rt}{(\pi R)^m} \end{aligned} \tag{2.7}$$

using the prior

$$P(\mathbf{w} \mid \sigma_w, M) = \frac{1}{Z_W(\sigma_w)} \exp\left[-\frac{1}{2\sigma_w^2}\sum_r w_r^2\right] \tag{2.8}$$

where $\mathbf{w} = \{\mathbf{a}, \mathbf{b}\} - a_0$. The term a_0 is set to the known DC level (the offset of the A/D converters). In the limit $R \to \infty$, $\sum_r w_r^2/2 = \int_0^1 [s^{(m)}(u)]^2\, du$ (Wahba 1990), which is the splines regularizer. We take $m = 1$ for linear splines.

The most probable parameter values, $\mathbf{w}^{\text{MP}}$, were determined as follows. Let $E_D = \sum_i [d_i - s(t_i)]^2/2$ and $E_W = \sum_r w_r^2/2$. Letting $\mathbf{B} = \nabla\nabla E_D$ and $\mathbf{C} = \nabla\nabla E_W$ (around $\mathbf{v}^{\text{ML}}$), we obtain $\mathbf{w}^{\text{MP}} = \sigma_\eta^{-2}\mathbf{A}^{-1}\mathbf{B}\mathbf{w}^{\text{ML}}$, where

$\mathbf{A} = \sigma_w^{-2}\mathbf{C} + \sigma_\eta^{-2}\mathbf{B}$. The maximum likelihood values, $\mathbf{v}^{\mathrm{ML}}$, can be determined efficiently by inverting a tridiagonal matrix. The Fourier coefficients can be computed efficiently with the fast Fourier transform.

The most probable values of σ_η and σ_w were obtained using the reestimation formulas $\sigma_\eta^2 = 2E_D/(I-\gamma)$ and $\sigma_w^2 = E_W/\gamma$, where $\gamma = \sum \lambda_r/(\lambda_r + \sigma_w^{-2})$ and λ_r is the rth eigenvalue of $\sigma_\eta^{-2}\mathbf{B}$. In terms of $\boldsymbol{\lambda}$, $w_r^{\mathrm{MP}} = \lambda_r w_r^{\mathrm{ML}}/(\lambda_r + \sigma_w^{-2})$.

Note that we could at this point apply the methods described by MacKay (1992) and discussed later on in Section 4 to compare alternative spike models, in essence to determine the *most probable* spike model given the data. For example, we might choose cubic splines instead of piecewise linear functions or choose priors that better represented our knowledge about spike shapes. The piece-wise linear spike models discussed here can be made to fit any fixed shape, since they can contain arbitrarily many segments. With 75 segments, the spike models have been descriptively sufficient for the all the data we have observed. Situations for which this is not the case will be discussed in Section 9. Figure 2a shows the result of fitting one spike model to data consisting of 40 APs.

2.2 Checking the Assumptions. Before proceeding to the more complicated cases of multiple spike models and overlapping spikes, we must check our assumptions on real data. Equation (2.1) assumes that the noise process is invariant throughout the duration of the AP, but in principle this need not be the case. For example, the noise might show larger variation at the extremes. The spike model residuals, $\eta_i = d_i - s(t_i)$, shown in Figure 2a, give no indication of an amplitude-dependent noise process.

A second assumption we have made is that the noise is gaussian. Figure 2b shows a gaussian distribution with the inferred width σ_η overlaid on a normalized histogram of the residuals from Figure 2a. The most significant deviation is in the tails of the distribution, which reflects the presence of overlapping spikes. In this case, the overlaps are evenly distributed over the range of the fitted event so they have little effect on the model's form in the limit of large amounts data. The model would be poorly inferred, however, if the overlaps were not uniformly distributed over the interval, for example if one cell tended to fire within a few milliseconds of another. This is a common problem in practice and will be addressed in Section 5.

An assumption that has not been tested is whether the residuals are independent. Figure 2c and d shows that the noise in these data is slightly correlated. This has little effect on the fit of the models but does affect the accuracy of the probabilities discussed in the later sections. A convenient way of reducing the correlation is to sample close to the Nyquist rate to avoid correlation introduced by the amplifier filters.

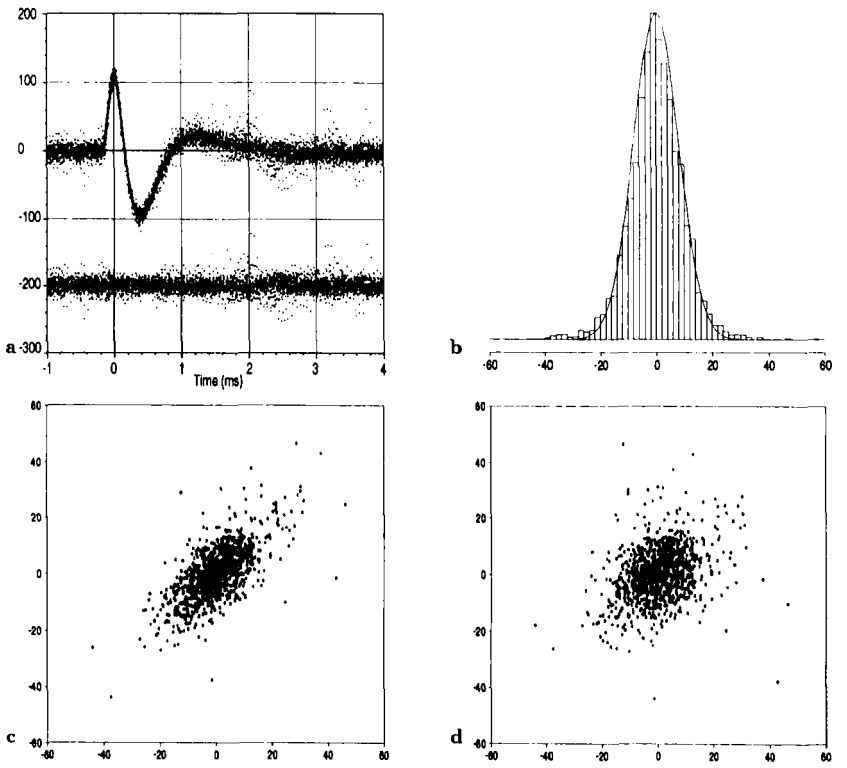

Figure 2: (a) Spike model fit to data consisting of 40 APs. The solid line is a 75 segment piece-wise linear model. Each AP is aligned with respect to the inferred spike peak. Each dot is one sample point. The residual error for each sample, $\eta_i = d_i - s(t_i)$, is offset by -200 μV and plotted below. The flat residuals indicate that the data is well-fit by the model. (b) Normalized histogram of the residuals from a. The curve is the gaussian inferred with the methods discussed in the text. The outliers result from overlapping APs, which can be seen in the data in a. (c and d) Lagged scatter plot of a sample of the residuals in a. (c) η_i vs η_{i+1}. (d) η_i vs η_{i+2}. These graphs indicate that there is some correlation between η_i and η_{i+1} (c), but little between η_i and η_{i+2} (d). This is expected for these data because the sampling rate (20 kHz) was higher than the Nyquist rate (14 kHz).

3 Multiple Spike Shapes

When a waveform contains multiple types of APs, determining the spike shapes is more difficult because the classes are not known a priori. We cannot infer the parameters for one spike model if we don't know what data is representative of its class. Furthermore, if two spike models are similar, it is possible that an observed event could have come from either class with equal probability. The uncertainty of which class an event belongs to can be incorporated with a mixture distribution (Duda and Hart 1973).

The probability of a particular event, $\mathbf{D}_n$, given all spike models, $M_{1:K}$, is

$$P(\mathbf{D}_n \mid \mathbf{v}_{1:K}, \boldsymbol{\pi}, \sigma_\eta, M_{1:K}) = \sum_{k=1}^{K} \pi_k P(\mathbf{D}_n \mid \mathbf{v}_k, \sigma_\eta, M_k) \tag{3.1}$$

where π_k is the a priori probability that a spike will be an instance of M_k ($\sum \pi_k = 1$). The joint probability for $\mathbf{D}_{1:N} = \{\mathbf{D}_1 \ldots \mathbf{D}_N\}$ is simply the product

$$\mathcal{L} = P(\mathbf{D}_{1:N} \mid \mathbf{v}_{1:K}, \boldsymbol{\pi}, \sigma_\eta, M_{1:K}) = \prod_{n=1}^{N} P(\mathbf{D}_n \mid \mathbf{v}_{1:K}, \boldsymbol{\pi}, \sigma_\eta, M_{1:K}) \tag{3.2}$$

The posterior for multiple spike models is then

$$\begin{aligned} &P(\mathbf{v}_{1:K}, \boldsymbol{\pi} \mid \mathbf{D}_{1:N}, \sigma_\eta, \boldsymbol{\sigma}_w, M_{1:K}) \\ &\quad = \frac{P(\mathbf{D}_{1:N} \mid \mathbf{v}_{1:K}, \boldsymbol{\pi}, \sigma_\eta, M_{1:K})\, P(\mathbf{v}_{1:K} \mid \boldsymbol{\sigma}_w, M_{1:K})\, P(\boldsymbol{\pi} \mid M_{1:K})}{P(\mathbf{D}_n \mid \sigma_\eta, \boldsymbol{\sigma}_w, M_{1:K})} \end{aligned} \tag{3.3}$$

We use $P(\mathbf{v}_{1:K} \mid \boldsymbol{\sigma}_w, M_{1:K}) = \prod_k P(\mathbf{v}_k \mid \sigma_{wk}, M_k)$ and take $P(\boldsymbol{\pi} \mid M_{1:K})$ to be flat over $[0,1]^K$ subject to the constraint $\sum_k \pi_k = 1$.

Note that we have implicitly assumed that the spike occurrence times are Poisson in nature with mean firing rates proportional to π_k. This assumes as little as possible about the temporal structure of the spikes. A more powerful description, for example, modeling the distribution of the interspike interval, would be obtained by incorporating this information into 3.2.

3.1 Maximizing the Posterior. We proceed as before to find the maxima of the posterior that will give us the most probable values for the whole set of spike models. The conditions satisfied at the maxima of $\mathcal{L}$ given in 3.2 are obtained by differentiating $\log \mathcal{L}$ with respect to $\mathbf{v}_k$ and equating the result to zero,

$$\begin{aligned} \frac{\partial \log \mathcal{L}}{\partial \mathbf{v}_k} &= \sum_{n=1}^{N} P(M_k \mid \mathbf{D}_n, \mathbf{v}_k, \boldsymbol{\pi}, \sigma_\eta) \frac{1}{\sigma_\eta^2} \\ &\quad \times \sum_i \left[d_{n,i} - s_k(t_i - \tau_n\,; \mathbf{v}_k) \right] \frac{\partial s_k(t_i\,; \mathbf{v}_k)}{\partial \mathbf{v}_k} \\ &= 0 \end{aligned} \tag{3.4}$$

where τ_n is the occurrence time of $\mathbf{D}_n$. Thus we obtain a soft clustering procedure in which the error for each event, $\mathbf{D}_n$, is weighted by the probability that it is an instance of M_k:

$$P(M_k \mid \mathbf{D}_n, \mathbf{v}_k, \boldsymbol{\pi}, \sigma_\eta) = \frac{\pi_k P(\mathbf{D}_n \mid \mathbf{v}_k, \sigma_\eta, M_k)}{\sum_k \pi_k P(\mathbf{D}_n \mid \mathbf{v}_k, \sigma_\eta, M_k)} \tag{3.5}$$

Although 3.4 can be solved exactly, it is still expensive to compute, because it uses all of the data. We adopt the approach of estimating each $\mathbf{v}_k$ by fitting each model to a reduced event list allowing the possibility of an event being in the lists of multiple models. These lists are obtained by sampling events from the whole data set and including an event in a model's reduced event list with probability proportional to $P(M_k \mid \mathbf{D}_n, \mathbf{v}_k, \boldsymbol{\pi}, \sigma_\eta)$. We apply the techniques used in the previous section to determine the values for $\boldsymbol{\sigma}_w$ and in turn the most probable values of $\mathbf{v}_{1:K}$.

Differentiating 3.2 and finding the condition satisfied at the maximum, we obtain the re-estimation formula

$$\pi_k = \frac{1}{N} \sum_n P(M_k \mid \mathbf{D}_n, \mathbf{v}_k, \boldsymbol{\pi}, \sigma_\eta) \tag{3.6}$$

For each model, σ_η can be estimated using the methods of the previous section. The mixture model estimate for σ_η is obtained by a weighted average of the individual estimates using weight π_k.

3.2 Selecting Events from the Data. For these demonstrations, any peak in the waveform that deviated from DC level by more than 4 times the estimated RMS noise level was labeled as an event, $\mathbf{D}_n$. Once an event is located, it is important to obtain accurate estimates of the occurrence time (with each spike model) by maximizing 2.4 over τ_n. For the largest models, deviations from the optimal value as little as one-tenth the sampling period will introduce misfit errors greater than σ_η. The τ_n's must be re-estimated as the spike models change for optimal results. An efficient way to perform this optimization is to use the *k–d* trees discussed in Section 5.

3.3 Initial Conditions. Since the re-estimation formulas derived here will find *local* maxima, it is critical to use good initial conditions for the spike models. Poor fits will result if there are too few spike models representing what are in fact several distinct APs. Conversely, if there are more spike models than distinct APs, not only will there be excess computational overhead, but there is no guarantee that each AP will be represented, since some spike functions may converge to represent the same AP class. Ideally, we want all potential spike shapes to be represented in the initial spike function set, $s_{1:K}(t)$. One approach toward obtaining an even representation of the AP shapes is to initialize each

spike function to single events so that $\max_t s(t) - \min_t s(t)$ is evenly distributed with a separation proportional to the estimated waveform RMS noise. This approach works well for present purposes, because the height of an AP captures much of the variability among classes. By erring on the side of starting with too many spike models, we can obtain a good initial representation of the AP shapes. There is still a need to decide if two different models should be combined and if one class should be split into two. How to choose the number of spike models objectively will be demonstrated in the next section.

4 Determining the Number of Spike Models

If we were to choose a set of spike models that best fit the data, we would wind up with a model for each event in the waveform. We might think of heuristics which would tell us when two spike models are distinct and when they are not, but ad hoc criteria are notoriously dependent on particular circumstances, and it is difficult to state precisely what information the rules take into account. A solution to this dilemma is provided by probability theory (Jeffreys 1939; Jaynes 1979; Gull 1988).

To determine the *most probable* number of spike models, we need to derive the probability of a set of spike models, denoted by $S_j = \{M_{1:K}^{(j)}\}$, conditioned only on the data and information known a priori, which we denote by H. From Bayes' rule, we obtain

$$P(S_j \mid \mathbf{D}_{1:N}, H) = \frac{P(S_j \mid H) P(\mathbf{D}_{1:N} \mid S_j, H)}{P(\mathbf{D}_{1:N} \mid H)} \tag{4.1}$$

The only data dependent term is $P(\mathbf{D}_{1:N} \mid S_j, H)$, which is called the *evidence* for S_j. If we assume all the hypotheses $S_{1:J}$ under consideration are equally probable, $P(\mathbf{D}_{1:N} \mid S_j, H)$ ranks alternative spike sets, since it is proportional to $P(S_j \mid \mathbf{D}_{1:N}, H)$. With equal priors, the ratio $P(\mathbf{D} \mid S_i, H)/P(\mathbf{D} \mid S_j, H)$ is equal to the Bayes factor in favor of hypothesis S_i over hypothesis S_j, which is the standard way to compare hypotheses in the Bayesian literature.

The evidence for S_j is obtained by integrating out the nuisance parameters in 3.3:

$$\begin{aligned} P(\mathbf{D}_{1:N} \mid S_j) &= \int d\mathbf{v}_{1:K} d\boldsymbol{\pi} d\sigma_\eta d\boldsymbol{\sigma}_w \, P(\mathbf{D}_{1:N} \mid \mathbf{v}_{1:K}, \boldsymbol{\pi}, \sigma_\eta, S_j) \\ &\quad \times P(\mathbf{v}_{1:K} \mid \boldsymbol{\sigma}_w, S_j) \, P(\boldsymbol{\pi} \mid S_j) \, P(\sigma_\eta, \boldsymbol{\sigma}_w \mid S_j) \end{aligned} \tag{4.2}$$

This integral is analytically intractable, but it is often well-approximated with a gaussian integral which for a function $f(\mathbf{w})$ is given by

$$\int d\mathbf{w} f(\mathbf{w}) \approx f(\hat{\mathbf{w}}) \, (2\pi)^{d/2} \, |-\nabla\nabla \log f(\mathbf{w})|^{-1/2} \tag{4.3}$$

where d is dimension of $\mathbf{w}$, $\hat{\mathbf{w}}$ is a (local) maximum of $f(\mathbf{w})$, $|\mathbf{A}|$ denotes the determinant of $\mathbf{A}$, and the derivatives are evaluated at $\hat{\mathbf{w}}$. With this

we obtain the evidence for spike set S_j,

$$\begin{aligned} P(\mathbf{D}_{1:N} \mid S_j, H) &= P(\mathbf{D}_{1:N} \mid \hat{\mathbf{v}}_{1:K}, \hat{\boldsymbol{\pi}}, \hat{\sigma}_\eta, S_j)\, P(\hat{\mathbf{v}}_{1:K} \mid \boldsymbol{\sigma}_w, S_j) \\ &\times P(\hat{\boldsymbol{\pi}} \mid S_j)\, P(\hat{\boldsymbol{\sigma}}_w, \hat{\sigma}_\eta \mid S_j) \\ &\times (2\pi)^{d/2} \left| -\nabla\nabla \log P(\mathbf{D}_{1:N} \mid \mathbf{v}_{1:K}, \boldsymbol{\pi}, \sigma_\eta, S_j) \right|^{-1/2} \\ &\times \Delta \log \hat{\boldsymbol{\sigma}}_w \Delta \log \hat{\sigma}_\eta \end{aligned} \tag{4.4}$$

where $\Delta \log \hat{\boldsymbol{\sigma}}_w = \prod_k \sqrt{2/\gamma_k}$, $\Delta \log \hat{\sigma}_\eta = \sqrt{2/(NI - \gamma)}$, and $d = KR + K + 1$. γ_k is the number of good degrees of freedom for M_k (MacKay 1992), which can be thought of as the number of parameters that are well-determined by the data. $\gamma = \sum_k \gamma_k$. $P(\boldsymbol{\sigma}_w, \sigma_\eta \mid S_j)$ is assumed to be separable and flat over $\log \sigma_w$ and $\log \sigma_\eta$. Since the labeling of the models is arbitrary, an additional factor of $1/K!$ must be included to estimate the posterior volume accurately. The Hessian $-\nabla\nabla \log P(\mathbf{D}_{1:N} \mid \mathbf{v}_{1:K}, \boldsymbol{\pi}, \sigma_\eta, S_j)$ (with respect to $\mathbf{v}_{1:K}$ and $\boldsymbol{\pi}$) was evaluated both analytically and using a diagonal approximation. Both methods produced similar results, and the latter, being much faster to compute, was used for these demonstrations. Notice that the approximation for the evidence decomposes into the best-fit likelihood for the best fit parameters times the other terms, which collectively constitute a complexity penalty called the Ockham factor (MacKay 1992). Since this factor is the ratio of the posterior accessible volume in parameter space to the prior accessible volume, it is smaller for more complicated models. Overly broad priors will introduce a bias toward simpler models. Unless the best-fit likelihood for complex models is sufficiently larger than the likelihood for simple ones, the simple models will be more probable.

A convenient way of collapsing the spike set is to compare spike models pairwise. Two models in the spike set are selected along with a sampled set of events fit by each model. We then evaluate $P(\mathbf{D} \mid S_1)$ and $P(\mathbf{D} \mid S_2)$. S_1 is the hypothesis that the data are modeled by a single spike shape, S_2 says there are two spike shapes. Included in the list of spike models should be a "null" model which is simply a flat line at DC. This hypothesis says that there are no events and that the data are a result of only the noise. Examples of this comparison are illustrated in Figure 3. If $P(\mathbf{D} \mid S_1) > P(\mathbf{D} \mid S_2)$, we replace both models in S_2 by the one in S_1. The procedure terminates when no more pairs can be combined to increase the evidence.

5 Decomposing Overlapping Events

The method of inferring the spike models we have discussed thus far is valid if the event occurrence times can be accurately determined and if the noise is gaussian and stationary. Often these conditions cannot be met without identifying and decomposing overlapping events. Even if

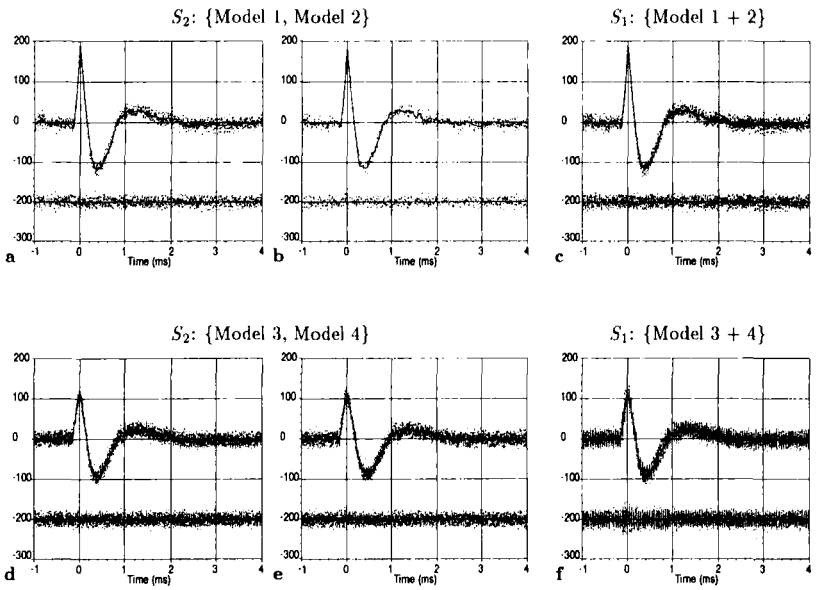

Figure 3: The most probable number of distinct spike models is determined by evaluating the evidence for alternative hypotheses for a given set of data. Simple hypotheses are generated by selecting similar shapes in a spike set. S_2 is the hypothesis that there are two distinct spike models; the fits of two such models two a sampled set of data are shown in a and b. S_1 is the hypothesis that there is only one spike model; the fit of this model is shown in c. In this case, even though the total misfit is less for S_2, the simpler hypothesis, S_1, is more probable by exp(111) to 1. In the second row, S_2 (d and e) is more probable than S_1 (f) by exp(343) to 1. Note the increase in residual error with the model shown in f. The difference between models 3 and 4 is better illustrated in Figure 8 (where they are labeled M_2 and M_3, respectively). The large log probability ratios reported here result mainly from the abundance of data and the nongaussian outliers in the noise. A more realistic noise model, such as heavy-tailed gaussian, would result in more accurate estimates of the true probability ratios.

the spike models are good, overlap decomposition is necessary to detect and classify individual events with accuracy.

For a given sequence of overlapping APs, there are potentially many spike model sequences that could account for the same data. An example is shown in Figure 4. We can calculate the probability of each alternative,

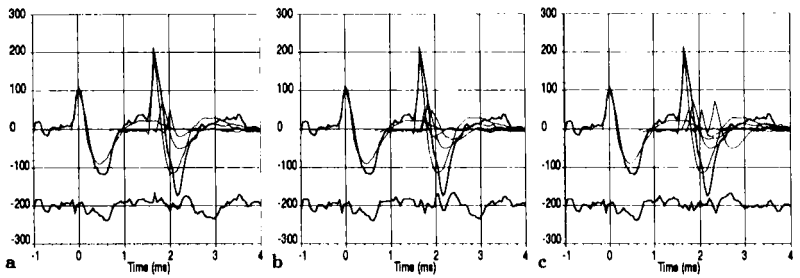

Figure 4: Overfitting also occurs in the case of decomposing overlapping events. Shown are three of many well-fitting solutions for a single region of data. Thick lines are drawn between the data samples. The thin lines are the spike functions (note that these examples were taken from the first iteration of the algorithm, so the spike functions are noisy estimates of the underlying AP shapes). The best-fitting overlap solution in this case is not the most probable: the solution with four spike functions shown in a is more than eight times more probable than either b (five spike functions) or c (six spike functions) even though these fit the data better. The simple approach of using the *best-fitting* overlap solution actually *increases* the classification error especially in the number of false positives for the smaller models. To minimize classification error, it is necessary to find the *most probable* overlap solution.

but there are an enormous number of sequences to consider, not only all possible models for each event but also all possible event times. A brute-force approach to this problem is to perform an exhaustive search of the space of overlapping spike functions and event times to find the sequence with maximum probability. This approach was used by Atiya (1992) in the case of two overlapping spikes with the times optimized to one sample period. Unfortunately, for many realistic situations this method is computationally too demanding even for off-line analysis. For overlap decomposition to be practical, we need an efficient way to fit and rank a large number of model potential spike sequences. In addition, we would like to state precisely what hypothesis subspace is searched, so we can say what model combinations *cannot* account for a given region of overlapping events.

We can obtain a more efficient decomposition algorithm by employing

Figure 5: As the peaks of two action potentials get closer together, it becomes more difficult to classify either one with accuracy. It is necessary in this case (b and c) to fit multiple models simultaneously.

two techniques. The first is to consider only AP sequences that occur with non-negligible probability. This allows us to obtain a large, but manageable hypothesis space in which to search. The second is to make the search itself efficient using appropriate data structures and dynamic programming.

5.1 Restricting the Overlap Hypothesis Space. The main difficulty with overlapping APs is that there is no simple way to determine the event times. For many overlaps, such as the one in Figure 5a, the event times can be determined directly, because the APs are separated enough so that the models can be fit independently. As the degree of overlap increases, as in Figure 5b and c, accurate classification of one event depends on accurate classification of the surrounding events. In this case, the overlapping models must be fit simultaneously. Moreover, since small misalignments of the model with respect to the event can introduce significant residual error, each model in the overlap sequence must be precisely aligned.

The continuum of possible event times is the major factor contributing to the multitude of potential overlap models. We can reduce this space significantly if we consider to what precision the τ_n's must be optimized. For a given spike model, $s_k(t)$, the maximum error resulting from a misalignment of δ_k is given by[2]

$$\epsilon = \delta_k \max_t \left| \frac{ds(t)}{dt} \right| \tag{5.1}$$

From this we obtain the precision necessary to ensure that the error introduced by the model alone is less than ϵ and only need to choose among a discrete set of points.[3]

[2]We ignore the discontinuities in the derivative of the piece-wise linear model.

[3]For these demonstrations we use $\epsilon = 0.5\sigma_\eta$, which results in δ_ks ranging from 0.05 to 0.3 sampling periods.

Even with this reduction, the number of possible sequences is still exponential in the number of overlapping models. This space can be reduced by considering only sequences that are likely to occur. For example, if there are five units with a Poisson firing rate of 20 Hz, the probability of observing three events within 0.5 msec is about 0.1%. Eliminating sequence models with more than two peaks within 0.5 msec of each other will introduce about 0.1% error. In this manner, the desired trade-off between classification accuracy and computational cost can be determined. In practice, however, spikes often do not fire in a Poisson manner but fire in bursts. The firing rate model in this case should be adapted accordingly so that the expected number of missed events is estimated accurately.

5.2 Searching the Overlap Hypothesis Space. Let us first outline the decomposition algorithm. To fit general model sequences, we use the methodology of dynamic programming. The event data is fit in sections from left to right. At every stage, a list is maintained of all plausible sequences[4] from the restricted hypothesis space determined by the methods described above. The length of data fit is extended by computing for each sequence on the list all plausible models that result by fitting the residual structure in the next region. The probabilities for all sequences are then recomputed, discarding any sequences below the probability threshold. The search terminates when no further overlaps are encountered in the most probable sequence model.

We now discuss each step in more detail. The primary operation in the algorithm is that of determining the most probable sequence models for a region of data. For efficiency, we precompute all possible waveform segments and store the set in a *k–d* tree (Bently 1975) with which a fixed-radius nearest-neighbor search can be performed in time logarithmic in the number of models (Friedman *et al.* 1977; Ramasubramanian and Paliwal 1992). $O(N \log N)$ time is required to construct the tree, but once it is set up, each nearest-neighbor search is very fast. The set of overlap functions for a region from a to b around the spike peak is defined by

$$\Lambda_{k_{1:L},n}(t) = \sum_{j=1}^{L} s_{k_j}(t - n\delta_{k_j}), \quad k_j = 1, \ldots, K, \quad k_1 < \cdots < k_L, \quad a < t - n\delta_{k_j} < b, \quad n \text{ integer} \tag{5.2}$$

where L is the maximum number of overlapping spike function segments in the peak region $[a, b]$, and δ_{k_j} is the τ-resolution for $s_k(t)$ defined in 5.1. The size of the peak region is somewhat arbitrary; the larger the region, the larger the number of waveform segments that must be considered, but the smaller the number of plausible overlap sequences found. In practice, the size of the peak region is largely limited by the memory

[4]By plausible sequences we mean sequences with probability greater than a specified threshold.

required for the *k–d* tree. For these demonstrations, we take $L = 2$ (up to two overlapping spike function segments within a peak region) with a peak region of 0.25 msec and include a "noise" model Λ_0 that has constant value equal to the DC voltage level. The number of waveform segments in the set can be reduced by eliminating overlapping spike functions for which the peak would have been (with high probability) detected at a sample position other than that of the data. Even with this reduction, an 11-model spike set results in about 50,000 waveform segments.

Once the best-fitting waveform segments for the first peak region are obtained, each segment is extended up to the next peak in the residuals for that segment. This peak is then fit using the *k–d* tree, which in turn generates additional overlap sequences. As long as the introduction of new waveform segments does not alter our conclusions about the ordering sequence list, for example, by fitting structure in a preceding region, we ensure either that one of the overlap sequences is true or that the sequences we are considering cannot account for the data.

After each sequence from the original list has been extended, the probability of each sequence model, $\mathbf{c}_i$, is recomputed. The exact relation is given by

$$P(\mathbf{c}_i \mid \mathbf{D}, \mathbf{S}) = \int d\boldsymbol{\tau}_i \frac{P(\mathbf{D} \mid \mathbf{c}_i, \boldsymbol{\tau}_i, \mathbf{S}) P(\mathbf{c}_i, \boldsymbol{\tau}_i \mid \mathbf{S})}{P(\mathbf{D} \mid \mathbf{S})} \tag{5.3}$$

where $\mathbf{D}$ is the subset of data common to all sequences, and $\mathbf{S} = \{\mathbf{v}_{1:K}, \boldsymbol{\pi}, \sigma_\eta, M_{1:K}\}$. The form of the probability density function, $P(\mathbf{D} \mid \mathbf{c}_i, \boldsymbol{\tau}_i)$, is the same as 2.4. Equation 5.3 can be approximated with a gaussian integral by treating each peak region as a separable component,

$$P(\mathbf{c}_i \mid \mathbf{D}, \mathbf{S}) \approx \frac{P(\mathbf{D} \mid \mathbf{c}_i, \hat{\boldsymbol{\tau}}_i, \mathbf{S})\,(2\pi)^{C/2} \prod_j d_j^{-1/2} P(\mathbf{c}_i \mid \mathbf{S})\, P(\boldsymbol{\tau}_i \mid \mathbf{S})}{P(\mathbf{D} \mid \mathbf{S})}, \tag{5.4}$$

where C is the number of total number of spike functions in the sequence, and d_j is the determinant of Hessian of the τ's for the jth peak region. The values needed to compute the Hessians can be obtained directly from the *k–d* tree. Note that integrating over $\boldsymbol{\tau}_i$ performs the function of Ockham's Razor by penalizing sequences with many spike models. Omitting this would reduce the solution to one of maximum likelihood, which chooses the sequence that best fits the data. For example, the solutions shown in Figure 4b and c both fit the data better than in 4a, but by 5.4, 4a is more than eight times more probable than the others. Use of the best-fitting solutions would result in an *increase* in the classification error due to the introduction of too many models. Classification error is minimized by using the most probable overlap sequences.

$P(\mathbf{c}_i, \boldsymbol{\tau}_i \mid \mathbf{S})$ describes the a priori probability of the sequence of models in $\mathbf{c}_i$ with associated occurrence times $\boldsymbol{\tau}_i$. For this discussion, we assume $P(\mathbf{c}_i \mid \mathbf{S})$ to be Poisson with rate proportional to $\langle \pi_k \rangle$ and $P(\boldsymbol{\tau}_i \mid \mathbf{S})$ to be proportional to $1/\langle \pi_k \rangle$. Useful alternatives for $P(\mathbf{c}_i, \boldsymbol{\tau}_i \mid \mathbf{S})$ include models

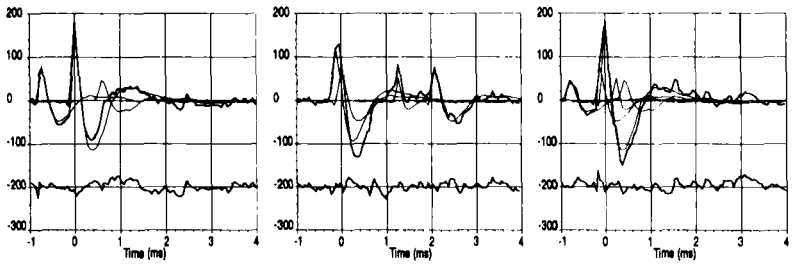

Figure 6: Example overlap solutions. Thick lines are drawn between the data samples. The thin lines are the spike models. The overlap sequence in (a) has three spike functions, (b) contains four spike functions, and (c) contains five spike functions.

which take into account a refractory period or describe different types of spiking patterns.

Once the probabilities for the sequence models have been computed, the improbable models are discarded. The decomposition algorithm iterates until no overlapping structure is found in the most probable model. The search can fail if an outlier is encountered or if the true sequence is outside the hypothesis space. Otherwise, upon termination the search results in a list of all plausible sequence models of the given data along with their associated probabilities. Example decompositions are shown in Figure 6.

6 Performance on Real Data

The algorithm was first tested on real data, a section of which was shown in Figure 1. The whole waveform consisted of 40 sec of data, filtered from 300 to 7000 Hz and sampled at 20 kHz. Three iterations of the algorithm were performed with overlap decomposition after the second (with $L = 1$) and third (with $L = 2$) iterations. Spike models that occurred fewer than 10 times were discarded for efficiency, and the remaining events were reclassified. The inferred spike models are shown in Figure 7. The residuals indicate that these spike models account for almost all events

in the 40 sec waveform. Out of about 1500 total events, only 6 were not fit to within $5\sigma_\eta$. By eye, these events looked very noisy and had no obvious composition in terms of the spike models. One possibility is that they resulted from animal movement. Such events were not present in the synthesized data set described in Section 7 where all the events were fit with the inferred spike models.

By eye, all the models look distinct except perhaps for M_2 and M_3. One way to see the difference between these two models is to fit the data from model 3 with model 2 as shown in Figure 8. With a single electrode it is difficult to determine whether or not these two shapes result from different neurons, but they are clearly two types of events. One possibility is that these are different states of the same neuron; another is that the shape in model 3 results from a tight coupling between two neurons. Recording with multiple electrodes from a local region of tissue would help resolve issues like this.

In spite of all the math, the algorithm is fast. Inferring the spike set with overlap decomposition takes a few minutes on a Sun Microsystems Sparc IPX. Classification of the 40 sec test waveform with overlap decomposition (using $L = 1$) takes about 10 sec.

7 Performance on Synthesized Data

The accuracy of the algorithm was tested by generating an artificial data set composed of the six inferred shapes shown in Figure 7. The event times were Poisson distributed with frequency equal the inferred firing rate of the real data set. Gaussian noise was then added with standard deviation equal to σ_η. The algorithm was run under the same conditions as above.

The algorithm chose 14 initial spike models, which were subsequently collapsed to 6 using the methods discussed in the previous section. Note that in this case, the number of inferred models matches the number of true models, but this need not be the case if some true models are too similar to be resolved, or if there are insufficient data to identify two distinct classes. The six-model spike set was preferred over the most probable five-model spike set by $\exp(34) : 1$ and over the most probable seven-model spike set by $\exp(19) : 1$. A summary of the accuracy of the spike shapes is shown in Table 1.

The results of inferring and classifying the synthesized data set are shown for the nonoverlapping spikes in Table 2 and for the overlapping spikes in Table 3. An event was considered an overlap if the extent[5] overlapped the extent of another event. Perfect performance would have all zeros in the off-diagonal entries and no undetected events. An event can be missed if it is not detected in an overlap sequence or if all its

[5]The extent of a event is defined as the minimum and maximum values in time at which the best-fitting spike function differs from DC by more than $0.5\sigma_\eta$.

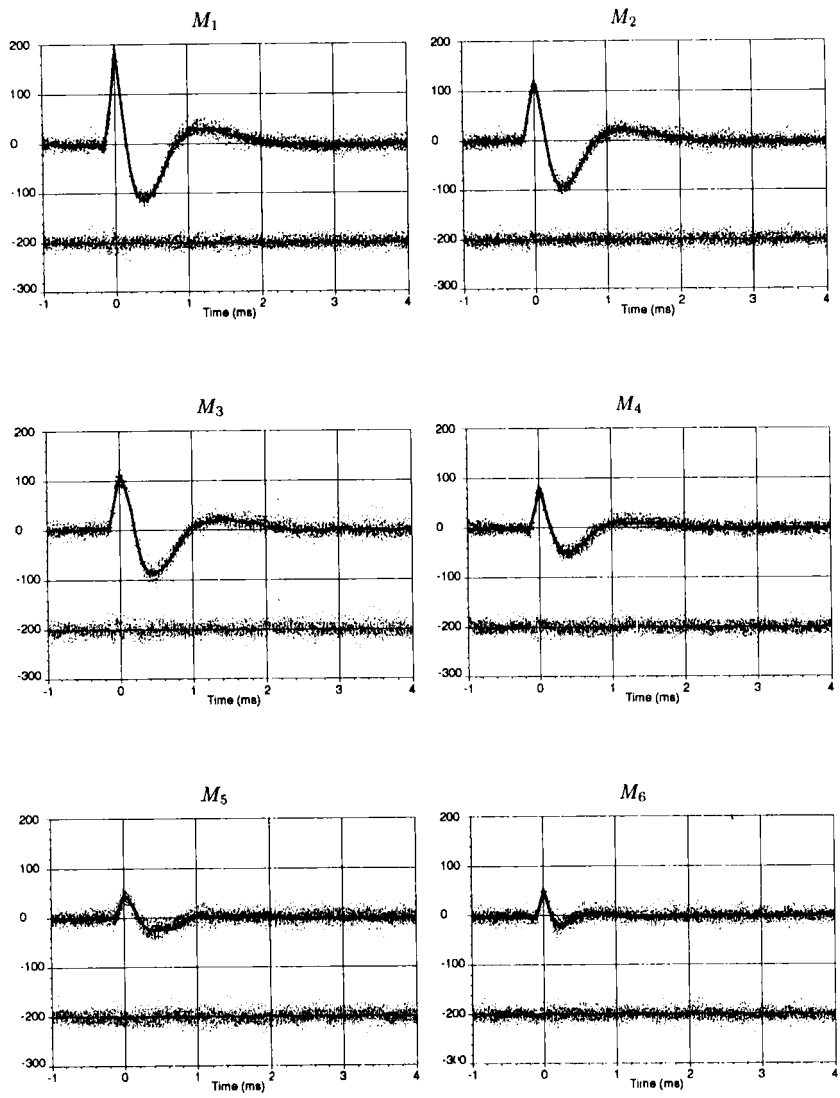

Figure 7: The results of applying the algorithm to a real data set. The solid lines are the inferred spike models. The data overlying each model are a sample of at most 40 events. The residual errors are plotted below each model.

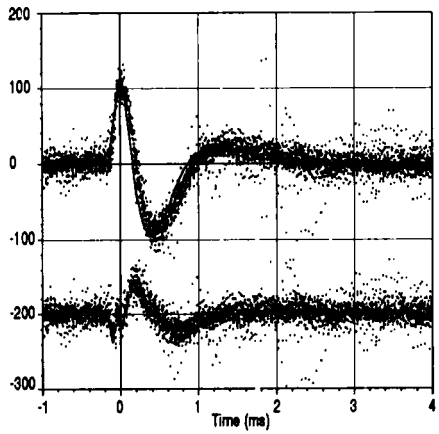

Figure 8: One way to see the difference between the spike models M_2 and M_3 is to fit the data from M_3 (points) with M_2 (solid line). The residual errors are plotted below. All the data from both spike models are plotted. If the noise level is constant throughout the duration of the AP, the large deviation in the residuals indicates that there are two distinct classes.

Table 1: Results of the Spike Model Inference Algorithm on the Synthesized Data Set.[a]

	Model					
	1	2	3	4	5	6
$\Delta_{\max}/\sigma_\eta$	0.44	0.36	1.07	0.78	0.84	0.40
$\max_t s_k(t)/\sigma_\eta$	17.9	11.1	10.6	7.4	4.4	5.0
Number of occurrences	39	63	45	238	155	1055

[a]Both the form and number of spike models were determined by the algorithm. The inferred number of spike models matched the true number (six models). The second row is the maximum absolute difference between the true spike model and the inferred model normalized by σ_η. The third row is the normalized peak of the inferred spike models, which is an indication of how far each type of AP is above the noise level. The last row shows the number of times each model occurred in the synthesized data.

sample values fall below the threshold for event detection ($4\sigma_\eta$). The tables indicate that for the largest four spikes, the performance is nearly perfect, even including the overlapping cases.

Performance is worst in the two smallest spike models where there are a large number of missed events. For these models, there are typ-

Table 2: Classification Results for the Nonoverlapping Events of the Synthesized Data Set.[a]

True	Inferred models						Missed	Total
models	1	2	3	4	5	6	events	events
1	17	0	0	0	0	0	0	17
2	0	25	1	0	0	0	0	26
3	0	0	15	0	0	0	0	15
4	0	0	0	116	0	0	1	117
5	0	0	0	0	56	0	17	73
6	0	0	0	0	0	393	254	647

[a]Each matrix component indicates the number of times true model i was classified as inferred model j. Events were missed if the true spikes were not detected in an overlap sequence or if all sample values for the spike fell below the event detection threshold ($4\sigma_\eta$). There was one false positive for M_5 and seven for M_6. See text for additional comments.

Table 3: Classification Results for the Overlapping Events of the Synthesized Data Set.[a]

True	Inferred models						Missed	Total
models	1	2	3	4	5	6	events	events
1	22	0	0	0	0	0	0	22
2	0	36	1	0	0	0	0	37
3	0	0	20	0	0	0	0	20
4	0	1	0	116	0	1	3	121
5	0	0	0	1	61	1	19	82
6	0	0	0	3	2	243	160	408

[a]Each matrix component indicates the number of times true model i was classified as inferred model j. Events were missed if the true spikes were not detected in an overlap sequence or if all sample values for the spike fell below the event detection threshold ($4\sigma_\eta$). There was one false positive for M_5 and seven for M_6. See text for additional comments.

ically only two or three samples that would be expected to exceed the noise level. As the threshold for event detection is lowered, there is a trade-off between the number of real spikes missed and the number of false positives resulting from common instance of when the noise contains a spike-like shape. The number of below threshold missed events can be minimized (with additional computational expense) by computing the probabilities at every sample point instead of only those that cross threshold. It is worth noting that this situation often does not pose

a problem in practice, since observed spikes just above the noise level frequently correspond to many different neurons.

8 Comparisons with Other Approaches

It is instructive to contrast the spike sorting algorithm presented here with other methods by comparing their performances on the synthesized data set used in the previous section. The most common method of classifying APs is through use of a hardware level detector, which detects an AP if the voltage exceeds a user-determined level. For the synthesized data set, a level detector is sufficient only to classify the largest AP (M_1) with accuracy. Another common hardware approach is a window discriminator with which APs are detected only if the peak value is within a voltage window. A window discriminator can classify M_1 accurately and classify M_4 with some error since the distribution of the M_4 peak voltages overlaps somewhat with other models, but it is not sufficient to discriminate between M_2 and M_3 or between M_5 and M_6. These discriminations demand more sophisticated methods.

A common software-based method for spike sorting is a feature clustering algorithm such as the one used in the commercial physiological data collection system *Brainwave*. The synthesized waveform was classified independently by an experienced Brainwave user (Matt Wilson). The features used to perform the classification were maximum spike amplitude, minimum spike amplitude, and time from the spike maximum to the spike minimum. Brainwave generates a list of occurrence times for each cluster but not explicit spike functions, so it was not possible to see how close the "inferred spike functions" were to the true spike functions. The occurrence times were compared to the known AP positions. Two separate classifications were performed (one using four clusters and another using six clusters), and the results of the most accurate classification (six clusters) are reported here.

Tables 4 and 5 show the classification results for the synthesized data set for the nonoverlapping and overlapping action potentials, respectively. A total of six clusters were found, but not all of these correspond to the true underlying clusters.

The tables show that true models M_1 and M_4 were accurately identified and classified. True models M_2 and M_3, however, were collapsed into a single cluster. This discrimination is difficult to make without accurately estimating the occurrence time of the APs. Brainwave uses the spike peak for the occurrence time, which is accurate to within 1 sample period and introduces a significant amount of noise into the features. In contrast, the Bayesian approach estimates the spike occurrence times with subsampling period accuracy. Note also that with no overlap decomposition, there are significantly more missed events for the larger APs.

Table 4: Brainwave Classification Results for the Nonoverlapping Events of the Synthesized Data Set.[a]

True	Cluster number						Missed	Total
models	1	2	3	4	5	6	events	events
1	17	0	0	0	0	0	0	17
2	0	26	0	0	0	0	0	26
3	0	15	0	0	0	0	0	15
4	0	0	116	0	0	0	1	117
5	0	0	1	24	0	6	42	73
6	0	0	0	22	13	188	424	647

[a]Each matrix component indicates the number of times true model i was classified as belong to Brainwave cluster j. An event was missed if a true AP did not correspond to any of the APs identified by Brainwave. The false positives counts were 2, 3, 4, and 2 for Brainwave clusters 3, 4, 5, and 6, respectively.

True models M_5 and M_6 were described with three clusters, with clusters 4 and 5 roughly corresponding to M_5 and cluster 6 corresponding to M_6. For these models, the features used make it difficult to choose the correct clusters, since the smaller models are not well separated in the three-dimensional feature space. There is less separation, because the occurrence times are not estimated accurately and no overlap decomposition is done. Even if the cluster centers were accurate, we would expect the Brainwave classification to be less accurate than the Bayesian approach. Using spike functions to perform the classification utilizes all significant sample points in the waveform, which for the smallest two models is between four and eight. In contrast, only three features are used by Brainwave.

9 Extensions

There are a number of possible directions for improvements to the general waveform model we have described. At the lowest level there are possibilities for alternative noise models. For example, real extracellular noise tends to be correlated and slightly nongaussian. Incorporating this information would make the probabilities more accurately reflect the real world.

The piece-wise linear model we have described is general enough to fit almost arbitrary shapes, but that generality is also part of its shortcoming. Since in the algorithm we have placed minimal restrictions on the form of the spike model, more data are required to infer the shape. Incorporating knowledge about the spike shapes would result in more accurate conclusions with the same amount of data. Overly weak spike

Table 5: Brainwave Classification Results for the Overlapping Events of the Synthesized Data Set.[a]

True	Cluster number						Missed	Total
models	1	2	3	4	5	6	events	events
1	22	0	0	0	0	0	0	22
2	0	34	1	0	0	0	2	37
3	0	18	1	0	0	0	1	20
4	0	3	106	3	0	2	7	121
5	0	0	1	26	8	10	45	82
6	0	3	9	15	2	108	264	408

[a]Each matrix component indicates the number of times true model i was classified as belong to Brainwave cluster j. An event was missed if a true AP did not correspond to any of the APs identified by Brainwave. The false positives counts were 2, 3, 4, and 2 for Brainwave clusters 3, 4, 5, and 6, respectively.

shape priors will also result in overly strong Occam factors, which will bias the results of model comparisons toward simpler models.

For some types of neurons the shape of an action potential is not constant. Bursting neurons, for example, have spikes that decay dramatically during a burst. Modeling the resulting shape is complicated because the interspike intervals during a burst are not constant over different bursts, and the degree of attenuation depends on the intervals. Another way in which APs can change their shape is due to electrode drift, which results in a slow change of the spike shapes over time. This can be handled readily by the algorithms since re-estimating previously inferred shapes is very fast.

Another limitation stems not from the algorithm but from the method of recording. Since a single electrode gives little information about a neuron's position, decisions about whether two shapes constitute two neurons must be made based on shape and firing frequency alone. The use of multiple electrodes in a local area resolves this issue by recording the same group of neurons from different sites. Thus even if two neurons have identical shapes when recorded from electrode, it is unlikely that those two neurons will generate the same AP shape when observed simultaneously from a different electrode. A trivial extension of the algorithm would be to run it on each electrode and then look for cross-correlations in the event times, but better results could be obtained by incorporating the information about multiple electrodes into a single model.

10 Discussion

Formulating the task as the inference of a probabilistic model made clear what was necessary to obtain accurate spike models. Optimizing the τ_n's

is crucial for both inference and classification, but this step is commonly ignored by algorithms which cluster the sample points or derive spike shapes from principal components. The soft clustering procedure makes it possible to determine the spike shapes with accuracy even when they are highly overlapping. Unless the spike shapes are well-separated, hard clustering procedures such as k-means will lead to inaccurate estimates of the spike shapes.

Probability theory also provided an objective means of determining the number of spike models, which is an essential reason for the success of this algorithm. With the incorrect number of spike models, overlap decomposition becomes especially difficult. If there are too few spike models, the overlap data cannot be fit. If there are too many, decomposition becomes a very expensive computation. The evidence has proved to be a sensitive indicator of when two classes are distinct, as was shown in Figure 8. Previous approaches have relied on ad hoc criteria or the user to make this decision, but such approaches cannot be relied upon to work under varying circumstances since their inherent assumptions are not explicit. An advantage of probability theory is that the assumptions are explicit, and given those assumptions, the answer provided by the evidence is optimal.

One might wonder if the user, having much more information than has been incorporated into the model, can make better decisions than the evidence about what constitutes distinct spike models. Probability theory provides a calculus for stating precisely what can be inferred from the data given the model. When the conclusions reached through probability theory do not fit our expectations, it is due to a failure of the model or a failure of the approximations (if approximations are made). From the performance on the synthesized data, however, the approximations appear to be reasonable. Thus when the conclusions reached through the evidence are at variance with the user's, information is at hand about possible shortcomings of the current model. In this manner, new models can be constructed, and, moreover, they can be compared objectively using the evidence.

Probability theory is also essential for accurate overlap decomposition. It is not sufficient just to fit data with compositions of spike models. That leads to the same overfitting problem encountered in determining the number of spike models and in determining the spike shapes. The Ockham penalty introduced by integrating out the τ's was a key to finding the most probable fits and consequently for achieving accurate classification. Previous approaches have been able to handle only a limited class of overlaps, mainly due to the difficulty in making the fit efficient. The algorithm we have described can fit an overlap sequence of virtually arbitrary complexity in milliseconds.

In practice, the algorithm we have described allows us to extract much more information from an experiment than with previous methods. Moreover, this information is qualitatively different from a simple

list of spike times. Having reliable estimates of the action potential shapes makes it possible to study the properties of these classes, since distinct neuronal types can have distinct neuronal spikes (Connors and Gutnick 1990). With stereotrodes this advantage would be amplified, since it is then possible to estimate somatic size, which is another distinguishing characteristic of cell type. Finally, accurate overlap decomposition makes it possible to investigate interactions among local neurons, which were previously very difficult to observe.

Acknowledgments

I thank David MacKay for helpful discussions and encouragement during the early stages of this work, Jamie Mazer for many conversations and extensive help with the development of the software, and Matt Wilson for classifying the synthesized data set with Brainwave. Thanks also to Allison Doupe and Ken Miller for helpful feedback on the manuscript. This work was supported by Caltech fellowships and an NIH Research Training Grant.

References

Atiya, A. F. 1992. Recognition of multiunit neural signals. *IEEE Transact. Biomed. Eng.* **39**(7), 723–729.

Bently, J. L. 1975. Multidimensional binary search trees used for associative searching. *Commun. ACM* **18**(9), 509–517.

Connors, B. W., and Gutnick, M. J. 1990. Intrinsic firing patterns of diverse neocortical neurons. *TINS* **13**(3), 99–104.

Duda, R. O., and Hart, P. E. 1973. *Pattern Classification and Scene Analysis.* Wiley-Interscience, New York.

Friedman, J. H., Bently, J. L., and Finkel, R. A. 1977. An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Software* **3**(3), 209–226.

Gull, S. F. 1988. Bayesian inductive inference and maximum entropy. In *Maximum Entropy and Bayesian Methods in Science and Engineering, Vol. 1: Foundations,* G. J. Erickson and C. R. Smith, eds. Kluwer.

Jaynes, E. T. 1979. Review of *Inference, Method, and Decision* (R. D. Rosenkrantz). *J. Am. Stat. Assoc.* **74**, 140.

Jeffreys, H. 1939. *Theory of Probability.* (3rd rev. ed 1961). Oxford University Press, New York.

MacKay, D. J. C. 1992. Bayesian interpolation. *Neural Comp.* **4**(3), 415–445.

Schmidt, E. M. 1984. Computer separation of multi-unit neuroelectric data: A review. *J. Neurosci. Methods* **12**, 95–111.

Ramasubramanian, V., and Paliwal, K. K. 1992. Fast k-dimensional tree algorithms for nearest-neighbor search with application to vector quantization encoding. *IEEE Trans. Signal Proc.* **40**(3), 518–531.

Wahba, G. 1990. *Spline Models for Observational Data*, SIAM.

Received June 29, 1993; accepted January 11, 1994.