J Comput Neurosci (2008) 24:69–79
DOI 10.1007/s10827-007-0042-x

Integral equation methods for computing likelihoods and their derivatives in the stochastic integrate-and-fire model

Liam Paninski · Adrian Haith · Gabor Szirtes

Received: 11 August 2006 / Revised: 13 March 2007 / Accepted: 19 April 2007 / Published online: 10 May 2007
© Springer Science + Business Media, LLC 2007

Abstract We recently introduced likelihood-based methods for fitting stochastic integrate-and-fire models to spike train data. The key component of this method involves the likelihood that the model will emit a spike at a given time t. Computing this likelihood is equivalent to computing a Markov first passage time density (the probability that the model voltage crosses threshold for the first time at time t). Here we detail an improved method for computing this likelihood, based on solving a certain integral equation. This integral equation method has several advantages over the techniques discussed in our previous work: in particular, the new method has fewer free parameters and is easily differentiable (for gradient computations). The new method is also easily adaptable for the case in which the model conductance, not just the input current, is time-varying. Finally, we describe how to incorporate large deviations approximations to very small likelihoods.

Keywords Volterra integral equation · Markov process · Large deviations approximation

Action Editor: Barry J. Richmond

L. Paninski (✉)
Department of Statistics, Columbia University
New York, NY, USA
e-mail: liam@stat.columbia.edu
URL: http://www.stat.columbia.edu/∼liam

A. Haith
Institute for Perception, Action and Behaviour,
University of Edinburgh, Edinburgh, UK

L. Paninski · G. Szirtes
Center for Theoretical Neuroscience, Columbia University
New York, NY, USA

1 Introduction

A classic and recurring problem in theoretical neuroscience is to estimate the probability that an integrate-and-fire-type neuronal model, driven by white Gaussian noise, that has fired at time $t = 0$ will not fire again until time $t = T$. This problem appears in a number of contexts, including firing rate computations (Plesser and Gerstner 2000; Plesser and Tanaka 1997), statistical model fitting (Iyengar and Liao 1997; Paninski et al. 2004b), and decoding (Pillow et al. 2005). In particular, Paninski et al. (2004b) recently introduced likelihood-based methods for fitting stochastic integrate-and-fire models to spike train data; these techniques rely on the numerical computation of these interspike interval (ISI) densities. Computing this likelihood is equivalent to computing a Markov first passage time density, the probability that the model voltage (a Markov process) crosses threshold for the first time at time $t = T$, given that the voltage was reset to some fixed subthreshold value at time $t = 0$. The main motivation for this paper is to develop efficient, robust techniques for computing this likelihood in the model-fitting framework, to facilitate the application of these models to real in vivo and in vitro data (Paninski et al. 2004a,b; Pillow et al. 2005).

Here we detail an improved numerical method for computing this likelihood, based on techniques introduced by Plesser and Tanaka (1997) and DiNardo et al. (2001). We begin by noting that the ISI density uniquely solves a certain linear Volterra integral equation, then provide details on approximating this integral equation by a lower-triangular matrix equation, which may be solved efficiently on a computer. In addition, the gradient of this solution with respect to the model parameters may be efficiently computed via

straightforward matrix perturbation techniques. This semi-analytic computation of the gradient speeds numerical optimization of the model parameters in a maximum-likelihood setting and therefore enables consideration of models with many more parameters than has previously been feasible.

This integral equation method has several advantages over the techniques discussed in our previous work (Paninski et al. 2004b) (where we discussed two methods: one based on Gaussian integrals over "boxes" in a high-dimensional space, and the other on the numerical solution to a Fokker–Planck partial differential equation): the new method has fewer free parameters and (as mentioned above) is much more easily differentiable. The new method is also easily adaptable for the case in which the model conductance, not just the input current, is allowed to vary as a function of time.

Finally, since the likelihood of a given spike train may be decomposed into a product over the likelihoods of each individual ISI, it is convenient to work with log-likelihoods. However, numerical errors in computing these small likelihoods can have a large deleterious effect on the overall likelihood computation in the log domain. (For example, numerical instabilities may occasionally convert very small probabilities into negative numbers, which is a disaster when taking logarithms.) Thus the computation of these very low-probability events must be handled carefully, both in the initialization stage of any maximization routine but also even near convergence to the maximum likelihood solution (since real data inevitably contains some outliers, when the neuron may have spiked at a highly unlikely time). In order to deal with this issue, we introduce a technique, based on the probabilistic theory of large deviations, which permits us to approximate these very small likelihoods on a logarithmic scale (Paninski 2006). Once again, this large deviation approximation (along with its gradient) may be computed efficiently using simple linear-algebraic techniques.

2 Previous approaches

We begin by reviewing two very different methods that have been proposed for the computation of the likelihood.

2.1 Gaussian integral method

The first method is based on a path-integral representation of the likelihood: the probability that a noisy LIF neuron spikes at some time t but at not at any previous times $0 < s < t$ may be computed formally as the fraction of all possible voltage paths $V(t)$ which lie in some constraint set C:

$$P(\text{spike at time t}) = P(V(s) < V_{th},\ s < t;\ V(t) \geq V_{th}) = \int_{\{V(t)\}\in C} dP(\{V(t)\});$$

here the constraint set C is defined as the set of all paths which cross the threshold for the first time at t

$$C = \{V : V(s) < V_{th},\ 0 < s < t;\ V(t) \geq V_{th}\},$$

(i.e., the set of all voltage paths consistent with the observed spiking data) and the probability measure $dP(\{V(t)\})$ is the Gaussian measure induced by the linear stochastic differential equation

$$dV(t) = (-gV(t) + I(t))dt + \sigma dN_t, \tag{1}$$

where $V(t)$ is driven by some time-varying input $I(t)$ and standard Gaussian white noise dN_t, scaled by σ. (We emphasize that dN_t is current noise here, not conductance noise; the latter case is more difficult, and will not be treated here.) Note that the "sides" of the box C are linear in $\{V(t)\}$, and that the mean and covariance of the Gaussian are easily-computed functions of the parameters $\theta = \{g, \sigma, I(t)\}$ (see Eqs. (3) and (4) below).

This representation suggests a simple direct approach to computing the likelihood: we discretize the time interval $[0, t]$ into d points $(s_1, s_2, \ldots s_{d-1}, t)$ and compute the finite-dimensional Gaussian integral

$$\int_C dP(\{V(s_1), V(s_2), \ldots, V(t)\}) = \int_{-\infty}^{V_{th}} dV(s_1) \int_{-\infty}^{V_{th}} dV(s_2) \ldots \int_{V_{th}}^{\infty} dV(t)\, P(\{V(s_1), V(s_2), \ldots, V(t)\}).$$

Efficient algorithms are available to compute integrals of this type (Genz 1992), for discretization depths d up to about 10. However, finer discretizations ($d \gg 10$) rapidly become numerically intractable; thus this direct approach can provide only a rough approximation of the true likelihood (and unfortunately no estimates of the approximation error are easily available).

The appendix of Paninski et al. (2004b) describes a method for computing the gradient of this integral with respect to the parameters $I(t)$ which requires the computation of d separate $(d-1)$-dimensional Gaussian integrals (this is more efficient than simple finite differences when the dimensionality of $\{I(t)\}$ is less than d); gradients with respect to σ and g must be computed by finite differences.

In summary, this direct Gaussian integral method provides a fairly crude approximation to the true likelihood and requires significant numerical effort to compute gradients.

2.2 Forward equation (Fokker–Planck) method

A more accurate approach may be developed using Fokker–Planck techniques (Haskell et al. 2001; Karlin and Taylor 1981; Knight et al. 2000; Paninski et al. 2004b). If we define

$$P(V,t) \equiv P(V(t) \cap V(s) < V_{th}\ \forall\, s < t)$$

then it is well-known that the first-passage time density, the likelihood of observing the first spike at time t, is given by

$$p(t) = -\frac{\partial}{\partial t}\int P(V,t)dV,$$

and $P(V,t)$ satisfies the partial differential (Fokker-Planck, or "forward") equation

$$\frac{\partial P(V,t)}{\partial t} = \frac{\sigma^2}{2}\frac{\partial^2 P(V,t)}{\partial V^2} + \frac{\partial[(g(t)V - I(t))P(V,t)]}{\partial V},$$

under boundary conditions

$$P(V_{th},t) = 0$$

and

$$P(V,0) = \delta(V - V_{reset}),$$

where 0, without loss of generality, denotes the time of the last observed spike, where the voltage V is reset deterministically to V_{reset}.

This linear advection–diffusion PDE for $P(V,t)$ may be solved efficiently, to any desired level of precision, via the usual numerical methods (e.g. Crank–Nicholson) (Press et al. 1992): $P(V,t)$ is discretized in time and voltage (the time discretization is between 0 and t; the voltage discretization is between the upper bound V_{th} and some sufficiently hyperpolarized lower bound V_{lb}), and the discretized PDE is solved by iteratively solving the corresponding set of tridiagonal linear equations. The likelihood $-\frac{\partial}{\partial t}\int P(V,t)dV$ may then be computed via standard methods (e.g. Euler or trapezoidal integration in V and finite differences in t). Arbitrarily accurate solutions are obtained by letting the time and voltage discretizations become arbitrarily fine, and simultaneously letting the voltage lower bound $V_{lb} \to -\infty$. Computing the likelihood takes $O(d_V d_t)$ steps, where d_V and d_t are the number of bins in the voltage and time discretization grids, respectively.

Methods for semi-analytically computing the gradient of the likelihood via this PDE technique are somewhat more involved and will be described elsewhere.

3 Integral equation approach

Now we turn to the main topic of this paper. A third method for computing the first-passage time density $p(t)$ may be derived by noting another well-known fact (Burkitt and Clark 1999; Plesser and Tanaka 1997; Ricciardi 1977; Siegert 1951): $p(t)$ solves the Volterra integral equation

$$G_\theta(y,t|V_{reset},0) = \int_0^t G_\theta(y,t|V_{th},s)p(s)ds, \tag{2}$$

under the initial condition $p(0) = 0$, for $y \geq V_{th}$. Here we have abbreviated

$$G_\theta(y,t|x,s) \equiv P(V(t) = y|V(s) = x,\theta),$$

the conditional probability that the voltage V, evolving under Eq. (1) given the model parameters θ, will be at level y at time t given that V has been observed to be equal to x at time s. As noted above, these conditional probability densities are Gaussian, and the relevant means and variances may be computed easily (see, e.g. Karlin and Taylor 1981, for details):

$$\sigma^2(t|s) \equiv \mathrm{Var}(V(t)|V(s)) = \sigma^2\int_s^t e^{-2\int_u^t g(v)dv}du \tag{3}$$

$$\begin{aligned}\mu(t|x,s) &\equiv E(V(t)|V(s) = x)\\ &= xe^{-\int_s^t g(v)dv} + \int_s^t I(u)e^{-\int_u^t g(v)dv}du.\end{aligned} \tag{4}$$

Here we are allowing $g(t)$ to vary with time, for increased generality and biophysical accuracy (Jolivet et al. 2004; Stevens and Zador 1998). Equation (2) is true for any $y \geq V_{th}$, but we will take $y = V_{th}$ when solving numerically since that gives rise to the best-conditioned linear system and therefore the most accurate solution.

It will be useful in the following to use the parameterization $V_{th} = 1$, $V_{reset} = 0$. Note that this entails

no loss of generality (since the voltage paths $V(t)$ are unobserved and therefore have a free offset and scale term).

3.1 Second-kind integral equation

The Volterra integral equation given above is of the "first kind." A related integral equation of the "second kind" is

$$p(t) = -2\psi[V_{th}, t|V_{reset}, 0] + 2\int_0^t \psi[V_{th}, t|V_{th}, s]p(s)ds \tag{5}$$

where ψ is defined by

$$\psi[y, t|x, s] = \frac{\partial}{\partial t}\int_{-\infty}^{y} G_\theta(y', t|x, s)dy'.$$

This can be derived straightforwardly from Eq. (2) by integrating both sides with respect to y from V_{th} to ∞, and then taking the derivative with respect to t (Buoncore et al. 1987).

Note that $\int_{-\infty}^{y} G_\theta(y', t|x, s)dy'$ is the cumulative distribution of the voltage at time t (conditional on the voltage being equal to x at time s). ψ represents the rate of change of this quantity and is equivalent to a probability current (or flux) across y at time t. The integral equation as a whole, therefore, states that the first-passage time density $p(t)$ is proportional to the difference between the total probability current through V_{th} at time t (the first term) and the contribution to this current from processes which have already hit V_{th} at some time prior to t (the second term).

Writing $\int_{-\infty}^{y} G_\theta$ as $\frac{1}{2}\left[1 + Erf\left\{\frac{y-\mu(t|x,s)}{\sqrt{2\sigma^2(t|s)}}\right\}\right]$ (since G_θ is Gaussian), and using the derivative of the error function

$$\frac{d}{dz}Erf(z) = \frac{2}{\sqrt{\pi}}e^{-z^2},$$

and the time derivatives of the mean and variance (derived from Eqs. (3) and (4) above)

$$\frac{\partial}{\partial t}\mu(t|x, s) = -g(t)\mu(t|x, s) + I(t)$$
$$\frac{\partial}{\partial t}\sigma^2(t|s) = \sigma^2 - 2g(t)\sigma^2(t|s),$$

we can show that

$$\psi(V_{th}, t|x, s) = G_\theta(V_{th}, t|x, s)\left[g(t)V_{th} - I(t) - \frac{\sigma^2}{2\sigma^2(t|s)} \times (V_{th} - \mu(t|x, s))\right].$$

3.2 Adjustment to remove singularity

In both the first and second-kind integral equations, there is a singularity in the kernel due to $G_\theta(V_{th}, t|V_{th}, s)$ diverging as $s \to t$. Although not too drastic a problem, this will cause difficulties when solving numerically (see numerical solution of the first kind equation; Section 4). However, as described in Buoncore et al. (1987), it is possible to remove this singularity. Note that any function of the form

$$\varphi(V_{th}, t|x, s) = \psi(V_t h, t|x, s) + \lambda(t)G_\theta(V_{th}, t|x, s)$$

will also satisfy Eq. (5) in place of ψ since the resulting extra terms will sum to zero, according to the first-kind equation (2). We can set λ so that the singularities in ψ and G_θ cancel one another out exactly, resulting in a non-singular φ. The appropriate value is given by

$$\lambda(t) = -\lim_{s\to t}\left[g(t)V_{th} - I(t) - \frac{\sigma^2}{2\sigma^2(t|s)}(V_{th} - \mu(t|x, s))\right]$$
$$= -\frac{1}{2}(g(t)V_{th} - I(t)).$$

We will therefore set

$$\varphi(V_{th}, t|x, s) = \frac{1}{2}\left[g(t)V_{th} - I(t) - \frac{\sigma^2}{\sigma^2(t|s)}(V_{th} - \mu(t|x, s))\right] \times G_\theta(V_{th}, t|V_{th}, s)$$

as the kernel for numerical solution of the second-kind integral equation.

An additional benefit of using φ rather than ψ is that in the nonleaky, constant-current case (i.e., $g(t) \equiv 0$ and $I(t) \equiv I$ for some constant I), $\varphi(V_{th}, t|V_{th}, s) = 0$ for all s and t (Buoncore et al. 1987). The integral term vanishes and we are left with the exact analytical solution for $p(t)$ (Karatzas and Shreve 1997; Tuckwell 1989):

$$p(t) = \frac{V_{th} - V_r}{\sqrt{2\pi\sigma^2 t^3}}e^{-((V_{th}-V_r)-It)^2/2\sigma^2 t}.$$

In this case then, the second-kind equation gives the exact density, no matter what the discretization level used.

4 Numerical solution of the integral equation

Standard methods exist for numerically solving Volterra equations (Press et al. 1992); the basic idea

is to use the fact that the integral on the right-hand-side of both the first- and second-kind equation only includes information about $p(t)$ up to time t; thus, just as in the numerical solution of an ordinary differential equation, we may begin with the initial condition $p(0) = 0$, then recursively compute $p(t + dt)$ given the value of $p(s)$ for $0 \le s \le t$, by computing the integral on the right-hand-side.

We implement a simple trapezoidal rule for computing this integral here. A good summary of one solution (for the first-kind equation) is given in Plesser and Tanaka (1997), borrowing a method from Press et al. (1992); an alternate approach is given in DiNardo et al. (2001) and Haith (2004) (see also references therein). We begin by following Plesser and Tanaka (1997) here, filling in some additional computational details along the way, but end up taking a slightly different approach which will permit us to differentiate the solution much more easily (Section 6).

To compute the necessary functions $G_\theta(y, t|x, s)$, we represent $I(t)$ and $g(t)$ as piecewise constant, with $d - 1$ equally-spaced discontinuities on $[0, T]$. Set the discretization width $\Delta = T/d$.

The means and variances of these Gaussian functions may be computed exactly and recursively, allowing computation in $O(d^2)$ time (instead of the $O(d^3)$ which would be required of a naive implementation). Computing the exponential function turns out to be the most time-consuming step here; using the recursive approach, only d exponentials need be computed to obtain the means $\mu(t|x, s)$ and variances $\sigma^2(t|s)$. For example, to compute the first term in the expression for $\mu(t|x, s)$ above,

$$xe^{-\int_s^t g(v)dv} = x \prod_{i=s/\Delta}^{t/\Delta-1} e^{-g(i\Delta)\Delta}$$

(remembering that $g(t)$ is piecewise constant on intervals of size Δ), which may be computed via an obvious backwards recursion. Similarly,

$$\begin{aligned}\sigma^2 \int_s^t e^{-2\int_u^t g(v)dv} du &= \sigma^2 \sum_{i=1}^{(t-s)/\Delta} \int_{s+(i-1)\Delta}^{s+i\Delta} e^{-2\int_u^t g(v)dv} du \\ &= \sigma^2 \sum_{i=1}^{(t-s)/\Delta} e^{-2\int_{s+i\Delta}^t g(v)dv} \\ &\quad \times \int_{s+(i-1)\Delta}^{s+i\Delta} e^{-2\int_u^{s+i\Delta} g(v)dv} du,\end{aligned}$$

so defining the vectors

$$u_2(i) = e^{-2g(i\Delta)\Delta}$$

and

$$v_2(i) = \begin{cases} \frac{1-u_2(i)}{2g(i\Delta)}, & g(i\Delta) > 0 \\ \Delta, & g(i\Delta) = 0 \end{cases}$$

for $i = 0, \ldots, d - 1$, we have

$$\sigma^2 \int_s^t e^{-2\int_u^t g(v)dv} du = \sigma^2 \sum_{i=s/\Delta}^{t/\Delta-1} v_2(i) \prod_{j=i+1}^{t/\Delta-1} u_2(j),$$

and defining

$$u_1(i) = u_2(i)^{1/2} = e^{-g(i\Delta)\Delta},$$

$$v_1(i) = \begin{cases} \frac{1-u_1(i)}{g(i\Delta)}, & g(i\Delta) > 0 \\ \Delta, & g(i\Delta) = 0 \end{cases},$$

we have

$$\begin{aligned}xe^{-\int_s^t g(v)dv} + \int_s^t I(u)e^{-\int_u^t g(v)dv} du &= x \prod_{i=s/\Delta}^{t/\Delta-1} u_1(i) \\ &\quad + \sum_{i=s/\Delta}^{t/\Delta-1} v_1(i) I(i\Delta) \prod_{j=i+1}^{t/\Delta-1} u_1(j).\end{aligned}$$

Once this matrix of means and variances has been obtained, we may compute the matrix of corresponding Gaussians $G(y, t|x, s)$ in the obvious way (unfortunately we can not avoid evaluating $O(d^2)$ exponentials to compute the Gaussian matrix), and once this matrix of Gaussians is in hand we are ready to compute the integrand on the right-hand-side.

Previous authors (DiNardo et al. 2001; Haith 2004; Plesser and Tanaka 1997) have described the solution of the integral equation in terms of iterated trapezoidal integrals, as indicated above. An alternative, slightly more symbolic viewpoint is to treat the integral equation as a lower-triangular linear system $Ap = b$, for some lower-triangular matrix A. In the case of the first-kind equation, this is straightforward: A consists of the matrix of $G(V_{th}, t|V_{th}, s)$ values, while b is a vector containing the discretized function $G(V_{th}, t|V_{reset}, 0)$. The second-kind case is slightly more subtle, but it is not hard to see that here the vector b corresponds to $2\varphi[V_{th}, t|V_{reset}, 0]$, and A is given by $2\varphi[V_{th}, t|V_{th}, s] - \delta(s - t)$.

Either approach to solving the integral approach—the iterated integration or the matrix approach—clearly leads to an $O(d^2)$ solution. Convergence issues as $\Delta \to 0$ may be addressed with standard approaches (DiNardo et al. 2001). In Matlab, the resulting matrix

equation $Ap = b$ is efficiently solved using the notation $p = A\backslash b$, which takes advantage of the lower triangular nature of the system (in contrast to $p = \text{inv}(A) * b$, which is clearly computed in $O(d^3)$ time). The one disadvantage of the matrix formulation versus the repeated-integral formulation is that the former requires $O(d^2)$ memory space, while the latter requires $O(d)$; however, this difference in memory requirements only has an impact on the computation speed for very large d (that is, very fine discretizations of the time interval $(0, T)$). Note that d is the only free parameter here (c.f. the Fokker-Planck approach, which requires the user to specify the time discretization d_t, the voltage discretization d_V, and the lower bound of the voltage discretization V_{lb}).

Note that a direct approach to substituting the integrand into A does not quite work in the first-kind case, because $G(x, t|x, s)$ diverges as $s \to t$. To handle this square-root singularity, we write

$$G_\theta(V_{th}, t|V_{th}, s) = f_1(s)(t-s)^{-1/2},$$

with

$$f_1(s) = \frac{(t-s)^{1/2}}{\sigma(t|s)} \exp\left[(V_{th} - \mu(t|V_{th}, s))^2/2\sigma^2(t|s)\right]$$

a uniformly smoothly differentiable function, with $f_1(s) \to 1, s \to t$.[1] Then $f_1(s)$ is expanded to first order around t:

$$\begin{aligned}\int_{t-\Delta}^{t} G_\theta(V_{th},t|V_{th},s)ds &= \int_{t-\Delta}^{t} f_1(s)(t-s)^{-1/2}ds \\ &\approx \int_{t-\Delta}^{t} [f_1(t)+f_1'(t)(s-t)](t-s)^{-1/2}ds \\ &= 2f_1(t)\Delta^{1/2} - \frac{2}{3}f_1'(t)\Delta^{3/2}\end{aligned}$$

If we approximate

$$f_1'(t) \approx \frac{f_1(t) - f_1(t-\Delta)}{\Delta},$$

we are left with

$$\int_{t-\Delta}^{t} G_\theta(V_{th}, t|V_{th}, s)ds \approx \Delta^{1/2}\left[\frac{4}{3}f_1(t) + \frac{2}{3}f_1(t-\Delta)\right].$$

In this case, we can take $f_1(t) = 1$, further simplifying the computation.

Finally, we incorporate the boundary condition $p(0) = 0$ to eliminate the first column of A and the first element of b, and use the fact that we have assumed that $V_{th} = 1$ and $V_{reset} = 0$, to obtain

$$A = \begin{bmatrix} 4\sqrt{\Delta}/3 & & & & \\ 7\Delta G(1,2\Delta|1,\Delta)/6 & 4\sqrt{\Delta}/3 & & & \\ \Delta G(1,3\Delta|1,\Delta) & 7\Delta G(1,3\Delta|1,2\Delta)/6 & 4\sqrt{\Delta}/3 & & \\ \vdots & & \ddots & \ddots & \\ & \dots & \Delta G(1,T|1,T-2\Delta) & 7\Delta G(1,T|1,T-\Delta)/6 & 4\sqrt{\Delta}/3 \end{bmatrix}$$

and

$$b = \left[\, G(1,\Delta|0,0)\ G(1,2\Delta|0,0) \cdots G(1,T|0,0) \,\right]^t.$$

The details of filling in A and b in the second-kind case follow similarly (although the second-kind case is much more straightforward, since we were able to adjust $\varphi(V_{th}, t|V_{th}, s)$ to eliminate the singularity).

5 Large-deviations approximation of very small likelihoods

As discussed in Haith (2004), the issue of very low-probability events is a nuisance with any of the three methods we have discussed so far. Small errors in the computation of these very small likelihoods cause problems both in the initialization stage of the likelihood optimization routine (where the initial parameters might match the true parameters poorly) and also near convergence (since real data inevitably contains some outliers, which correspond to low-likelihood events).

To handle these low-likelihood events, we make use of the well-known theory of large deviations (Dembo and Zeitouni 1993; Freidlin and Wentzell 1984; Paninski 2006), which provides approximations to $\log p(t)$ that are asymptotically exact in the limit that $D(\mu, C) \to \infty$, where $D(\mu, C)$ denotes a kind of Mahalanobis distance between the mean μ of the Gaussian

[1] Note that the term $\frac{1}{\sigma\sqrt{2\pi}}$ appears as a constant factor on both sides of the first-kind integral equation, and thus we have divided it out here to simplify the computation.

measure on $\{V(t)\}$ and the constraint set C of valid voltage paths.[2]

Specifically, the large-deviation approximation states that

$$\log p(t) = \left(-\frac{1}{2}\inf_{V\in C} D(\mu, V)\right)\left(1 + o(1)\right)$$

as $\inf_{V\in C} D(\mu, V) \to \infty$, with

$$D(\mu, V) \equiv \frac{1}{\sigma^2}\int_0^t \left[\dot{V}(s) - (I(s) - g(s)V(s))\right]^2 ds,$$

where $\dot{V}(s)$ denotes the time-derivative of the voltage path V at time s.

Thus to compute the approximation we must solve the optimization problem $\inf_{V\in C} D(\mu, V)$. As emphasized in Paninski (2006), this is a quadratic programming problem in the vector $\{V(\Delta), V(2\Delta), \ldots, V(T-\Delta)\}$, with a unique global optimum which can be computed via standard and efficient ascent algorithms (Boyd and Vandenberghe 2004) (the uniqueness of the optimizer here is due to the strict convexity of the function $D(.,.)$ and the convexity of the set C). In this case we may start with the analytic guess

$$V_{opt'}(t) \equiv \mu(t|V_{reset}, 0) + e^{-\int_t^T g(u)du}\frac{\sigma^2(t|0)}{\sigma^2(T|0)}\left(V_{th} - \mu(T|V_{reset}, 0)\right);$$

this was derived in Paninski (2006) as the solution to the optimization problem $\inf_{V\in C'} D(\mu, V)$, where the set C' is defined as

$$C' = \{V : V(0) = V_{reset}, V(t) = V_{th}\}$$

(note that $C \subset C'$). It turns out we have already computed all the pieces of the above formula in the previous section, which makes $V_{opt'}$ a convenient initializer for the optimization.

Now if $V_{opt'}(t) \le V_{th}$ for $t < T$ (that is, if $V_{opt'} \in C$), then the optimization problem is complete; otherwise, we need to ascend (via quadratic programming, e.g., quadprog.m in Matlab) into the feasible set C.

Once we have obtained the optimizer

$$V_{opt} = \arg\min_{V\in C} D(\mu, V),$$

we compute the large-deviations approximation as

$$\log p(T) \approx -\frac{1}{2\sigma^2}\int_0^T \left(\dot{V}_{opt}(t) + g(t)V_{opt}(t) - I(t)\right)^2 dt.$$

This integral may be computed by a straightforward Euler rule; note that here it is slightly more convenient to use a rectangular than a trapezoidal integration rule, since $\dot{V}_{opt}$, g, and I are most conveniently defined on the intervals between the points $(0, \Delta, 2\Delta, \ldots, T)$.

6 Computing gradients

The key advantage of the matrix formulation of the integral equation described in Section 4 is that the equation $A_\theta p_\theta = b_\theta$ is easy to differentiate with respect to θ; that is, this formulation permits the efficient computation of likelihood gradients. In particular, we can use the chain rule to compute the gradient of $p_\theta(t)$ with respect to θ via simple matrix perturbation techniques: we need only compute the derivative

$$\frac{\partial}{\partial\epsilon}[(A + \epsilon A')^{-1}(b + \epsilon b')]_{\epsilon=0},$$

where A' and b' are arbitrary perturbations of A and b, respectively. For this we have

$$\begin{aligned}(A + \epsilon A')^{-1}(b + \epsilon b') &= \left[A(I + \epsilon A^{-1}A')\right]^{-1}(b + \epsilon b')\\ &= \left[I + \epsilon A^{-1}A'\right]^{-1}A^{-1}(b + \epsilon b')\\ &= \left[I - \epsilon A^{-1}A' + o(\epsilon)\right]\\ &\quad \times A^{-1}(b + \epsilon b')\\ &= A^{-1}b + \epsilon\left[A^{-1}b' - A^{-1}\right.\\ &\qquad \left.\times A'A^{-1}b\right] + o(\epsilon).\end{aligned}$$

Alternatively, we may simply define the implicit derivatives

$$\begin{aligned}Ap &= b\\ \frac{\partial A}{\partial\theta}p + A\frac{\partial p}{\partial\theta} &= \frac{\partial b}{\partial\theta}\\ \frac{\partial p}{\partial\theta} &= A^{-1}\left[\frac{\partial b}{\partial\theta} - \frac{\partial A}{\partial\theta}p\right]\\ &= A^{-1}\left[\frac{\partial b}{\partial\theta} - \frac{\partial A}{\partial\theta}A^{-1}b\right]\end{aligned}$$

to obtain the same result in a more symbolic manner.

So given A^{-1}, the directional derivatives with respect to A' and b' are cheap to compute, and therefore the

[2] A different method, based on the theory of "small-ball" probabilities for Gaussian measures, exists to handle the opposite extreme, when the probability of C becomes small not because the mean is distant from C, but rather because σ is large compared to the scale of C. However, these large-σ asymptotics are less relevant in the neural setting (since it is known that somatic current noise is relatively small compared to the other nondeterministic components of the neural response Mainen and Sejnowski 1995) and will not be discussed further here.

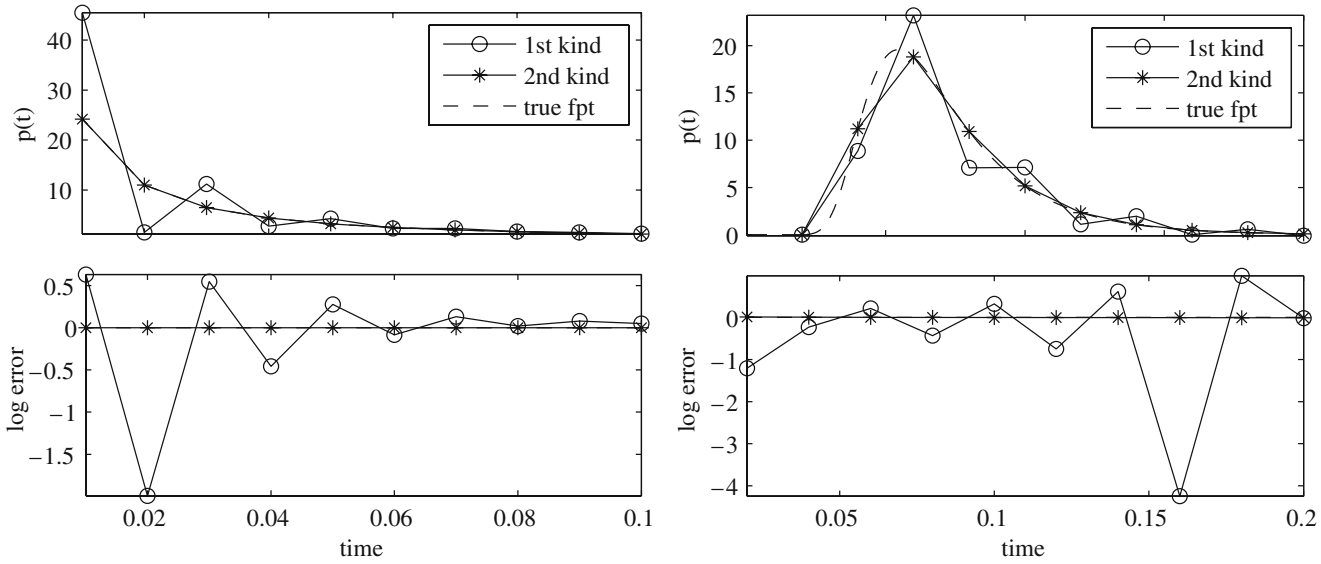

Fig. 1 Stability comparisons between first-kind and second-kind approaches. *Top*: first-passage time density $p(t)$ computed using a discretization depth $d = 10$. I and g were each constant here; *left*: $I = g = 0$ and $\sigma = 10$; *right*: $I = g = 40$, and $\sigma = 1$. The condition number of the first-kind matrix was > 20 times as large as was the condition number of the second-kind matrix in the case on the *right*. *Bottom*: log-ratio of computed to true $p(t)$. Note that the first-kind method has larger errors than does the second-kind method; in fact, the second-kind method is exact in the $g = 0$ case shown in the *left panels*

gradient with respect to the parameters may be obtained easily as well. Again, the most efficient notation for A^{-1} in Matlab is $A\backslash\text{eye}(\text{size}(A))$, which exploits the lower-triangular nature of A.

Note in particular that the likelihood is given by the last element of the vector p; thus, we do not need to compute the derivative of the full vector p with respect to A' and b', but rather only that of the last element. Thus, letting a_0 denote the bottom row of A^{-1},

$$\nabla_b p = a_0^t,$$

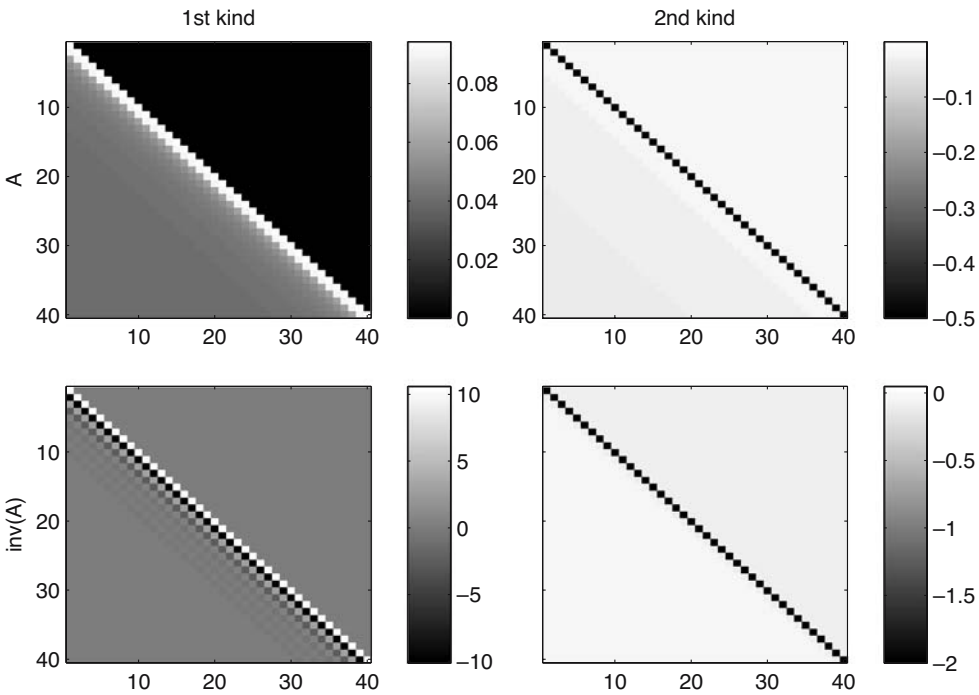

Fig. 2 Explanation of the stability differences between first-kind and second-kind approaches. *Top left*: A matrix in first-kind equation. *Top right*: A in second-kind equation. Note that the second-kind A is close to diagonal, decaying much more quickly away from the diagonal than does the first-kind A, with a much smaller condition number and therefore more stable estimates. *Bottom*: inverse matrices A^{-1}; note the oscillations in the first-kind case. Parameters: $I = 30$, $g = 40$, $\sigma = 5$, and $d = 40$

and

$$\nabla_A p = -(A^{-1}b)a_0.$$

No major conceptual difficulties arise in computing the gradients of A and b with respect to $I(s), g(s)$, and σ. As expected, everything can be computed in $O(d^3)$ time (although since no additional exponentials need to be computed, it turns out that even for d as large as 50, computing the gradient turns out to be only about as expensive as a single additional likelihood computation); an implementation of the code is available at http://www.stat.columbia.edu/~liam/research/code/basic-fpt-code.zip. We do need the gradients of the Gaussian function:

$$f(x, s, v) = v^{-1/2} \exp[-s(a-x)^2/2v];$$

$$\frac{\partial f}{\partial x} = s\frac{a-x}{v} f$$

$$\frac{\partial f}{\partial s} = -\frac{(a-x)^2}{2v} f$$

$$\frac{\partial f}{\partial v} = \left[-\frac{1}{2v} + s\frac{(a-x)^2}{2v^2}\right] f.$$

(Note that these gradients do not require any additional calls to the exponential function.) We also need the derivatives of u_1, etc., with respect to g:

$$\frac{\partial u_1}{\partial g} = -\Delta u_1$$

$$\frac{\partial u_2}{\partial g} = -2\Delta u_2$$

$$\frac{\partial v_1}{\partial g} = \begin{cases} \frac{g\Delta u_1 - 1 + u_1}{g^2}, & g(i\Delta) > 0 \\ -\Delta^2/2, & g(i\Delta) = 0 \end{cases}$$

$$\frac{\partial v_2}{\partial g} = \begin{cases} \frac{2g\Delta u_2 - 1 + u_2}{2g^2}, & g(i\Delta) > 0 \\ -\Delta^2, & g(i\Delta) = 0 \end{cases}$$

Again, note that we have avoided any further exponentiation. The gradients of φ with respect to σ^2, $I(s)$, and $g(s)$ follow similarly, albeit with a few more steps and applications of the chain rule; we skip the (unenlightening) details.

The gradient of the large deviation approximation described in the last section is comparatively easy to compute. Defining the function

$$Q(V, \theta) = -\frac{1}{2\sigma^2}\int_0^T \left(\dot{V}(t) + g(t)V(t) - I(t)\right)^2 dt,$$

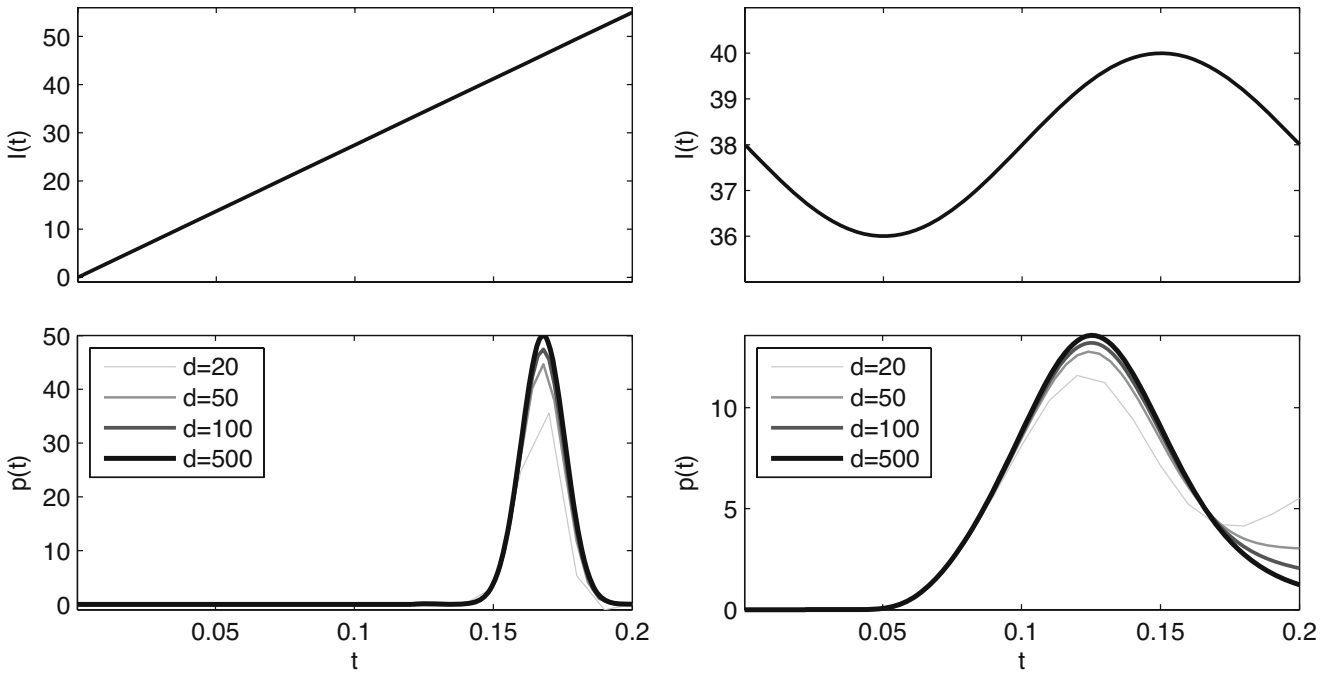

Fig. 3 Two examples of $p(t)$ computed via the second-kind method, given time-varying input current $I(t)$. *Left*: linear ramp $I(t)$. *Right*: sinusoidal $I(t)$. In each case, $\sigma = 0.5$ and $g = 40$. *Top*: input $I(t)$. *Bottom*: computed $p(t)$, at various settings for the discretization depth d. Note that $p(t)$ converges somewhat slowly in the case of a rapidly-varying $I(t)$; thus it is reasonable to assign large d to longer interspike intervals or on intervals where $I(t)$ or $g(t)$ vary quickly, and assign small d to intervals where I and g are relatively constant (where the faster small-d approximation is expected to be more adequate)

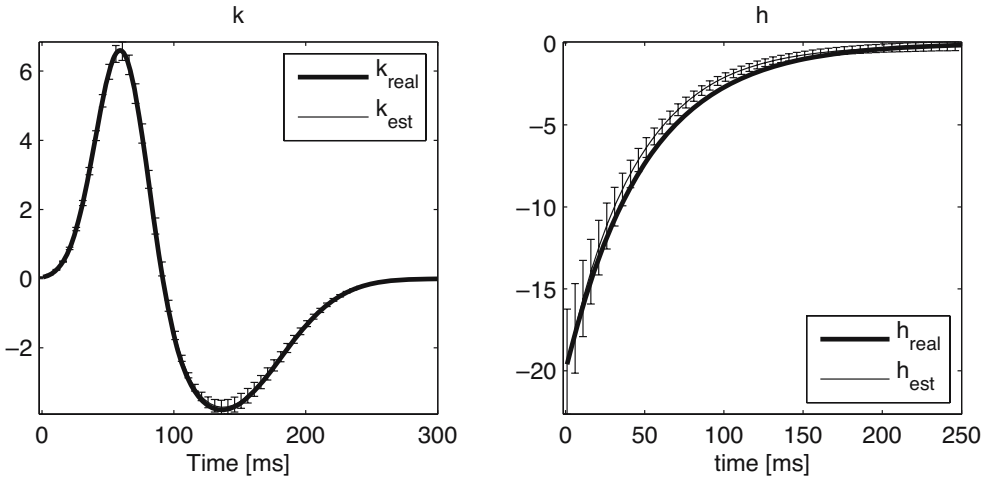

Fig. 4 Examples of the estimated parameters, $k(.)$ and $h(.)$. White noise current was filtered by the function $k(.)$ shown in *bold* in the *left panel*, then injected into an IF cell with parameters $g = 40, \sigma = 0.05$, DC input current $I_0 = 35$, and afterhyperpolarizing current $h(.)$ shown in *bold* on the *right*. Seventy-five spikes were generated and used to fit the unknown model parameters. Estimated $k(.)$ and h are shown for comparison; each function was estimated in a lower-dimensional subspace via maximum likelihood. *Errorbars* correspond to the estimated standard error, and were computed using the observed Fisher information method (van der Vaart 1998); we numerically computed the negative Hessian of the log-likelihood as a function of the model parameters, evaluated at the maximum likelihood value of the model parameters, then approximated the covariance of the estimate as the inverse of this negative Hessian matrix. Finally, we obtained the errorbars by the usual propagation-of-error formulas, followed by a square-root

we use the chain rule to write the gradient of the approximation as

$$\begin{aligned}\nabla_\theta \log p(T) &= \nabla_\theta Q(V_{opt}(\theta), \theta) \\ &= \frac{\partial V_{opt}}{\partial \theta} \nabla_1 Q(V_{opt}(\theta), \theta) + \nabla_2 Q(V_{opt}(\theta), \theta) \\ &= \nabla_2 Q(V_{opt}(\theta), \theta),\end{aligned}$$

where the last equality follows from the fact that V_{opt} optimizes $D(\mu, V)$, and therefore $\frac{\partial V_{opt}}{\partial \theta}$ is orthogonal to $\nabla_1 Q(V_{opt}(\theta), \theta)$, by the Karush–Kuhn–Tucker conditions. Computing

$$\nabla_2 Q(V_{opt}(\theta), \theta) = \nabla_\theta \left(-\frac{1}{2\sigma^2} \int_0^T \left(\dot{V}_{opt}(t) + g(t) V_{opt}(t) - I(t) \right)^2 dt \right)$$

with V_{opt} held fixed, is now straightforward; e.g., $Q(V_{opt}(\theta), \theta)$ is simply quadratic in $I(t)$ and $g(t)$.

7 Numerical results

We found that the second-kind integral equation method due to DiNardo et al. (2001) is more robust than the basic first-kind equation (Fig. 1; c.f. Plesser and Tanaka 1997); this relative instability of the first-kind equation is well-known (Lamm 2000). Note that this behavior is not universal; for some parameter settings, the first-kind equation gives more accurate results. However, on balance it appears that the second-kind equation has a more stable solution, as predicted (Lamm 2000). In particular, the matrix A in the second-kind equation consistently has a smaller (more stable) condition number than does the first-kind A; this gap becomes particularly pronounced when σ is large, where $f_1(t-s)$ falls off slowly as $|t-s|$ becomes large. See Fig. 2 for a comparison of these two matrices; Fig. 3 shows estimates of $p(t)$ in the case of two simple example time-varying currents $I(t)$ (a linear ramp and a sine-wave current).

Since the solution to the second-kind equation is generally more stable, and is guaranteed to give the correct result in the special case of $g = 0$, we use the second-kind equation in our demonstration of the performance of the MLE in simulated data (Fig. 4). We fit the model described in Paninski et al. (2004b) and Pillow et al. (2005), with $V(t)$ solving Eq. (1) (with constant membrane conductance g) and the input current given by

$$I(t) = I_0 + \int_{-\infty}^{t} x(\tau) k(t-\tau) d\tau + \sum_j h(t - t_j),$$

where $x(t)$ was a (fully-observed) white noise stimulus, the sum over $h(.)$ is over all past spike times t_j, and the true values of the stimulus filter $k(.)$ and spike-history filter $h(.)$ are shown in Fig. 4. We see that the integral equation method provides good estimates

of the true parameters, with the advantage of easily-computed gradients (note that the gradient of the likelihood with respect to $I(t)$ can easily be translated into gradients with respect to $k(.)$, I_0, and $h(.)$, since $I(t)$ is a linear function of these parameters), which speeds optimization and also the computation of the negative Hessian of the log-likelihood (the observed Fisher information, Schervish 1995), for the purposes of computing confidence intervals around the MLE.

8 Conclusions

We have adapted exact integral equation methods (DiNardo et al. 2001; Plesser and Tanaka 1997) and approximate quadratic-programming methods (Freidlin and Wentzell 1984; Paninski 2006) for computing spiking likelihoods in the stochastic integrate-and-fire neuron in order that the gradient of the likelihood may be efficiently computed, for optimization purposes. We found that an integral equation of the second kind required the computation of a few additional terms but provided solutions that were significantly more stable than the first-kind method. In particular, the second-kind integral equation method is acceptably fast and stable for slowly-changing $I(t)$ and $g(t)$, with speed $O((\frac{T}{dt})^2)$ (or $O((\frac{T}{dt})^3)$ if the gradient is computed) and accuracy $O(\frac{1}{dt})$ (and in fact perfect accuracy in the special case of constant input current and zero leak). Finally, this second-kind method provided us with an accurate and efficient code for computing the ML estimator for the integrate-and-fire model used in Paninski et al. (2004b) and Pillow et al. (2005).

Acknowledgements This work was partially supported by funding from the Gatsby Charitable Trust and by a Royal Society International Research Fellowship to LP. AH is funded by UK EPSRC/MRC at the Neuroinformatics Doctoral Training Centre, University of Edinburgh. We thank C.K.I. Williams for valuable conversations throughout the project and J. Fisher for helpful feedback on an early draft of the manuscript.

References

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. New York: Oxford University Press.

Buoncore, A., Nobile, A., & Ricciardi, L. (1987). A new integral equation for the evaluation of first-passage-time probability densities. *Advances in Applied Probability*, *19*, 784–800.

Burkitt, A., & Clark, G. (1999). Analysis of integrate-and-fire neurons: Synchronization of synaptic input and spike output. *Neural Computation*, *11*, 871–901.

Dembo, A., & Zeitouni, O. (1993). *Large deviations techniques and applications*. New York: Springer.

DiNardo, E., Nobile, A., Pirozzi, E., & Ricciardi, L. (2001). A computational approach to first-passage-time problems for Gauss–Markov processes. *Advances in Applied Probability*, *33*, 453–482.

Freidlin, M., & Wentzell, A. (1984). *Random perturbations of dynamical systems*. Berlin Heidelberg New York: Springer.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, *1*, 141–149.

Haith, A. (2004). *Estimating the parameters of a stochastic integrate-and-fire neural model.* Master's thesis, University of Edinburgh.

Haskell, E., Nykamp, D., & Tranchina, D. (2001). Population density methods for large-scale modelling of neuronal networks with realistic synaptic kinetics. *Network: Computation in Neural Systems*, *12*, 141–174.

Iyengar, S., & Liao, Q. (1997). Modeling neural activity using the generalized inverse Gaussian distribution. *Biological Cybernetics*, *77*, 289–295.

Jolivet, R., Lewis, T., & Gerstner, W. (2004). Generalized integrate-and-fire models of neuronal activity approximate spike trains of a detailed model to a high degree of accuracy. *Journal of Neurophysiology*, *92*, 959–976.

Karatzas, I., & Shreve, S. (1997). *Brownian motion and stochastic calculus*. Berlin Heidelberg New York: Springer.

Karlin, S., & Taylor, H. (1981). *A second course in stochastic processes*. New York: Academic.

Knight, B., Omurtag, A., & Sirovich, L. (2000). The approach of a neuron population firing rate to a new equilibrium: an exact theoretical result. *Neural Computation*, *12*, 1045–1055.

Lamm, P. (2000). A survey of regularization methods for first-kind Volterra equations. In *Surveys on solution methods for inverse problems* (pp. 53–82). Berlin Heidelberg New York: Springer.

Mainen, Z., & Sejnowski, T. (1995). Reliability of spike timing in neocortical neurons. *Science*, *268*, 1503–1506.

Paninski, L. (2006). The most likely voltage path and large deviations approximations for integrate-and-fire neurons. *Journal of Computational Neuroscience*, *21*, 71–87.

Paninski, L., Pillow, J., & Simoncelli, E. (2004a). Comparing integrate-and-fire-like models estimated using intracellular and extracellular data. *Neurocomputing*, *65*, 379–385.

Paninski, L., Pillow, J., & Simoncelli, E. (2004b). Maximum likelihood estimation of a stochastic integrate-and-fire neural model. *Neural Computation*, *16*, 2533–2561.

Pillow, J., Paninski, L., Uzzell, V., Simoncelli, E., & Chichilnisky, E. (2005). Accounting for timing and variability of retinal ganglion cell light responses with a stochastic integrate-and-fire model. *Journal of Neuroscience*, *25*, 11003–11013.

Plesser, H., & Gerstner, W. (2000). Noise in integrate-and-fire neurons: From stochastic input to escape rates. *Neural Computation*, *12*, 367–384.

Plesser, H., & Tanaka, S. (1997). Stochastic resonance in a model neuron with reset. *Physics Letters. A*, *225*, 228–234.

Press, W., Teukolsky, S., Vetterling, W., & Flannery, B. (1992). *Numerical recipes in C*. Cambridge: Cambridge University Press.

Ricciardi, L. (1977). *Diffusion processes and related topics in biology*. Berlin Heidelberg New York: Springer.

Schervish, M. (1995). *Theory of statistics*. New York: Springer.

Siegert, A. (1951). On the first passage time probability problem. *Physical Review*, *81*, 617–623.

Stevens, C., & Zador, A. (1998). Novel integrate-and-fire-like model of repetitive firing in cortical neurons. In *Proc. 5th Joint Symp. Neural Computation, UCSD*.

Tuckwell, H. (1989). *Stochastic processes in the neurosciences*. Philadelphia, PA: SIAM.

van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge: Cambridge University Press.