10

Estimating parameters of probabilistic neuron models

It is helpful to break neural data analysis into two basic problems. The "encoding" problem concerns how information is encoded in neural spike trains: can we predict the spike trains of a neuron (or population of neurons), given an arbitrary synaptic input, current injection, or sensory stimulus? Conversely, the "decoding" problem concerns how much we can learn from the observation of a sequence of spikes: in particular, how well can we estimate the stimulus that gave rise to the spike train?

The problems of encoding and decoding are difficult both because neural responses are stochastic and because we want to identify these response properties given any possible stimulus in some very large set (e.g., all images that might occur in the world), and there are typically many more such stimuli than we can hope to sample by brute force. Thus the neural coding problem is fundamentally *statistical*: given a finite number of samples of noisy physiological data, how do we estimate, in a global sense, the neural codebook?

This basic question has taken on a new urgency as neurophysiological recordings allow us to peer into the brain with ever greater facility: with the development of fast computers, inexpensive memory, and large-scale multineuronal recording and high-resolution imaging techniques, it has become feasible to directly observe and analyze neural activity at a level of detail that was impossible in the twentieth century. Experimenter now routinely record from hundreds of neurons simultaneously, providing great challenges for data analysis by computational neuroscientists and statisticians. Indeed, it has become clear that sophisticated statistical techniques are necessary to understand the neural code: many of the key questions cannot be answered without powerful statistical tools.

This chapter describes statistical model-based techniques that provide a unified approach to both encoding and decoding. These *statistical* models can capture stimulus dependencies as well as spike history and interneuronal interaction effects in population of spike trains, and are intimately related to the generalized integrate-and-fire models discussed in previous chapters.

In Section 10.1, we establish the notation that enables us to identify relevant model parameters and introduce the concept of parameter optimization in a linear model. We then leave the realm of linear models and turn to the models that we have discussed in preceding chapters (e.g., the Spike Response Model with escape noise in Chapter 9) where spike generation is stochastic and nonlinear.

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

In Section 10.2, we describe the same neuron models of spike trains in the slightly more abstract language of statistics. The likelihood of a spike train given the stimulus plays a central role in statistical models of encoding. As we have seen in Chapter 9, the stochasticity introduced by "escape noise" in the Spike Response Model (SRM) or other generalized integrate-and-fire models enables us to write down the likelihood that an observed spike train could have been generated by the neuron model. Likelihood-based optimization methods for fitting these neuron models to data allow us to predict neuronal spike timing for future, unknown stimuli. Thus, the SRM and other generalized integrate-and-fire models can be viewed as *encoding* models. In Chapter 11 we shall see that the same models can also be used to perform optimal decoding.

The emphasis of this chapter is on likelihood-based methods for model optimization. The likelihood-based optimization methods are computationally tractable, due to a key concavity property of the model likelihood (Paninski, 2004). However, likelihood is just one of several quantities that can be chosen to compare spike trains, and other measures to quantify the performance of models can also be used. In Section 10.3 we review different performance measures for the "goodness-of-fit" of a model. In particular, we present the notion of "spike train similarity" and the "time rescaling theorem" (Brown *et al.*, 2002).

Finally, in Section 10.4 we apply the ideas developed in this chapter to adaptively choose the optimal stimuli for characterizing the response function.

10.1 Parameter optimization in linear and nonlinear models

Before we turn to the statistical formulation of models of encoding and decoding, we need to introduce the language of statistics into neuron modeling. In particular, we will define what is meant by convex problems, optimal solutions, *linear* models and *generalized linear* models.

When choosing a neuron model for which we want to estimate the parameters from data, we must satisfy three competing desiderata:

(i) The model must be flexible and powerful enough to fit the observed data. For example, a linear model might be easy to fit, but not powerful enough to account for the data.

(ii) The model must be tractable: we need to be able to fit the model given the modest amount of data available in a physiological recording (preferably using modest computational resources as well); moreover, the model should not be so complex that we cannot assign an intuitive functional role to the inferred parameters.

(iii) The model must respect what is already known about the underlying physiology and anatomy of the system; ideally, we should be able to interpret the model parameters and predictions not only in statistical terms (e.g., confidence intervals, significance tests) but also in biophysical terms (membrane noise, dendritic filtering, etc.). For example, with a purely statistical "black box" model we might be able to make predictions and test their significance, but we will not be able to make links to the biophysics of neurons.

While in general there are many varieties of encoding models that could conceivably

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

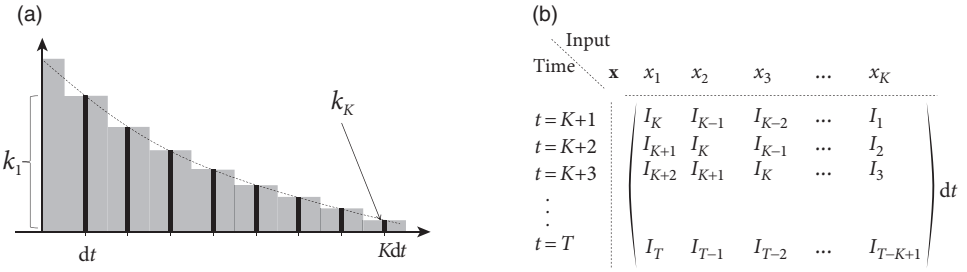

Fig. 10.1 Measurement of membrane filter. (a) Schematic of the linear membrane filter κ in discrete time. (b) Matrix of temporal inputs (schematic). At each moment in time, the last K time steps of the input current I_t serve as an input vector. Rows of the matrix correspond to different input samples.

satisfy these three conditions, in this chapter we will mainly focus on the SRM with escape noise (Chapter 9). In a more general setting (see Chapter 11), the linear filter can be interpreted not just as local biophysical processes within the neuron, but as a summary of the whole signal processing chain from sensory inputs to the neuron under consideration. In such a general setting, the SRM may also be seen as an example of a "generalized linear" model (GLM) (Paninski, 2004; Truccolo *et al.*, 2005). In the following two subsections, we review linear and Generalized Linear Models from the point of view of neuronal dynamics: how is a stimulus $I(t)$ encoded by the neuron?

10.1.1 Linear models

Let us suppose that an experimenter injects, with a first electrode, a time-dependent current $I(t)$ in an interval $0 < t \leq T$ while recording with a second electrode the membrane voltage $u^{\exp}(t)$. The maximal amplitude of the input current has been chosen small enough for the neuron to stay in the subthreshold regime. We may therefore assume that the voltage is well described by our linear model

$$u(t) = \int_0^\infty \kappa(s) I(t-s)\,\mathrm{d}s + u_{\rm rest}; \tag{10.1}$$

see Section 1.3.5. In order to determine the filter κ that describes the linear properties of the experimental neuron, we discretize time in steps of $\mathrm{d}t$ and denote the voltage measurement and injected current at time t by $u_t^{\exp}$ and I_t, respectively. Here the time subscript $t \in \mathbb{Z}$ is an integer time step counter. We set $K = s^{\max}/\mathrm{d}t$ where $\max(s) \in \mathbb{N}$ and introduce a vector

$$\boldsymbol{k} = (\kappa(\mathrm{d}t), \kappa(2\mathrm{d}t), \ldots, \kappa(K\mathrm{d}t)) \tag{10.2}$$

which describes the time course κ in discrete time; see Fig. 10.1a. Similarly, the input current I during the last K time steps is given by the vector

$$\boldsymbol{x}_t = (I_{t-1}, I_{t-2}, \ldots, I_{t-K})\,\mathrm{d}t\,. \tag{10.3}$$

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

The discrete-time version of the integral equation (10.1) is then a simple scalar product

$$u_t = \sum_{l=1}^{K} k_l I_{t-l} \mathrm{d}t + u_{\mathrm{rest}} = \boldsymbol{k} \cdot \boldsymbol{x}_t + u_{\mathrm{rest}} \,. \tag{10.4}$$

Note that $\boldsymbol{k}$ is the vector of parameters $k_1, k_2, \ldots, k_K$ that need to be estimated. In the language of statistics, Eq. (10.4) is a linear model because the observable u_t is *linear in the parameters*. Moreover, u_t is a continuous variable so that the problem of estimating the parameters falls in the class of *linear regression* problems. More generally, regression problems refer to the prediction or modeling of continuous variables whereas classification problems refer to the modeling or prediction of discrete variables.

To find a good choice of parameters $\boldsymbol{k}$, we compare the prediction u_t of the model equation (10.4) with the experimental measurement u_t^{exp}. In a least-square error approach, the components of the vector $\boldsymbol{k}$ will be chosen such that the squared difference between model voltage and experimental voltage

$$E(\boldsymbol{k}) = \sum_{t=K+1}^{T} \left(u_t^{\mathrm{exp}} - u_t\right)^2 \tag{10.5}$$

is minimal. An important insight is the following. For any model that is linear in the parameters, the function E in Eq. (10.5) is *quadratic and convex* in the parameters $\boldsymbol{k}$ of the model. Therefore

(i) the function E has no non-global local minima as a function of the parameter vector $\boldsymbol{k}$ (in fact, the set of minimizers of E forms a linear subspace in this case, and simple conditions are available to verify that E has a single global minimum, as discussed below);

(ii) the minimum can be found either numerically by gradient descent or analytically by matrix inversion.

While the explicit solution is only possible for linear models, the numerical gradient descent is possible for all kinds of error functions E and yields a unique solution if the error has a *unique* minimum. In particular, for all error functions which are convex, gradient descent converges to the optimal solution (Fig. 10.2) – and this is what we will exploit in Section 10.2.

Example: Analytical solution

For the analytical solution of the least-square optimization problem, defined by Eqs. (10.4) and (10.5), it is convenient to collect all time points u_t, $K+1 < t < T$ into a single vector $u = (u_{K+1}, u_{K+2}, \ldots, u_T)$ which describes the membrane voltage of the model. Similarly, the observed voltage in the experiment is summarized by the vector $u^{\mathrm{exp}} = (u_{K+1}^{\mathrm{exp}}, \ldots, u_T^{\mathrm{exp}})$. Furthermore, let us align the observed input vectors $\boldsymbol{x}$ into a matrix X. More precisely, the matrix has $T-K$ rows consisting of the vector $\boldsymbol{x}_t$; see Fig. 10.1b. With this notation, Eq. (10.4) can be written as a matrix equation

$$u = X\boldsymbol{k}^{\mathrm{T}} + u_{\mathrm{rest}}, \tag{10.6}$$

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

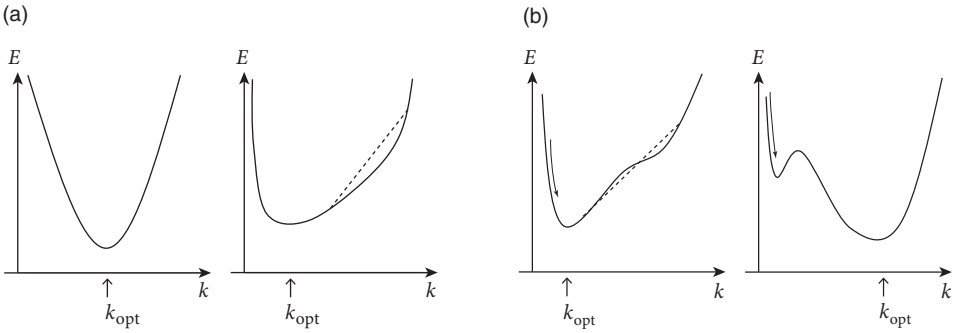

Fig. 10.2 Convex function and global minimum. (a) A quadratic function (left) and an arbitrary convex function (right). A convex function is curved upward so that any straight line (dashed) connecting two points on the curve stays above the curve. A convex function cannot have a non-global minimum. (b) A non-convex function without (left) and with (right) a non-global minimum. Gradient descent refers to a change of the parameter k that leads to a downward move (arrow) on the error surface. In the case on the right, a gradient-descent method can get stuck in the local minimum.

where u_{rest} is a vector with all components equal to u_{rest}. We suppose that the value of u_{rest} has already been determined in an earlier experiment.

We search for the minimum of Eq. (10.5), defined by $\nabla_{\boldsymbol{k}} E = 0$ (where $\nabla_{\boldsymbol{k}} E$ denotes the gradient of E with respect to $\boldsymbol{k}$), which gives rise a single linear equation for each component of the parameter vector $\boldsymbol{k}$, i.e., a set of K linear equations. With our matrix notation, the error function is a scalar product

$$E(\boldsymbol{k}) = [u^{\text{exp}} - X\boldsymbol{k} - u_{\text{rest}}]^{\text{T}} \cdot [u^{\text{exp}} - X\boldsymbol{k} - u_{\text{rest}}] \tag{10.7}$$

and the unique solution of the set of linear equations is the parameter vector

$$\hat{\boldsymbol{k}}_{\text{LS}} = (X^{\text{T}}X)^{-1} X^{\text{T}}(u^{\text{exp}} - u_{\text{rest}}), \tag{10.8}$$

assuming the matrix $(X^{\text{T}}X)$ is invertible. (If this matrix is non-invertible, then a unique minimum does not exist.) The subscript highlights that the parameter $\hat{\boldsymbol{k}}_{\text{LS}}$ has been determined by least-square optimization.

10.1.2 Generalized Linear Models

The above linearity arguments not only work in the subthreshold regime, but can be extended to the case of spiking neurons. In the deterministic formulation of the Spike Response Model, the membrane voltage is given as

$$u(t) = \int_0^\infty \eta(s) S(t-s) \mathrm{d}s + \int_0^\infty \kappa(s) I(t-s)\, \mathrm{d}s + u_{\text{rest}}, \tag{10.9}$$

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

where $S(t) = \sum_f \delta(t - t^f)$ is the spike train of the neuron; see Eq. (1.22) or (9.1). Similarly to the passive membrane, the input current enters linearly with a membrane filter κ. Similarly, past output spikes $t^f < t$ enter linearly with a "refractory kernel" or "adaptation filter" η. Therefore, spike history effects are treated in the SRM as linear contributions to the membrane potential. The time course of the spike history filter η can therefore be estimated analogously to that of κ.

Regarding the subthreshold voltage, we can generalize Eqs. (10.2)–(10.4). Suppose that the spike history filter η extends over a maximum of J time steps. Then we can introduce a new parameter vector

$$\boldsymbol{k} = (\kappa(\mathrm{d}t), \kappa(2\mathrm{d}t), \ldots, \kappa(K\mathrm{d}t), \eta(\mathrm{d}t), \eta(2\mathrm{d}t), \ldots, \eta(J\mathrm{d}t), u_{\mathrm{rest}}) \tag{10.10}$$

which includes both the membrane filter κ and the spike history filter η. The spike train in the last J time steps is represented by the spike count sequence $n_{t-1}, n_{t-2}, \ldots, n_{t-J}$, where $n_t \in \{0, 1\}$, and included into the "input" vector

$$\boldsymbol{x}_t = (I_{t-1}\mathrm{d}t, I_{t-2}\mathrm{d}t, \ldots, I_{t-K}\mathrm{d}t, n_{t-1}, n_{t-2}, \ldots, n_{t-J}, 1). \tag{10.11}$$

The discrete-time version of the voltage equation in the SRM is then again a simple scalar product

$$u_t = \sum_{j=1}^{J} k_{K+j} n_{t-j} + \sum_{k=1}^{K} k_k I_{t-k}\mathrm{d}t + u_{\mathrm{rest}} = \boldsymbol{k} \cdot \boldsymbol{x}_t \,. \tag{10.12}$$

Thus, the membrane voltage during the interspike intervals is a linear regression problem that can be solved as before by minimizing the mean square error.

Spiking itself, however, is a *nonlinear* process. In the SRM with escape rate, the firing intensity is

$$\rho(t) = f(u(t) - \vartheta) = f(\boldsymbol{k} \cdot \boldsymbol{x}_t - \vartheta) \,. \tag{10.13}$$

We have assumed that the firing threshold is constant, but this is no limitation since, in terms of spiking, any dynamic threshold can be included into the spike-history filter η.

We emphasize that the firing intensity in Eq. (10.13) is a *nonlinear* function of the parameters $\boldsymbol{k}$ and b that we need to estimate. Nevertheless, rapid parameter estimation is still possible if the function f has properties that we will identify in Section 10.2. The reason is that in each time step firing is stochastic with an instantaneous firing intensity f that only depends on the momentary value of the membrane potential – where the membrane potential can be written as a *linear* function of the parameters. This insight leads to the notion of Generalized Linear Models (GLM). For an SRM with exponential escape noise $\rho(t) = f(u(t) - \vartheta) = \rho_0 \exp(u(t) - \vartheta)$ the likelihood of a spike train

$$L^n(\{t^{(1)}, t^{(2)}, \ldots, t^{(n)}\}) = \rho(t^{(1)}) \cdot \rho(t^{(2)}) \cdot \ldots \cdot \rho(t^{(n)}) \exp\left[- \int_0^T \rho(s)\mathrm{d}s \right] , \tag{10.14}$$

which we have already defined in Eq. (9.10), is a log-concave function of the parameters (Paninski, 2004); i.e., the loglikelihood is a concave function. We will discuss this result in

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

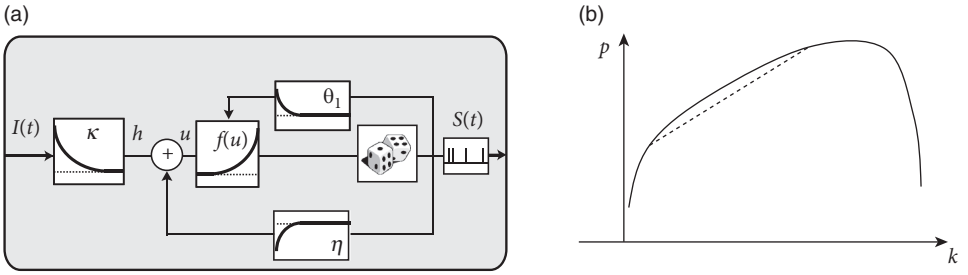

Fig. 10.3 SRM revisited. (a) The SRM takes as input a time-dependent current $I(t)$ and generates a spike train $S(t)$ at the output. The parameters of the model control the shape of the filters κ, θ_1 and η. (b) If the escape rate $f(u-\vartheta)$ is exponential and the parameters k enter linearly into the voltage equation u and the threshold ϑ, then the likelihood p that a specific spike train is generated by the model is a concave (i.e., downward curving) function of the parameters (Paninski, 2004).

the next section as it is the fundamental reason why parameter optimization for the SRM is computationally efficient (Fig. 10.3).

GLMs are fundamental tools in statistics for which a great deal of theory and computational methods are available. In what follows we exploit the elegant mathematical properties of GLMs.

10.2 Statistical formulation of encoding models

Let us denote the observed spike train data by D. In general, D could represent measurements from a population of neurons. To keep the arguments simple, we focus for the moment on a single neuron, but at the end of the section we will extend the approach to a population of interacting neurons. If the time bins are chosen smaller than the absolute refractory period, the discretized spike train is described by a sequence of scalar variables n_t in an interval $0 < t \leq T$

$$D = \{n_1, n_2, \ldots, n_T\} \tag{10.15}$$

see Fig. 9.6. If time bins are large, spike counts can take values larger than 1.

A neural "encoding model" is a model that assigns a conditional probability, $p(D|\boldsymbol{x})$, to any possible neural response D given a stimulus $\boldsymbol{x}$. The vector $\boldsymbol{x}_t$ can include the momentary stimulus presented at time t, or more generally the concatenated spatio-temporal stimulus history up to time t. Examples have been given in Section 10.1.

As emphasized in the introduction to this chapter, it is not feasible to directly measure this probability $p(D|\boldsymbol{x})$ for all stimulus-response pairs $(\boldsymbol{x}, D)$, simply because there are infinitely many potential stimuli. We therefore hypothesize some encoding model,

$$p(D|\boldsymbol{x}, \theta)\,. \tag{10.16}$$

Here θ is a short-hand notation for the set of all model parameters. In the examples of the previous section, the model parameters are $\theta = \{\boldsymbol{k}, b\}$.

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

Our aim is to estimate the model parameters θ so that the model "fits" the observed data D. Once θ is in hand we may compute the desired response probabilities as

$$p(D|\boldsymbol{x}) \approx p(D|\boldsymbol{x},\theta)\,, \tag{10.17}$$

i.e., knowing θ allows us to interpolate between the observed (noisy) stimulus-response pairs, in order to predict the response probabilities for novel stimuli $\boldsymbol{x}$ for which we have not yet observed any responses.

10.2.1 Parameter estimation

How do we find a good estimate for the parameters θ for a chosen model class? The general recipe is as follows. The first step is to introduce a model that makes sense biophysically, and incorporates our prior knowledge in a tractable manner. Next we write down the likelihood of the observed data given the model parameters, along with a prior distribution that encodes our prior beliefs about the model parameters. Finally, we compute the posterior distribution of the model parameters given the observed data, using Bayes' rule, which states that

$$p(\theta|D) \propto p(D|\theta)p(\theta); \tag{10.18}$$

the left-hand side is the desired posterior distribution, while the right-hand side is just the product of the likelihood and the prior.

In the current setting, we need to write down the likelihood $p(D|X,\boldsymbol{k})$ of the observed spike data D given the model parameter $\boldsymbol{k}$ and the observed set of stimuli summarized in the matrix X, and then we may employ standard likelihood optimization methods to obtain the maximum likelihood (ML) or maximum a posteriori (MAP) solutions for $\boldsymbol{k}$, defined by

$$\text{ML:}\quad \hat{\boldsymbol{k}}_{\text{ML}} = \text{argmax}_{\boldsymbol{k}}\{p(D|X,\boldsymbol{k})\}, \tag{10.19}$$

$$\text{MAP:}\quad \hat{\boldsymbol{k}}_{\text{MAP}} = \text{argmax}_{\boldsymbol{k}}\{p(D|X,\boldsymbol{k})\,p(\boldsymbol{k})\}, \tag{10.20}$$

where the maximization runs over all possible parameter choices.

We assume that spike counts per bin follow a conditional Poisson distribution, given $\rho(t)$:

$$n_t \sim \text{Poiss}[\rho(t)\text{d}t]\,; \tag{10.21}$$

see text and exercises of Chapter 7. For example, with the rate parameter of the Poisson distribution given by a GLM or SRM model $\rho(t) = f(\boldsymbol{k}\cdot\boldsymbol{x}_t)$, we have

$$p(D|X,\boldsymbol{k}) = \prod_t \left\{ \frac{[f(\boldsymbol{k}\cdot\boldsymbol{x}_t)\text{d}t]^{n_t}}{(n_t)!} \exp[-f(\boldsymbol{x}_t\cdot\boldsymbol{k})\text{d}t] \right\}. \tag{10.22}$$

Here Π_t denotes the product over all time steps. We recall that, by definition, $n_t! = 1$ for $n_t = 0$.

Our aim is to optimize the parameter $\boldsymbol{k}$. For a given observed spike train, the spike count numbers n_t are fixed. In this case, $(\text{d}t)^{n_t}/(n_t)!$ is a constant which is irrelevant for the

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

parameter optimization. If we work with a fixed time step $\mathrm{d}t$ and drop all units, we can therefore reshuffle the terms and consider the logarithm of the above likelihood

$$\log p(D|X,\boldsymbol{k}) = c_0 + \sum_t (n_t \log f(\boldsymbol{k}\cdot\boldsymbol{x}_t) - f(\boldsymbol{x}_t\cdot\boldsymbol{k})\mathrm{d}t)\,. \tag{10.23}$$

If we choose the time step $\mathrm{d}t$ shorter than the half-width of an action potential, say 1 ms, the spike count variable n_t can only take the values zero or 1. For small $\mathrm{d}t$, the likelihood of Eq. (10.22) is then identical to that of the SRM with escape noise, defined in Chapter 9; see Eqs. (9.10) and (9.15)–(9.17).

We don't have an analytical expression for the maximum of the likelihood defined in Eq. (10.22), but nonetheless we can numerically optimize this function quite easily if we are willing to make two assumptions about the nonlinear function $f(.)$. More precisely, if we assume that

(i) $f(u)$ is a convex (upward-curving) function of its scalar argument u, and

(ii) $\log f(u)$ is concave (downward-curving) in u,

then the log likelihood above is guaranteed to be a concave function of the parameter $\boldsymbol{k}$, since in this case the log-likelihood is just a sum of concave functions of $\boldsymbol{k}$ (Paninski, 2004).

This implies that the likelihood has no non-global maximum (also called local maximum). Therefore the maximum likelihood parameter $\hat{\boldsymbol{k}}_{\mathrm{ML}}$ may be found by numerical ascent techniques; see Fig. 10.2. Functions $f(.)$ satisfying these two constraints are easy to think of: for example, the standard linear rectifier and the exponential function both work.

Fitting model parameters proceeds as follows: we form the (augmented) matrix X where each row is now

$$\boldsymbol{x}_t = (1, I_{t-1}\mathrm{d}t, I_{t-2}\mathrm{d}t, \ldots, I_{t-K}\mathrm{d}t, n_{t-1}, n_{t-2}, \ldots, n_{t-J})\,. \tag{10.24}$$

Similarly, the parameter vector is in analogy to Eq. (10.10)

$$\boldsymbol{k} = (b, \kappa(\mathrm{d}t), \kappa(2\mathrm{d}t), \ldots, \kappa(K\mathrm{d}t), \eta(\mathrm{d}t), \eta(2\mathrm{d}t), \ldots, \eta(J\mathrm{d}t))\,; \tag{10.25}$$

here $b = u_{\mathrm{rest}} - \vartheta$ is a constant offset term which we want to optimize.

We then calculate the log-likelihood

$$\log p(D|X,\boldsymbol{k}) = \sum_t (n_t \log f(X_t\cdot\boldsymbol{k}) - f(X_t\cdot\boldsymbol{k})\mathrm{d}t) \tag{10.26}$$

and compute the ML or maximum *a posteriori* (MAP) solution for the model parameters $\boldsymbol{k}$ by a concave optimization algorithm. Note that, while we still assume that the conditional spike count n_t within a given short time bin is drawn from a one-dimensional Poiss $(\rho(t)\mathrm{d}t)$ distribution given $\rho(t)$, the resulting model displays strong history effects (since $\rho(t)$ depends on the past spike trains) and therefore the output of the model, considered as a vector of counts $D = \{n_t\}$, is no longer a Poisson process, unless $\eta = 0$. Importantly, because of the refractory effects incorporated by a strong negative η at small times, the spike count variable n_t cannot take a value larger than 1 if $\mathrm{d}t$ is in the range of one or a few milliseconds. Therefore, we can expect that interspike intervals are correctly reproduced in the model; see Fig. 10.5 for an example application.

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

Finally, we may expand the definition of X to include observations of other spike trains, and therefore develop GLMs not just of single spike trains, but network models of how populations of neurons encode information jointly (Chornoboy *et al.*, 1988; Paninski *et al.*, 2004; Truccolo *et al.*, 2005; Pillow *et al.*, 2008). The resulting model is summarized as follows: Spike counts are conditionally Poisson distributed given $\rho_i(t)$ $n_{i,t} \sim \text{Poiss}(\rho_i(t)\mathrm{d}t)$ with a firing rate

$$\rho_i(t) = f\left(\boldsymbol{k}_i \cdot \boldsymbol{x}_t + \sum_{i' \neq i, j} \varepsilon_{i',j} n_{i',t-j}\right). \tag{10.27}$$

Here, $\rho_i(t)$ denotes the instantaneous firing rate of the ith cell at time t and $\boldsymbol{k}_i$ is the cell's linear receptive field including spike-history effects; see Eq. (10.25). The net effect of a spike of neuron i' onto the membrane potential of neuron i is summarized by $\varepsilon_{i',j}$; these terms are summed over all past spike activity $n_{i',t-j}$ in the population of cells from which we are recording simultaneously. In the special case that we record from all neurons in the population, $\varepsilon_{i',j}$ can be interpreted as the excitatory or inhibitory postsynaptic potential caused by a spike of neuron i' a time $j\,\mathrm{d}t$ earlier.

Example: Linear regression and voltage estimation

It may be helpful to draw an analogy to linear regression here. We want to show that the standard procedure of least-square minimization can be linked to statistical parameter estimation under the assumption of Gaussian noise.

We consider the linear voltage model of Eq. (10.1). We are interested in the temporal filter properties of the neuron when it is driven by a time-dependent input $I(t)$. Let us set $\boldsymbol{x}_t = (I_t, I_{t-1}, \ldots, I_{t-K})\,\mathrm{d}t$ and $\boldsymbol{k} = (\kappa(\mathrm{d}t), \ldots, \kappa(K\,\mathrm{d}t))$. If we assume that the discrete-time voltage measurements have a Gaussian distribution around the mean predicted by the model of Eq. (10.4), then we need to maximize the likelihood

$$\log p(D|X, \boldsymbol{k}) = c_1 - c_2 \sum_t \left(u_t - (\boldsymbol{k} \cdot \boldsymbol{x}_t)\right)^2, \tag{10.28}$$

where $X = (x_1, x_2, \ldots, x_T)$ is the matrix of observed stimuli, c_1, c_2 are constants that do not depend on the parameter $\boldsymbol{k}$, and the sum in t is over all observed time bins. Maximization yields $\boldsymbol{k}_{\text{opt}} = (X^{\mathrm{T}}X)^{-1}\left(\sum_t u_t \boldsymbol{x}_t / \mathrm{d}t\right)$ which determines the time course of the filter $\kappa(s)$ that characterizes the passive membrane properties. The result is identical to Eq. (10.8):

$$\hat{\boldsymbol{k}}_{\text{opt}} = \hat{\boldsymbol{k}}_{\text{LS}}. \tag{10.29}$$

10.2.2 Regularization: maximum penalized likelihood

In the linear regression case it is well known that estimates of the components of the parameter vector $\boldsymbol{k}$ can be quite noisy when the dimension of $\boldsymbol{k}$ is large. The noisiness of the estimate $\hat{\boldsymbol{k}}_{\text{ML}}$ is roughly proportional to the dimensionality of $\boldsymbol{k}$ (the number of parameters

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

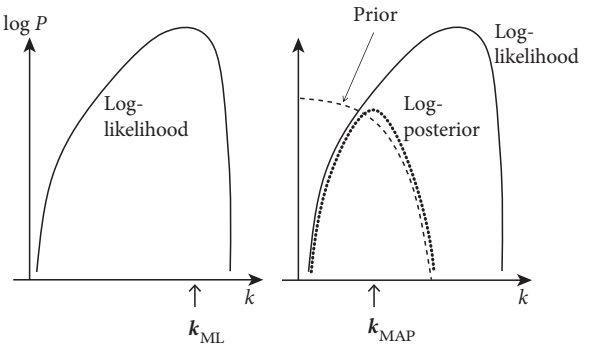

Fig. 10.4 Regularization. Because of the concavity, there is only a single global maximum of the log-likelihood which defines the optimal parameter choice $\boldsymbol{k}_{\mathrm{ML}}$. Right: Example of regularization by a prior (dashed line) that favors a smaller value for $\boldsymbol{k}_{\mathrm{MAP}}$.

in $\boldsymbol{k}$ that we need to estimate from data) divided by the total number of observed samples (Paninski, 2003). The same "overfitting" phenomenon (estimator variability increasing with number of parameters) occurs in the GLM context. A variety of methods have been introduced to "regularize" the estimated $\boldsymbol{k}$, to incorporate prior knowledge about the shape and/or magnitude of the true $\boldsymbol{k}$ to reduce the noise in $\hat{\boldsymbol{k}}_{\mathrm{ML}}$. One basic idea is to restrict $\boldsymbol{k}$ to lie within a lower-dimensional subspace; we then employ the same fitting procedure to estimate the coefficients of $\boldsymbol{k}$ within this lower-dimensional basis (model selection procedures may be employed to choose the dimensionality of this subspace (Truccolo *et al.*, 2005)).

A slightly less restrictive approach is to maximize the posterior

$$p(\boldsymbol{k}|X,D) \propto p(D|X,\boldsymbol{k})p(\boldsymbol{k}) \tag{10.30}$$

(with $\boldsymbol{k}$ allowed to take values in the full original basis), instead of the likelihood $p(D|X,\boldsymbol{k})$; here $p(\boldsymbol{k})$ encodes our *a priori* beliefs about the true underlying $\boldsymbol{k}$.

It is easy to incorporate this maxima *a posteriori* idea in the GLM context (Paninski, 2004): we simply maximize

$$\log p(\boldsymbol{k}|X,D) = c + \log p(\boldsymbol{k}) + \log p(D|X,\boldsymbol{k}) \tag{10.31}$$

$$= c + \log p(\boldsymbol{k}) + \sum_t \left(n_t \log f(\boldsymbol{x}_t \cdot \boldsymbol{k}) - f(\boldsymbol{x}_t \cdot \boldsymbol{k})\mathrm{d}t\right). \tag{10.32}$$

Whenever $\log p(\boldsymbol{k})$ is a concave function of $\boldsymbol{k}$, this "penalized" likelihood (where $\log p(\boldsymbol{k})$ acts to penalize improbable values of $\boldsymbol{k}$) is a concave function of $\boldsymbol{k}$, and ascent-based maximization may proceed (with no local maximum) as before; see Fig. 10.4.

Example: Linear regression and Gaussian prior

In the linear regression case, the computationally simplest prior is a zero-mean Gaussian, $\log p(\boldsymbol{k}) = c - \boldsymbol{k}^{\mathrm{T}} A \boldsymbol{k}/2$, where A is a positive definite matrix (the inverse

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

covariance matrix). The Gaussian prior can be combined with the Gaussian noise model of Eq. (10.28). Maximizing the corresponding posterior analytically (Sahani and Linden, 2003; Smyth *et al.*, 2003) then leads directly to the regularized least-square estimator

$$\hat{\boldsymbol{k}}_{\mathrm{RLS}} = (X^{\mathrm{T}}X + A)^{-1}\left(\sum_t u_t \boldsymbol{x}_t / \mathrm{d}t\right). \tag{10.33}$$

10.2.3 Fitting generalized integrate-and-fire models to data

Suppose an experimenter has injected a time-dependent current $I(t)$ into a neuron and has recorded with a second electrode the voltage $u^{\exp}(t)$ of the same neuron. The voltage trajectory contains spikes $S(t) = \sum_f \delta(t - t^f)$ with firing times $t^{(1)}, t^{(2)}, \ldots, t^{(N)}$. The natural approach would be to write down the joint likelihood of observing both the spike times and the subthreshold membrane potential (Paninski *et al.*, 2005). A simpler approach would be to maximize the likelihood of observing the spike separately from the likelihood of observing the membrane potential.

Given the input $I(t)$ and the spike train $S(t)$, the voltage of an SRM is given by Eq. (10.9) and we can adjust the parameters of the filter κ and η so as to minimize the squared error Eq. (10.5). We now fix the parameters for the membrane potential and maximize the likelihood of observing the spike times given our model voltage trajectory $u(t)$. We insert $u(t)$ into the escape rate function $\rho(t) = f(u(t) - \vartheta(t))$ which contains the parameters of the threshold

$$\vartheta(t) = \vartheta_0 + \int_0^\infty \theta_1(s) S(t-s) \mathrm{d}s. \tag{10.34}$$

We then calculate the log-likelihood

$$\log p(D|X, \boldsymbol{k}) = c + \sum_t (n_t \log f(X_t \cdot \boldsymbol{k}) - f(X_t \cdot \boldsymbol{k}) \mathrm{d}t) \tag{10.35}$$

and compute the ML or maximum *a posteriori* (MAP) solution for the model parameters $\boldsymbol{k}$ (which are the parameters of the threshold – the subthreshold voltage parameters are already fixed) by an optimization algorithm for concave functions.

Fig. 10.5 shows an example application. Both voltage in the subthreshold regime and spike times are nicely reproduced. Therefore, we can expect that interspike intervals are correctly reproduced as well. In order to quantify the performance of neuron models, we need to develop criteria of "goodness-of-fit" for subthreshold membrane potential, spike timings, and possibly higher-order spike-train statistics. This is the topic of Section 10.3; we will return to similar applications in the next chapter.

10.2.4 Extensions (*)

The GLM encoding framework described here can be extended in a number of important directions. We briefly describe two such directions here.

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

First, as we have described the GLM above, it may appear that the model is restricted to including only linear dependencies on the stimulus $\boldsymbol{x}_t$, through the $\boldsymbol{k} \cdot \boldsymbol{x}_t$ term. However, if we modify our input matrix X once again, to include nonlinear transformations $\mathscr{F}_j(\boldsymbol{x})$ of the stimulus $\boldsymbol{x}$, we may fit nonlinear models of the form

$$\rho(t) = f\left(\sum_j a_j \mathscr{F}_j(\boldsymbol{x})\right) \tag{10.36}$$

efficiently by maximizing the log-likelihood $\log p(D|X,a)$ with respect to the weight parameter a (Wu *et al*., 2006; Ahrens *et al*., 2008). Mathematically, the nonlinearities $\mathscr{F}_j(\boldsymbol{x})$ may take essentially arbitrary form; physiologically speaking, it is clearly wise to choose $\mathscr{F}_j(\boldsymbol{x})$ to reflect known facts about the anatomy and physiology of the system (e.g., $\mathscr{F}_j(\boldsymbol{x})$ might model inputs from a presynaptic layer whose responses are better-characterized than are those of the neuron of interest (Rust *et al*., 2006)).

Second, in many cases it is reasonable to include terms in X that we may not be able to observe or calculate directly (e.g., intracellular noise, or the dynamical state of the network); fitting the model parameters in this case requires that we properly account for these "latent," unobserved variables in our likelihood. While inference in the presence of these hidden parameters is beyond the scope of this chapter, it is worth noting that this type of model may fit tractably using generalizations of the methods described here, at the cost of increased computational complexity, but the benefit of enhanced model flexibility and realism (Smith and Brown, 2003; Yu *et al*., 2009; Vidne *et al*., 2012).

Example: Estimating spike triggered currents and dynamic threshold

In Fig. 10.1a, we have suggested estimating the filter $\kappa(s)$ by extracting its values $k_j = \kappa(j\mathrm{d}t)$ at discrete equally spaced time steps: the integral $\int_0^\infty \kappa(s) I(t-s)\mathrm{d}s = \sum_{j=1}^K k_l I_{t-j}\mathrm{d}t$ is linear in the K parameters.

However, the observables remain *linear in the parameters* (k_j) if we set $\kappa(s) = \sum_{j=1}^4 k_j \exp(-s/\tau_j)$ with fixed time constants, e.g., $\tau_j = 10^j$ ms. Again, the integral $\int_0^\infty \kappa(s) I(t-s)\mathrm{d}s = \sum_{j=1}^4 k_j \int_0^\infty \exp(-s/\tau_j) I(t-s)\mathrm{d}s$ is linear in the parameters. The exponentials play the role of "basis functions" F_j.

Similarly, the threshold filter $\vartheta_1(s)$ or the spike-afterpotential $\eta(s)$ can be expressed with basis functions. A common choice is to take rectangular basis functions F_j which take a value of unity on a finite interval $[t_j, t_{j+1}]$. Exponential spacing $t_j = 2^{j-1}$ms of time points allows us to cover a large time span with a small number of parameters. Regular spacing leads back to the naive discretization scheme.

10.3 Evaluating goodness-of-fit

No single method provides a complete assessment of goodness-of-fit; rather, model fitting should always be seen as a loop, in which we start by fitting a model, then examine the

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

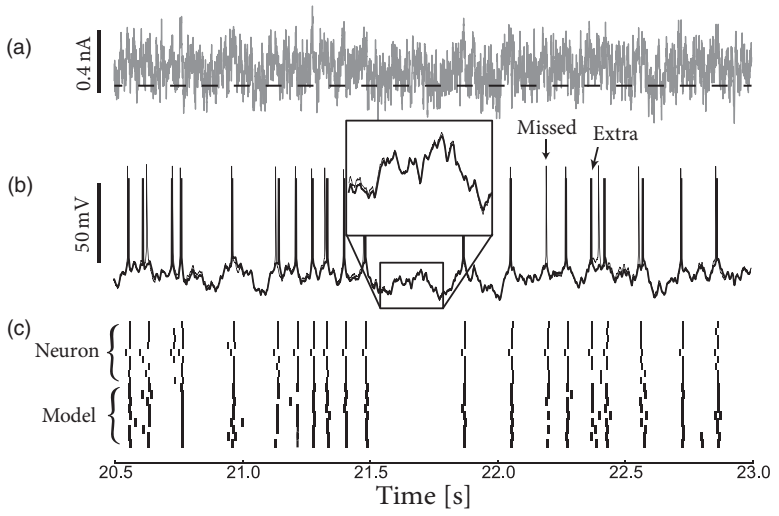

Fig. 10.5 Comparing models with intracellular recordings. (a) A noisy time-dependent current is used to stimulate the neurons experimentally (dashed line corresponds to zero current). (b) Recording from the neuron (thin black line) shows membrane potential fluctuations and action potentials. Simulating an SRM (thick black line) with the same current and using parameters previously optimized on a different dataset shows similar membrane potential fluctuations (inset) and action potentials. Some of the spikes are missed, some are added, but most coincide with the recorded ones. (c) Multiple repeated stimulations with the same current shows the intrinsic variability of neural responses (the first nine rows are recorded action potentials indicated by thick black ticks). The variability is matched by the model (the last nine rows are model action potentials). Data and models from Mensi *et al*. (2012).

results, attempt to diagnose any model failures, and then improve our model accordingly. In the following, we describe different methods for assessing the goodness-of-fit.

Before beginning with specific examples of these methods, we note that it is very important to evaluate the goodness-of-fit on data that was not used for fitting. The part of the data used for fitting model parameters is called the *training set* and the part of the data reserved to evaluate the goodness-of-fit is called the *test set*. Data in the test set is said to be *predicted* by the model, while it is simply *reproduced* in the training set. By simply adding parameters to the model, the quality of the fit on the training set increases. Given a sufficient number of parameters, the model might be able to reproduce the training set perfectly, but that does not mean that data in the test set is well predicted. In fact it is usually the opposite: overly complicated models that are "overfit" on the training data (i.e., which fit not only the reproducible signal in the training set but also the noise) will often do a bad job generalizing and predicting new data in the test set. Thus in the following we assume that the goodness-of-fit quantities are computed using "cross-validation": parameters are estimated using the training set, and then the goodness-of-fit quantification is performed on the test set.

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

10.3.1 Comparing spiking membrane potential recordings

Given a spiking membrane potential recording, we can use traditional measures such as the squared error between model and recorded voltage to evaluate the goodness-of-fit. This approach, however, implicitly assumes that the remaining error has a Gaussian distribution (recall the close relationship between Gaussian noise and the squared error, discussed above). Under diffusive noise, we have seen (Chapter 8) that membrane potential distributions are Gaussian only when all trajectories started at the same point and none have reached threshold. Also, a small jitter in the firing time of the action potential implies a large error in the membrane potential, much larger than the typical subthreshold membrane potential variations. For these two reasons, the goodness-of-fit in terms of subthreshold membrane potential away from spikes is considered separately from the goodness-of-fit in terms of the spike times only.

To evaluate how the model predicts subthreshold membrane potential we must compare the average error with the intrinsic variability. To estimate the first of these two quantities, we compute the squared error between the recorded membrane potential $u_t^{\exp}$ and model membrane potential u_t^{mod} with forced spikes at the times of the observed ones. Since spike times in the model are made synchronous with the experimental recordings, all voltage traces start at the same point. A Gaussian assumption thus justified, we can average the squared error over all recorded times t that are not too close to an action potential:

$$\mathrm{RMSE}_{\mathrm{nm}} = \sqrt{\frac{1}{T_{\Omega_1} N_{\mathrm{rep}}} \sum_{i=1}^{N_{\mathrm{rep}}} \int_{\Omega_1} \left(u_i^{\exp}(t) - u_i^{\mathrm{mod}}(t)\right)^2 \mathrm{d}t}, \tag{10.37}$$

where Ω_1 refers to the ensemble of time bins at least 5 ms before or after any spikes and T_{Ω_1} is the total number of time bins in Ω_1. $\mathrm{RMSE}_{\mathrm{nm}}$ has index n for "neuron" and index m for "model." It estimates the error between the real neuron and the model.

To evaluate the second quantity, we compare recorded membrane potential from multiple repeated stimulations having the same stimulus. Despite the variability in spike timings, it is usually possible to find times which are sufficiently away from a spike in any repetition and compute the averaged squared error

$$\mathrm{RMSE}_{\mathrm{nn}} = \sqrt{\frac{2}{T_{\Omega_2} N_{\mathrm{rep}}(N_{\mathrm{rep}}-1)} \sum_{i=1}^{N_{\mathrm{rep}}} \sum_{j=1}^{i-1} \int_{\Omega_2} \left(u_j^{\exp}(t) - u_i^{\exp}(t)\right)^2 \mathrm{d}t}, \tag{10.38}$$

where Ω_2 refers to the ensemble of time bins far from the spike times in any repetition and T_{Ω_2} is the total number of time bins in Ω_2. Typically, 20 ms before and 200 ms after the spike is considered sufficiently far. Note that with this approach we implicitly assume that spike-afterpotentials have vanished 200 ms after a spike. However, as we shall see in Chapter 11, the spike-afterpotentials can extend for more than one second, so that the above assumption is a rather bad approximation. Because the earlier spiking history will affect the membrane potential, the $\mathrm{RMSE}_{\mathrm{nn}}$ calculated in Eq. (10.38) is an overestimate.

To quantify the predictive power of the model, we finally compute the model error with

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

the intrinsic error by taking the ratio

$$\text{RMSER} = \frac{\text{RMSE}_{\text{nn}}}{\text{RMSE}_{\text{nm}}}. \tag{10.39}$$

The root-mean-squared-error ratio (RMSER) reaches 1 if the model precision is matched with intrinsic error. When smaller than 1, the RMSER indicates that the model could be improved. Values larger than 1 are possible because RMSE_{nn} is an overestimate of the true intrinsic error.

10.3.2 Spike train likelihood

The likelihood is the probability of generating the observed set of spike times $S(t)$ with the current set of parameters in our stochastic neuron model. It was defined in Eq. (9.10), which we reproduce here

$$L^n(S|\theta) = \prod_{t^{(i)} \in S} \rho(t^{(i)}|S,\theta) \exp\left[-\int_0^T \rho(s|S,\theta)\mathrm{d}s\right], \tag{10.40}$$

where we use $\rho(t^{(i)}|S,\theta)$ to emphasize that the firing intensity of a spike at $t^{(i)}$ depends on both the stimulus and spike history as well as the model parameters θ.

The likelihood L^n is a conditional probability density and has units of inverse time to the power of n (where n is the number of observed spikes). To arrive at a more interpretable measure, it is common to compare L^n with the likelihood of a homogeneous Poisson model with a constant firing intensity $\rho_0 = n/T$, i.e., a Poisson process which is expected to generate the same number of spikes in the observation interval T. The difference in log-likelihood between the Poisson model and the neuron model is finally divided by the total number n of observed spikes in order to obtain a quantity with units of "bits per spike":

$$\frac{1}{n}\log_2 \frac{L^n(S|\theta)}{\rho_0^n e^{-\rho T}}. \tag{10.41}$$

This quantity can be interpreted as an instantaneous mutual information between the spike count in a single time bin and the stimulus given the parameters. Hence, it is interpreted as a gain in predictability produced by the set of model parameters θ. One advantage of using the log-likelihood of the conditional firing intensity is that it does not require multiple stimulus repetitions. It is especially useful to compare on a given dataset the performances of different models: better models achieve higher cross-validated likelihood scores.

10.3.3 Time-rescaling theorem

For a spike train with spikes at $t^1 < t^2 < \ldots < t^n$ and with firing intensity $\rho(t|S,\theta)$, the time-rescaling transformation $t \to \Lambda(t)$ is defined as

$$\Lambda(t) = \int_0^t \rho(x|S,\theta)\mathrm{d}x. \tag{10.42}$$

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

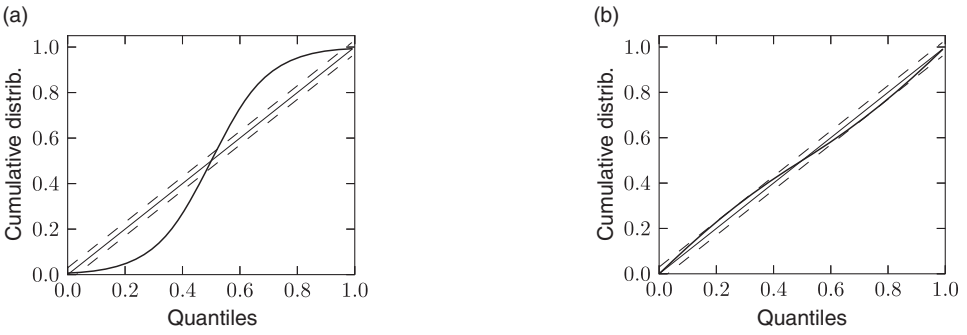

Fig. 10.6 Time-rescaling theorem as a goodness-of-fit. Illustrating the K–S test by plotting the cumulative probability of z_k as a function of quantiles. (a) Rescaling time using an inadequate model does not result in a uniform distribution of z_k as can be seen by comparing the empirical distribution (thick black line) with the diagonal. Dashed lines illustrate 95% confidence bounds. (b) As in (a) but with a better rescaling of time. The empirical distribution follows the cumulative of the uniform distribution within the confidence bounds.

It is a useful and somewhat surprising result that $\Lambda(t^k)$ (evaluated at the measured firing times) is a Poisson process with unit rate (Brown *et al.*, 2002). A correlate of this time-rescaling theorem is that the time intervals

$$\Lambda(t^k) - \Lambda(t^{k-1}) \tag{10.43}$$

are independent random variables with an exponential distribution (see Chapter 7). Re-scaling again the time axis with the transformation

$$z_k = 1 - \exp\left[-\left(\Lambda(t^k) - \Lambda(t^{k-1})\right)\right] \tag{10.44}$$

forms independent uniform random variables on the interval zero to 1.

Therefore, if the model $\rho(t|S,\theta)$ is a valid description of the spike train $S(t)$, then the resulting z_k should have the statistics of a sequence of independent uniformly distributed random variables. As a first step, one can verify that the z_ks are independent by looking at the serial correlation of the interspike intervals or by using a scatter plot of z_{k+1} against z_k. Testing whether the z_ks are uniformly distributed can be done with a Kolmogorov–Smirnov (K–S) test. The K–S statistic evaluates the distance between the empirical cumulative distribution function of z_k, $P(z)$, and the cumulative distribution of the reference function. In our case, the reference function is the uniform distribution, so that its cumulative is simply z. Thus,

$$D = \sup_z |P(z) - z| \,. \tag{10.45}$$

The K–S statistic converges to zero as the empirical probability $P(z)$ converges to the reference. The K–S test then compares D with the critical values of the Kolmogorov distribution. Figure 10.6 illustrates two examples: one where the empirical distribution was

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

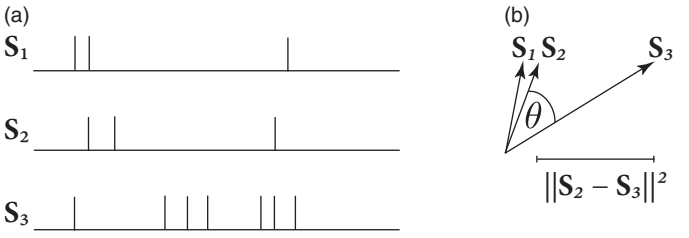

Fig. 10.7 Distance and angular separation between spike trains seen as vectors. (a) Three schematic spike trains where the first two have the same number of spikes and roughly the same timing. (b) The spike trains in (a) can be represented as vectors where $\mathbf{S}_1$ and $\mathbf{S}_2$ have the same length but a slightly different orientation due to small differences in spike timing. The third spike train $\mathbf{S}_3$ is much longer due to the larger number of spikes. It is at a squared distance $D_{23} = ||\mathbf{S}_2 - \mathbf{S}_3||^2$ from $\mathbf{S}_2$ and at angle θ.

far from a uniform distribution and the other where the model rescaled time correctly. See (Gerhard *et al*., 2011) for a multivariate version of this idea that is applicable to the case of coupled neurons. To summarize, the time-rescaling theorem along with the K–S test provide a useful goodness-of-fit measure for spike train data with confidence intervals that does not require multiple repetitions.

10.3.4 Spike-train metric

Evaluating the goodness-of-fit in terms of log-likelihood or the time-rescaling theorem requires that we know the conditional firing intensity $\rho(t|S,\theta)$ accurately. For biophysical models as seen in Chapter 2 but complemented with a source of variability, the firing intensity is unavailable analytically. The intensity can be estimated numerically by simulating the model with different realizations of noise, or by solving a Fokker–Planck equation, but this is sometimes impractical.

Another approach for comparing spike trains involves defining a metric between spike trains. Multiple spike timing metrics have been proposed, with different interpretations. A popular metric was proposed by Victor and Purpura (1996). Here, we describe an alternative framework for the comparison of spike trains that makes use of vector space ideas, rather than more general metric spaces.

Let us consider spike trains as vectors in an abstract vector space, with these vectors denoted with boldface: $\mathbf{S}$. A vector space is said to have an inner (or scalar) product if for each vector pair $\mathbf{S}_i$ and $\mathbf{S}_j$ there exists a unique real number $(\mathbf{S}_i, \mathbf{S}_j)$ satisfying the following axioms: commutativity, distributivity, associativity, and positivity. There are multiple candidate inner products satisfying the above axioms. The choice of inner product will be related to the type of metric being considered. For now, consider the general form

$$(\mathbf{S}_i, \mathbf{S}_j) = \int_0^T \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} K_\Delta(s,s') S_i(t-s) S_j(t-s') \mathrm{d}s \mathrm{d}s' \mathrm{d}t, \tag{10.46}$$

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

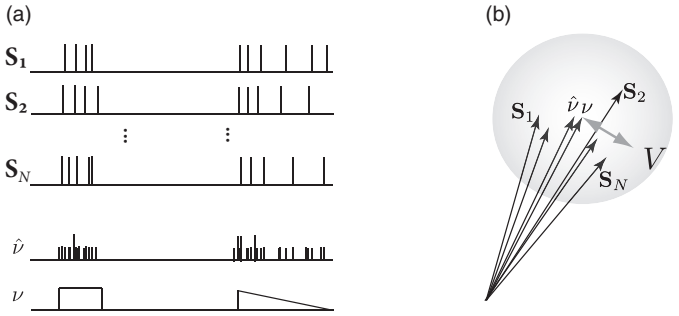

Fig. 10.8 The spike density vector. (a) A set of N spike trains ($\mathbf{S}_1, \mathbf{S}_2, ..., \mathbf{S}_N$) is combined to yield an estimate of the spike density $\hat{\nu}$. At the limit $N \to \infty$ the spike density converges to the instantaneous firing rate ν. (b) Schematic representation of the quantities in (a). The variability V measures the scatter of the individual spike trains around their mean $\hat{\nu}_X$.

where K_Δ is a two-dimensional coincidence kernel with a scaling parameter Δ, and T is the maximum length of the spike trains. Here K_Δ is required to be a non-negative function with a global maximum at $s = s' = 0$. Moreover, $K_\Delta(s,s')$ should fall off rapidly so that $K_\Delta(s,s') \approx 0$ for all $s,s' > \Delta$. Examples of kernels include $K_\Delta(s,s') = k_1(s)k_2(s')$. For instance, $k_1(s) = k_2(s) = \frac{1}{\Delta}e^{-s/\Delta}\Theta(s)$ is a kernel that was used by van Rossum (2001). The scaling parameter Δ must be small, much smaller than the length T of the spike train.

For a comparison of spike trains seen as vectors the notions of angular separation, distance, and norm of spike trains are particularly important. The squared norm of a spike train will be written $||\mathbf{S}_i||^2 = (\mathbf{S}_i, \mathbf{S}_i)$. With $K_\Delta(s,s') = \delta(s)\delta(s')$, we observe that $(\mathbf{S}_i, \mathbf{S}_i) = \int_0^T S_i(t)\mathrm{d}t = n_i$ where n_i is the number of spikes in $\mathbf{S}_i$. Therefore the norm of a spike train is related to the total number of spikes it contains. The Victor and Purpura metric is of a different form than the form discussed here, but it has similar properties (see exercies).

The norm readily defines a distance, D_{ij}, between two spike trains

$$D_{ij}^2 = ||\mathbf{S}_i - \mathbf{S}_j||^2 = (\mathbf{S}_i - \mathbf{S}_j, \mathbf{S}_i - \mathbf{S}_j) = ||\mathbf{S}_i||^2 + ||\mathbf{S}_j||^2 - 2(\mathbf{S}_i, \mathbf{S}_j). \tag{10.47}$$

The right-hand side of Eq. (10.47) shows that that the distance between two spike trains is maximum when $(\mathbf{S}_i, \mathbf{S}_j)$ is zero. On the other hand, D_{ij}^2 becomes zero only when $S_i = S_j$. This implies that $(\mathbf{S}_i, \mathbf{S}_j) = (\mathbf{S}_i, \mathbf{S}_i) = (\mathbf{S}_j, \mathbf{S}_j)$. Again consider $K_\Delta(s,s') = \delta(s)\delta(s')$, then D_{ij} is the total number of spikes in both $\mathbf{S}_i$ and $\mathbf{S}_j$ reduced by 2 for each spike in $\mathbf{S}_i$ that coincided with one in $\mathbf{S}_j$. For the following, it is useful to think of a distance between spike trains as a number of non-coincident spikes.

The cosine of the angle between $\mathbf{S}_i$ and $\mathbf{S}_j$ is

$$\cos\theta_{ij} = \frac{(\mathbf{S}_j, \mathbf{S}_i)}{||\mathbf{S}_i||\,||\mathbf{S}_j||}. \tag{10.48}$$

This angular separation relates to the fraction of coincident spikes. Fig. 10.7 illustrates the concepts of angle and distance for spike trains seen as vectors.

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

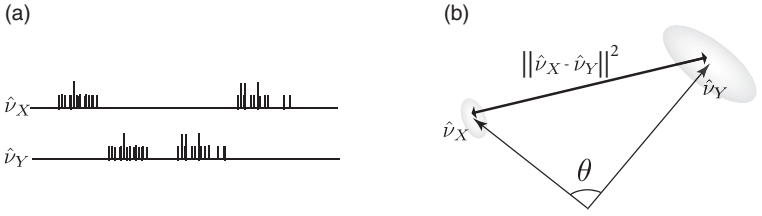

Fig. 10.9 Distance and angular separation between spike densities. (a) Two spike densities corresponding to a sum of spike trains labeled X and Y. (b) Schematic representation of the densities $\hat{\nu}_X$ and $\hat{\nu}_Y$ seen as vectors. Both are separated by a distance $||\hat{\nu}_X - \hat{\nu}_Y||^2$ and angle θ. The variability is shown as a gray cloud centered on the vectors.

10.3.5 Comparing sets of spike trains

Metrics such as D_{ij} described above can quantify the similarity between two spike trains. In the presence of variability, however, a simple comparison of spike trains is not sufficient. Instead, the spike train similarity measure must be maximally sensitive to differences in the underlying stochastic processes.

We want to know if spike trains generated from a neuron model could well have been generated by a real neuron. We could simply calculate the distance between a spike train from the neuron and a spike train from the model, but neurons are noisy and we will find a different distance each time we repeat the recording. To achieve better statistics, we can compare a set of spike trains from the model with a set of spike trains from the neuron.

Let the two sets of spike trains be denoted by X and Y, containing N_X and N_Y spike trains, respectively. First, it is useful to define some characteristics of such sets of spike trains. A natural quantity to consider is the average of the norms of each spike train within a set, say X,

$$\hat{L}_X = \frac{1}{N_X} \sum_{i=1}^{N_X} ||\mathbf{S}_i^{(x)}||^2, \tag{10.49}$$

where we have used $\hat{}$ to denote that the quantity is an experimental estimate. We note that $\hat{L}_X$ is related to the averaged spike count. L_X is exactly the averaged spike count if the inner product satisfies (i) $\int\int K_\Delta(s,s')\mathrm{d}s\mathrm{d}s' = 1$ and (ii) $K_\Delta(s,s') = 0$ whenever $|s-s'|$ is greater than the minimum interspike interval of any of the spike trains considered. The interpretation $L_X \sim$ *spike count* is helpful for the discussion in the remainder of this section. Furthermore, the vector of averaged spike trains,

$$\hat{\nu}_X = \frac{1}{N_X} \sum_{i=1}^{N_X} \mathbf{S}_i^{(x)}, \tag{10.50}$$

is another occurrence of the spike density seen in Chapter 7. It defines the instantaneous firing rate of the the spiking process, $\nu(t) = \langle \hat{\nu} \rangle$. In the vector space, $\hat{\nu}_X$ can be thought of as lying at the center of the spike trains seen as vectors (Fig. 10.9); note that other "mean

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

spike trains" could be defined (Wu and Srivastava, 2012). The size of the cloud quantifies the variability in the spike timing. The variability is defined as the variance

$$\hat{V}_X = \frac{1}{N_X - 1} \sum_{i=1}^{N_X} ||\mathbf{S}_i^{(x)} - \hat{\nu}_X||^2. \tag{10.51}$$

A set of spike trains where spikes always occur at the same times has low variability. When the spikes occur with some jitter around a given time, the variability is larger. Variability relates to reliability. While variability is a positive quantity that cannot exceed L_X, reliability is usually defined between zero and 1, where 1 means perfectly reliable spike timing: $\hat{R}_X = 1 - \hat{V}_X / \hat{L}_X$.

Finally, we come to a measure of match between the set of spike trains X and Y. The discussion in Chapter 7 about the neural code would suggest that neuron models should reproduce the detailed time structure of the PSTH. We therefore define

$$\hat{M} = \frac{2\,(\hat{\nu}_X, \hat{\nu}_Y)}{\hat{R}_X \hat{L}_X + \hat{R}_Y \hat{L}_Y}. \tag{10.52}$$

We have M (for match) equal to 1 if X and Y have the same instantaneous firing rate. The smaller M is, the greater the mismatch between the spiking processes. The quantity $R_X L_X$ can be interpreted as a number of reliable spikes. Since $(\hat{\nu}_X, \hat{\nu}_Y)$ is interpreted as a number of coincident spikes between X and Y, we can still regard M as a factor counting the fraction of coincident spikes. A similar quantity can be defined for metrics that cannot be cast into an inner product (Naud *et al.*, 2011).

If the kernel $K_\Delta(s, s')$ is chosen to be $k_g(s)k_g(s')$ and k_g is a Gaussian distribution of width Δ, then M relates to a mean square error between PSTHs that were filtered with k_g. Therefore, the kernel used in the definition of the inner product (Eq. (10.46)) can be related to the smoothing filter of the PSTH (see Exercises).

10.4 Closed-loop stimulus design

In the previous sections we have developed robust and tractable approaches to understand neural encoding, based on GLMs, and quantifying the performance of models. The framework we have developed is ultimately data-driven; both our encoding and decoding methods fail if the observed data do not sufficiently constrain our encoding model parameters θ. Therefore we will close by describing how to take advantage of the properties of the GLM to optimize our experiments: the objective is to select, in an online, closed-loop manner, the stimuli that will most efficiently characterize the neuron's response properties.

An important property of GLMs is that not all stimuli will provide the same amount of information about the unknown coefficients $\boldsymbol{k}$. As a concrete example, we can typically learn much more about a visual neuron's response properties if we place stimulus energy within the receptive field, rather than "wasting" stimulus energy outside the receptive field. To make this idea more rigorous and generally applicable, we need a well-defined objective function that will rank any given stimulus according to its potential informativeness.

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

Numerous objective functions have been proposed for quantifying the utility of different stimuli (Mackay, 1992; Nelken *et al*., 1994; Machens, 2002). When the goal is estimating the unknown parameters of a model, it makes sense to choose stimuli $\boldsymbol{x}_t$ which will on average reduce the uncertainty in the parameters θ as quickly as possible (as in the game of 20 questions), given $D = \{\boldsymbol{x}(s), n_s\}_{s<t}$, the observed data up to the current trial. This posterior uncertainty in θ can be quantified using the information-theoretic notion of "entropy"; see Cover and Thomas (1991), Mackay (1992), Paninski (2005) for further details.

In general, information-theoretic quantities such as the entropy can be difficult to compute and optimize in high-dimensional spaces. However, Lewi *et al*. (2009) show that the special structure of the GLM can be exploited (along with a Gaussian approximation to $p(\theta|D)$) to obtain a surprisingly efficient procedure for choosing stimuli optimally in many cases. Indeed, a closed-loop optimization procedure leads to much more efficient experiments than does the standard open-loop approach of stimulating the cell with randomly chosen stimuli that are not optimized adaptively for the neuron under study.

A common argument against online stimulus optimization is that neurons are highly adaptive: a stimulus which might be optimal for a given neuron in a quiescent state may quickly become suboptimal due to adaptation (in the form of short- and long-term synaptic plasticity, slow network dynamics, etc.). Including spike-history terms in the GLM allows us to incorporate some forms of adaptation (particularly those due to intrinsic processes including, for example, sodium channel inactivation and calcium-activated potassium channels), and these spike-history effects may be easily incorporated into the derivation of the optimal stimulus (Lewi *et al*., 2009). However, extending these results to models with more profound sources of adaptation is an important open research direction; see Lewi *et al*. (2009) and DiMattina and Zhang (2011) for further discussion.

10.5 Summary

With modern statistical methods, we have fast and computationally tractable schemes to fit models of neural encoding and decoding to experimental data. A key insight is that, for a suitable chosen model class, the likelihood of the data being generated by the model is a concave function of the model parameters, i.e., there are no local maxima. Because of this, numerical methods of gradient ascent are bound to lead to the global maximum.

Generalized Linear Models (GLMs) are the representative of this model class. Importantly, a large ensemble of generalized integrate-and-fire models, in particular the SRM with escape noise, belong to the family of GLMs. As we have seen in previous chapters, the SRM can account for a large body of electrophysiological data and firing patterns such as adaptation, burst firing, time-dependent firing threshold, hyperpolarizing spike-afterpotential, etc. The link from SRM to GLM implies that there are systematic and computationally fast methods to fit biologically plausible neuron models to data.

Interestingly, once neuron models are phrased in the language of statistics, the problems of coding and stimulus design can be formulated in a single unified framework. In the

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

following chapter we shall see that the problem of decoding can also be analyzed in the same statistical framework.

Literature

An early application of maximum likelihood approaches to neuronal data can be found in Brillinger (1988). The application of the framework of Generalized Linear Models to the field of neuroscience has been made popular by Truccolo *et al*. (2005) and Pillow *et al*. (2008). A review of Generalized Linear Models can be found in Dobson and Barnett (2008).

The influential book Rieke *et al*. (1997) gives a broad introduction to the field of neural coding. The time-rescaling theorem was exploited by Brown *et al*. (2002) to develop useful goodness-of-fit methods for spike trains. Spike-train metrics were introduced in Victor and Purpura (1996, 1997), but comparisons of spike trains in terms of PSTHs and other features has been commonly used before (Perkel *et al*., 1967a,b; Gerstein and Perkel, 1972; MacPherson and Aldridge, 1979; Eggermont *et al*., 1983; Gawne *et al*., 1991). Many other spike-train distances were also proposed (Kistler *et al*., 1997; van Rossum, 2001; Quiroga *et al*., 2002; Hunter and Milton, 2003; Schreiber *et al*., 2003; Naud *et al*., 2011) which can be cast in the general framework of a vector space as outlined in Schrauwen and Campenhout (2007), Paiva *et al*. (2009a) and Naud *et al*. (2011); see also Paiva *et al*. (2009b, 2010) and Park *et al*. (2012). Nonlinear functions of the spike trains can also be used to relate to different features of the spiking process such as the interval distribution or the presence of definite firing patterns (Victor and Purpura, 1996; Quiroga *et al*., 2002; Tiesinga, 2004; Kreuz *et al*., 2007, 2009; Druckmann *et al*., 2007).

Exercises

1. **Concave function and non-global optima**
   (a) Suppose a function $G(x)$ has a global maximum at location x_0. Suppose that $f(y)$ is a strictly increasing function of y (i.e., $\mathrm{d}f/\mathrm{d}y > 0$).
   Show that $f(G(x))$ has a maximum at x_0. Is it possible that $f(G(x))$ has further maxima as a function of x?
   (b) A strictly concave function G can be defined as a curve with negative curvature $\mathrm{d}^2G/\mathrm{d}x^2 < 0$ for all x. Show that a concave function can have at most one maximum.
   (c) Give an example of a concave function which does not have a maximum. Give an example of a function G which has a global maximum, but is not concave. Give an example of a function G which is concave and has a global maximum.
2. **Sum of concave functions.** *Consider a quadratic function $f_k(x) = 1 - (x - \vartheta_k)^2$.*
   (a) Show that f_k is a concave function of x for any choice of parameter ϑ_k.
   (b) Show that $f_1(x) + f_2(x)$ is a concave function.
   (c) Show that $\sum_k b_k f_k(x)$ with $b_k > 0$ is a concave function.
   (d) Repeat the steps (b) and (c) for a family of functions f_k which are concave, but not necessarily quadratic.
3. **Comparing PSTHs and spike train similarity measures.** *Experimentally the PSTH is constructed from a set of N_{rep} spike trains, $S_i(t)$, measured from repeated presentations of the same*

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press

stimulus. The ensemble average of the recorded spike trains:

$$\frac{1}{N_{\text{rep}}}\sum_{i=1}^{N_{\text{rep}}} S_i(t) \tag{10.53}$$

is typically convolved with a Gaussian function $h_g(x) = (2\pi\sigma^2)^{-1/2}\exp\left(-x^2/2\sigma^2\right)$ *with* σ *around 5 ms, such that* $A_1(t) = (h_g * \frac{1}{N_{\text{rep}}}\sum S_i)(t)$ *is a smoothed PSTH. Suppose that two sets of experimental spike trains were recorded in two different conditions, resulting in two smoothed PSTHs* $A_1(t)$ *and* $A_2(t)$.

(a) Show that the sum of the squared error $(A_1(t) - A_2(t))^2$ *can be written as a distance between sets of spike train* D^2 *with the kernel* $K(t,t') = h_g(t)h_g(t')$.

(b) Recall that the correlation coefficient between datasets x and y is

$$c = \text{cov}(x,y)/\sqrt{\text{cov}(x,x)\text{cov}(y,y)}. \tag{10.54}$$

Show that the correlation coefficient between the two smoothed PSTHs can be written as a angular separation between the sets of spike trains with kernel $K(t,t') = h_g(t)h_g(t')$.

4. **Victor and Purpura metric.** *Consider the minimum cost C required to transform a spike train* S_i *into another spike train* S_j *if the only transformations available are:*
- removing a spike has a cost of 1,
- adding a spike has a cost of 1,
- shifting a spike by a distance d has a cost q d where q is a parameter defining temporal precision.
The C defines a metric that measures the dissimilarity between spike train S_i *and spike train* S_j. *The smaller C is the more alike the spike trains are in terms of spike timing.*
(a) For $q = 0$ *units of cost per seconds, show that C becomes the difference in number of spikes in spike trains* S_i *and* S_j.
(b) For q greater than four times the maximum firing frequency (i.e., the inverse of the shortest observed interspike interval), show that C can be written as a distance D^2_{ij} *with kernel* $K(t,t') = h_t(t)\delta(t')$ *and triangular function* $h_t(t) = (1 - |t|q/2)\Theta(1 - |t|q/2)$ *where* $\delta(\cdot)$ *is the Dirac delta function and* $\Theta(\cdot)$ *is the Heaviside function.*

https://doi.org/10.1017/CBO9781107447615.012 Published online by Cambridge University Press