11

Encoding and decoding with stochastic neuron models

In the ten preceding chapters, we have come a long way: starting from the biophysical basis of neuronal dynamics we arrived at a description of neurons that we called generalized integrate-and-fire models. We have seen that neurons contain multiple types of ion channels embedded in a capacitive membrane (Chapter 2). We have seen how basic principles regulate the dynamics of electrical current and membrane potential in synapses, dendrites and axons (Chapter 3). We have seen that sodium and potassium ion channels form an excitable system characterized by a threshold mechanism (Chapter 4) and that other ion channels shape the spike after-effects (Chapter 6). Finally, we have seen in Chapters 4, 5 and 6 how biophysical models can be reduced by successive approximations to other, simpler, models such as the LIF, EIF, AdEx, and SRM. Moreover, we have added noise to our neuron models (Chapters 7 and 9). At this point, it is natural to step back and check whether our assumptions were too stringent, whether the biophysical assumptions were well-founded, and whether the generalized models can explain neuronal data. We laid out the mathematical groundwork in Chapter 10; we can now set out to apply these statistical methods to real data.

We can test the performance of these, and other, models by using them as predictive models of *encoding*. Given a stimulus, will the model be able to predict the neuronal response? Will it be able to predict spike times observed in real neurons when driven by the same stimulus – or only the mean firing rate or PSTH? Will the model be able to account for the variability observed in neuronal data across repetitions?

Testing the performance of models addresses an even bigger question. What information is discarded in the neural code? What features of the stimulus are most important? If we understand the neural code, will we be able to reconstruct the image that the eye is actually seeing at any given moment from spike trains observed in the brain? The problem of *decoding* neuronal activity is central both for our basic understanding of neural information processing (Rieke *et al*., 1997) and for engineering "neural prosthetic" devices that can interact with the brain directly (Donoghue, 2002). Given a spike train observed in the brain, can we read out intentions, thoughts, or movement plans? Can we use the data to control a prosthetic device?

In Section 11.1 we use the generalized integrate-and-fire models of Chapters 6 and 9 to predict membrane voltage and spike timings of real neurons during stimulation with an

https://doi.org/10.1017/CBO9781107447615.013 Published online by Cambridge University Press

arbitrary time-dependent input current in vitro. In Section 11.2, we use the same model class to predict spike timings *in vivo*. Finally, in Section 11.3 we examine the question of decoding: given a measured spike train can we reconstruct the stimulus, or control a prosthetic arm?

11.1 Encoding models for intracellular recordings

We will focus the discussion on generalized integrate-and-fire models with escape noise, also called soft-threshold integrate-and-fire models (Fig. 10.3a). The vast majority of studies achieving good predictions of voltage and spike timing use some variant of this model. The reasons lie in the model's ease of optimization and in its flexibility; see Chapter 6. Also, the possibility of casting them into the GLM formalism allows efficient parameter optimization; see Chapter 10. In Section 11.1.1 we use the SRM as well as soft-threshold integrate-and-fire models to predict the subthreshold voltage of neurons in slices driven by a time-dependent external current. We then use these models to also predict the spike timings of the same neurons (Section 11.1.2).

11.1.1 Predicting membrane potential

The SRM describes somatic membrane potential in the presence of an external current (Section 6.4)

$$u(t) = \sum_f \eta(t - t^f) + \int_0^\infty \kappa(s)\, I^{\text{ext}}(t-s)\, \mathrm{d}s + u_{\text{rest}}. \tag{11.1}$$

The parameters of this model define the functional shape of the functions $\eta(t)$ and $\kappa(t)$. Other parameters such as threshold or, in a stochastic model, the sharpness of threshold β do not contribute to the mean squared error of the membrane potential as defined in Section 10.3.1. Following the methods of Chapter 10, we can estimate the functions $\kappa(t)$ and $\eta(t)$ from recordings of cortical neurons. We note that the spike-afterpotential has units of voltage, whereas the membrane filter κ has units of resistance over time.

Using *in vitro* intracellular recordings of cells in layer 2–3 of the somatosensory cortex, Mensi *et al*. (2012) optimized the functions $\kappa(t)$ and $\eta(t)$ on the recorded potential. For both the main type of excitatory neurons and the main type of inhibitory neurons, the membrane filter $\kappa(t)$ is well described by a single exponential (Fig. 11.1a). Different cell types have different amplitudes and time constants. The inhibitory neurons are typically faster, with a smaller time constant than the excitatory neurons, suggesting we could discriminate between excitatory and inhibitory neurons in terms of the shape of $\kappa(t)$. Discrimination of cell types, however, is much improved when we take into account the spike-afterpotential. The shape of $\eta(t)$ in inhibitory cells is very different than that in excitatory ones (Fig. 11.1b).

While the spike-afterpotential is a monotonically decreasing function in the excitatory cells, in the inhibitory cells the function $\eta(t)$ is better fitted by two exponentials of opposite

https://doi.org/10.1017/CBO9781107447615.013 Published online by Cambridge University Press

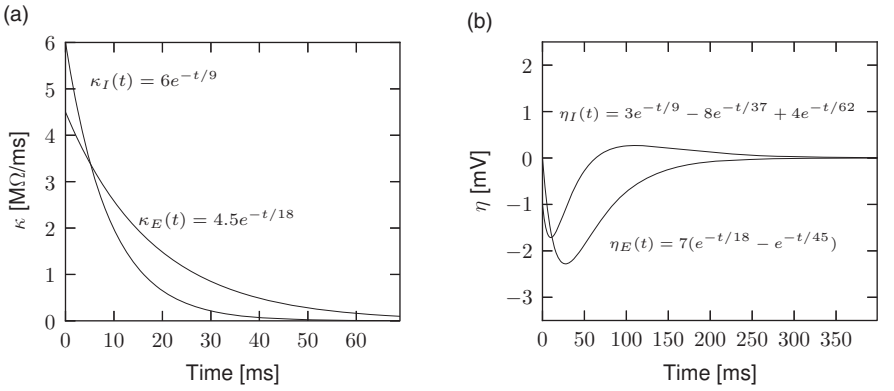

Fig. 11.1 Parameters of voltage recordings in the main excitatory and inhibitory neuron type of cortical layer 2–3. (a) Membrane filter for the main excitatory cell type $\kappa_E(t)$ and the fast-spiking inhibitory cell type $\kappa_I(t)$. (b) Spike-afterpotential for the main excitatory cell type $\eta_E(t)$ and the main inhibitory cell type $\eta_I(t)$. Equations have units of ms for time, mV for η and MΩ/ms for κ. Modified from Mensi *et al*. (2012).

polarity. This illustrates that different cell types differ in the functional shape of the spike-afterpotential $\eta(t)$. This finding is consistent with the predictions of Chapter 6 where we discussed the role of different ion channels in shaping $\eta(t)$. Similar to Fig. 6.14, the spike-afterpotential of inhibitory neurons is depolarizing 30–150 ms after the spike time, another property providing fast-spiking dynamics. Therefore, we conclude that the spike-afterpotential of inhibitory neurons has an *oscillatory* component.

Once the parameters have been extracted from a first set of data, how well does the neuron model predict membrane potential recordings? Qualitatively, we have already seen in Chapter 10 (Fig. 10.5) an example of a typical prediction. Quantitatively, the membrane potential fluctuations of the inhibitory and excitatory neuron have a RMSER (Eq. (10.39)) below one, meaning that the prediction error is smaller than our estimate of the intrinsic error. This indicates that our estimate of the intrinsic error is slightly too large, probably because the actual spike-afterpotential is even longer than a few hundred milliseconds – as we shall see below.

Subthreshold mechanisms that can lead to a resonance (Chapter 6) would cause $\kappa(t)$ to oscillate in time. Mensi *et al.* (2012) have tested for the presence of a resonance in $\kappa(t)$ and found none. Using two exponentials to model $\kappa(t)$ does not improve the prediction of subthreshold membrane potential. Thus, the membrane potential filter is well described by a *single exponential* with time constant $\tau_m = RC$ where R is the passive membrane resistance and C the capacity of the membrane. If we set $\kappa(s) = (1/C)\exp(-s/\tau_m)$, we can take the derivative of Eq. (11.1) and write it in the form of a differential equation

$$C\frac{\mathrm{d}u(t)}{\mathrm{d}t} = -\frac{1}{R}(u - u_{\mathrm{rest}}) + \sum_f \tilde{\eta}(t - t^f) + I^{\mathrm{ext}}(t), \tag{11.2}$$

where $\tilde{\eta}(s)$ is the time course of the net *current* triggered after a spike.

https://doi.org/10.1017/CBO9781107447615.013 Published online by Cambridge University Press

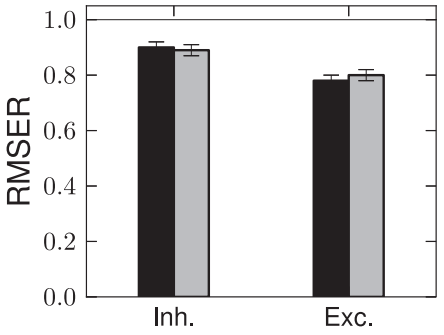

Fig. 11.2 Voltage prediction. Goodness-of-fit of voltage recordings in the main excitatory and inhibitory neuron types of cortical layer 2–3. RMSER (see Chapter 10) for generalized soft-threshold integrate-and-fire models of excitatory and inhibitory neurons (black bars). The gray bars indicate models where the spike-afterpotential is mediated by a spike-triggered change in conductance instead of current. Modified from Mensi *et al.* (2012).

We have seen in Chapters 2 and 6 that spike-triggered adaptation is mediated by ion channels that change the *conductance* of the membrane. Biophysics would therefore suggest a spike-triggered change in conductance, such that after every spike the total current that can charge the membrane capacitance is

$$C\frac{\mathrm{d}u}{\mathrm{d}t} \propto \eta_C(t-\hat{t})(u-E_{\mathrm{rev}}), \tag{11.3}$$

where η_C is the spike-triggered change in conductance and E_{rev} its reversal potential. The reversal potential and the time course η_C can be optimized to yield the best goodness-of-fit. In the excitatory neurons, the resulting conductance change follows qualitatively the current-based $\tilde{\eta}(t)$. The prediction performance, however, is not significantly improved (Fig. 11.2), indicating that describing spike after-effects in terms of current is a good assumption.

11.1.2 Predicting spikes

Using the same intracellular recordings as in Fig. 11.1 (Mensi *et al.*, 2012), we now ask whether spike firing can be predicted from the model. The results of the previous subsection provide us with the voltage trajectory $u(t)$ of the generalized integrate-and-fire model. Assuming a moving threshold that can undergo a stereotypical change at every spike $\vartheta(t) = \vartheta_0 + \sum_f \theta_1(t-t^f)$ we can model the conditional firing intensity as follows, given the spike train, S (compare Eq. (9.27))

$$\rho(t|S) = \frac{1}{\tau_0}\exp\left[\beta\left(u(t)-\vartheta_0-\sum_{t^f\in S}\theta_1(t-t^f)\right)\right]. \tag{11.4}$$

Since the parameters regulating $u(t)$ were optimized using the subthreshold membrane potential in Section 11.1.1, the only free parameters left are those of the threshold, i.e., ϑ_0, β, and the function $\theta_1(t)$. Once the function θ_1 is expanded in a linear combination of basis functions, maximizing the likelihood Eq. (10.40), can be done through a convex gradient descent because Eq. (11.4) can be cast into a GLM.

https://doi.org/10.1017/CBO9781107447615.013 Published online by Cambridge University Press

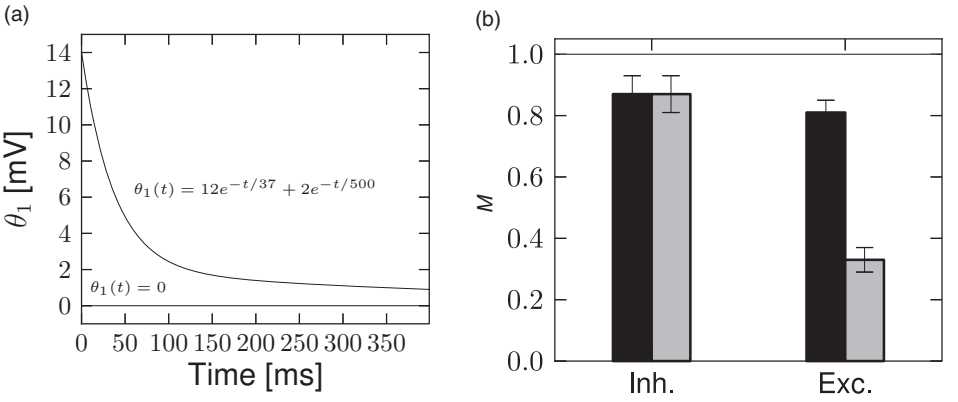

Fig. 11.3 Parameters and spike time prediction in the main excitatory and inhibitory neuron type of cortical layer 2–3. (a) Moving threshold for the main excitatory cell type was found to be an exponentially decaying function (top curve and equation). For the main inhibitory cell type, the fitted moving threshold was not significantly different from zero (bottom curve and equation). Equations have units of ms for time and mV for θ_1. (b) The spike-timing prediction in terms of the similarity measure M (Eq. (10.52)) for models with the moving threshold (black bars) and without the moving threshold (gray bars). Modified from Mensi *et al*. (2012).

Is the dynamic threshold necessary? Optimizing the parameters on a training dataset, we find no need for a moving threshold in the inhibitory neurons (Fig. 11.3a). The threshold in those cells is constant in time. However, the excitatory cells have a strongly moving threshold (Fig. 11.3a) which is characterized by at least two decay time constants. A moving threshold can have several potential biophysical causes. Inactivation of sodium channels is a likely candidate (Fleidervish *et al*., 1996).

How good is the prediction of spike times in inhibitory and excitatory cortical neurons? Qualitatively, the model spike trains resemble the recorded ones with a similar intrinsic variability (Fig. 10.5). Quantitatively, Mensi *et al*. (2012) used the measure of match M (see Eq. (10.52)) and $K(t,t') = \Theta(t+\Delta)\Theta(\Delta - t)\delta(t')$ with $\Delta = 4$ ms. They found $M = 87\%$ for the inhibitory neurons and $M = 81\%$ for the excitatory neurons (Fig. 11.3b). Intuitively, this result means that these models predict more than 80% of the "predictable" spikes.

These numbers are averaged over a set of cells. Some cells were predicted better than others such that the M reached 95% for inhibitory neurons and 87% for excitatory neurons. Similar results are found in excitatory neurons of layer 5. Spikes from these neurons can be predicted with $M = 81\%$ on average (Pozzorini *et al*., 2013). Other optimization methods but with similar models could improve the spike-timing prediction of inhibitory neurons, reaching $M = 100\%$ for some cells (Kobayashi *et al*., 2009). Thus, the case of inhibitory neurons seems well resolved. The almost perfect match between predicted and experimental spike trains leaves little place for model refinement. Unless the stimulus is specifically

https://doi.org/10.1017/CBO9781107447615.013 Published online by Cambridge University Press

designed to probe bursting or postinhibitory rebound, the generalized integrate-and-fire model is a sufficient description of the fast-spiking inhibitory neuron.

One important feature of the model for spike-timing prediction is adaptation. Optimizing a generalized integrate-and-fire model with refractory effects but no adaptation reduces the prediction performance by 20–30% (Jolivet *et al*., 2008b), for both the excitatory and inhibitory cortical cells. How long do the effects of spike-triggered adaptation last? Surprisingly, a single spike has a measurable effect more than 10 seconds after the action potential has occurred (Fig. 11.4). Thus, adaptation is not characterized by a single time scale (Lundstrom *et al*., 2008) and shows up as a power-law decay in both spike-triggered current and threshold (Pozzorini *et al*., 2013).

11.1.3 How good are generalized integrate-and-fire models?

For excitatory neurons, a value of $M = 81\%$ implies that there remains nevertheless 19% unexplained PSTH variance. Using state-of-the-art model optimization for a full biophysical model with ion channels and extended dendritic tree does not improve the model performance (Druckmann *et al*., 2007). Considering a dependence on the voltage derivative in the escape rate (Chapter 9) can slightly improve the performance (Kobayashi and Shinomoto, 2007) but is not sufficient to achieve a flawless prediction. Similarly, taking into account very long spike-history effects (Fig. 11.4) and experimental drifts improves mostly the prediction of time-dependent rate performance on long time scales, and only slightly spike-time prediction at short time scales (Pozzorini *et al*., 2013). Overall, the situation gives the impression that a mechanism might be missing in the generalized integrate-and-fire model and perhaps in the biophysical description as well.

Nevertheless, more than 80% of PSTH variance *is predicted* by generalized soft-threshold integrate-and-fire models during current injection into the soma. This result holds for a time-dependent current which changes on fast as well as slow time scales – a challenging scenario. The effective current driving single neurons in an awake animal *in vivo* might have comparable characteristics in that it comprises slow fluctuations of the mean as well as fast fluctuations (Crochet *et al*., 2011; Pozzorini *et al*., 2013). Similarly, the net driving current in connected model networks (see Part III), typically also shows fluctuations around a mean value that changes on a slower time scale. Taken together, generalized integrate-and-fire models are valid models in the physiological input range observed *in vivo*, and are good candidates for large-scale network simulation and analysis.

Linear dendritic effects show up in the membrane filter and spike-afterpotential; but strongly nonlinear dendrites as observed with multiple recordings from the same neuron (Larkum *et al*., 2001) cannot be accounted for by a generalized soft-threshold integrate-and-fire model or GLM. If nonlinear interactions between different current injection sites along the dendrite are important, a different class of neuron models needs to be considered (Chapter 3).

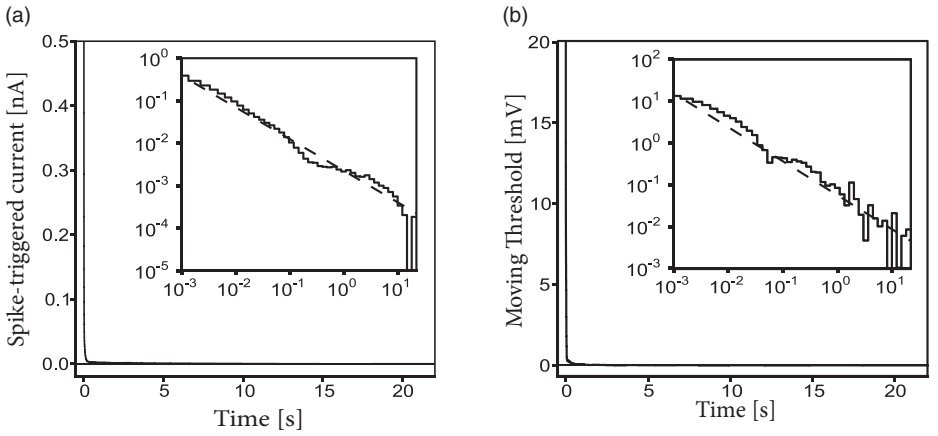

Fig. 11.4 Long spike after-effects in excitatory cortical cells of the layer 5. (a) The spike-triggered current fitted on the membrane potential of layer 5 pyramidal neurons is shown as a function of time since spike emission. Although the effect of a spike appears to be over after a few tens of milliseconds, the log–log scale (inset) reveals that the spike after-current decays with a power law $\tilde{\eta}(t) \propto -t^{-0.8}$ over four orders of magnitude. (b) The moving threshold fitted on the spike timing of layer 5 pyramidal neurons is shown as a function of time since spike emission. As in (a), the log–log scale (inset) reveals a power law $\theta_1(t) \propto t^{-0.8}$ (Pozzorini *et al.*, 2013).

11.2 Encoding models in systems neuroscience

Generalized integrate-and-fire models have been used not only for spike prediction of neurons in brain slices, but also for measurements in systems neuroscience, i.e., in the intact brain driven by sensory stimuli or engaged in a behavioral task. Traditionally, electrophysiological measurements *in vivo* have been performed with extracellular electrodes or multi-electrode probes. With extracellular recording devices, the presence of spikes emitted by one or several neurons can be detected, but the membrane potential of the neuron is unknown. Therefore, in the following we aim at predicting spikes in extracellular recordings from a single neuron, or in groups of connected neurons.

11.2.1 Receptive fields and linear-nonlinear-poisson model

The linear properties of a simple neuron in the primary visual cortex can be identified with its receptive field, i.e., the small region of visual space in which the neuron is responsive to stimuli (see Chapters 1 and 12). Receptive fields as linear filters have been analyzed in a wide variety of experimental preparations.

Experimentally, the receptive field of a simple cell in visual cortex can be determined by presenting a random sequence of spots of lights on a gray screen while the animal is watching the screen (Fig. 11.5a). In a very limited region of the screen, the spot of light leads to an increase in the probability of firing of the cell, in an adjacent small region to a

https://doi.org/10.1017/CBO9781107447615.013 Published online by Cambridge University Press

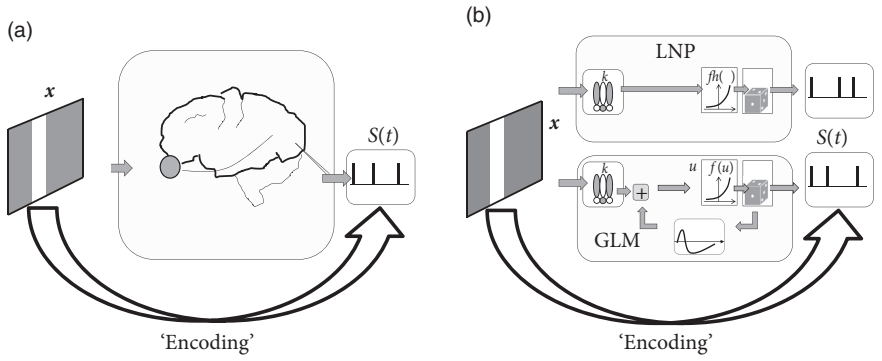

Fig. 11.5 The encoding problem in the visual neuroscience. (a) A stimulus is presented on a screen while a spike train is recorded from an area in the visual cortex. (b) Models designed to predict the spike train first filter the stimulus $\boldsymbol{x}$ with a spatial filter $\boldsymbol{k}$ (linear processing step), pass the result $u = \boldsymbol{k} \cdot \boldsymbol{x}$ through a nonlinearity f and then generate spikes stochastically with Poisson statistics. The main difference between a Linear-Nonlinear-Poisson (LNP, top) and a soft-threshold generalized integrate-and-fire model (GLM, bottom) is the presence of spike-triggered currents $\tilde{\eta}(s)$ in the latter.

decrease. The spatial arrangement of these regions defines the spatial receptive field of the cell and can be visualized as a two-dimensional spatial linear filter (Fig. 11.5b).

Instead of a two-dimensional notation of screen coordinates, we choose in what follows a vector notation where we label all pixels with a single index k. For example, on a screen with 256×256 pixels we have $1 \leq k \leq K$ with $K = 65536$. A full image corresponds to a vector $\boldsymbol{x} = (x_1, \ldots, x_K)$ while a single spot of light corresponds to a vector with all components equal to zero except one (Fig. 11.6a).

The spatial receptive field of a neuron is a vector $\boldsymbol{k}$ of the same dimensionality as $\boldsymbol{x}$. The response of the neuron to an *arbitrary* spatial stimulus $\boldsymbol{x}$ depends on the total drive $\boldsymbol{k} \cdot \boldsymbol{x}_t$, i.e., the similarity between the stimulus and the spatial filter.

More generally, the receptive field filter $\boldsymbol{k}$ can be described not only by a spatial component, but also by a temporal component: an input 100 ms ago has less influence on the spiking probability now than an input 30 ms ago. In other words, the scalar product $\boldsymbol{k} \cdot \boldsymbol{x}_t$ is a short-hand notation for integration over space as well as over time. Such a filter $\boldsymbol{k}$ is called a *spatio-temporal receptive field*.

In the Linear-Nonlinear-Poisson (LNP) model, one assumes that spike trains are produced by an inhomogeneous Poisson process with rate

$$\rho(t) = f(\boldsymbol{k} \cdot \boldsymbol{x}_t), \tag{11.5}$$

given by a cascade of two simple steps (Fig. 11.5b). The linear stage, $\boldsymbol{k} \cdot \boldsymbol{x}_t$, is a linear projection of $\boldsymbol{x}_t$, the (vector) stimulus at time t, onto the receptive field $\boldsymbol{k}$; this linear stage is then followed by a simple scalar nonlinearity $f(.)$ which shapes the output (and in particular enforces the non-negativity of the output firing rate $\rho(t)$). A great deal of the systems neuroscience literature concerns the quantification of the receptive field parameters $\boldsymbol{k}$.

https://doi.org/10.1017/CBO9781107447615.013 Published online by Cambridge University Press

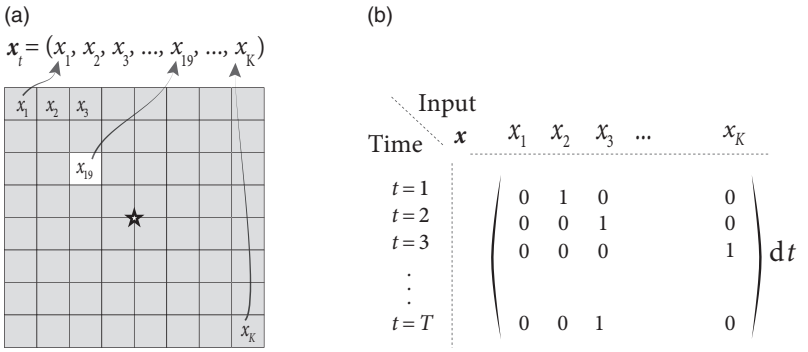

Fig. 11.6 Spatial receptive field measurement. (a) While the animal focuses on the center (star), light dots are presented at random positions on a gray screen; in the present trial, pixel 19 lights up. The input is denoted as a vector $\boldsymbol{x}_t$. (b) Schematic of input matrix X. Matrix representing a sparse input, such as a single spot of light. Rows of the matrix correspond to different trials, marked by the observation time t.

Note that the LNP model neglects the spike-history effects that are the hallmark of the SRM and the GLM – otherwise the two models are surprisingly similar; see Fig. 11.5b. Therefore, an LNP model cannot account for refractoriness or adaptation, while a GLM in the form of a generalized soft-threshold integrate-and-fire model does. The question arises whether a model with spike-history effects yields a better performance than the standard LNP model.

Both models, LNP and GLM, can be fitted using the methods discussed in Chapter 10. For example, the two models have been compared on a dataset where retinal ganglion cells have been driven by full-field light stimulus, i.e., the stimulus did not have any spatial structure (Pillow *et al.*, 2005). Prediction performance had a similar range of values as for cortical neurons driven by intracellular current injection, with up to 90% of the PSTH variance predicted in some cases. LNP models in this context have significantly worse prediction accuracy; in particular, LNP models greatly overestimate the variance of the predicted spiking responses. See Fig. 11.7 for an example.

Example: Detour on reverse correlation for receptive field estimation

Reverse correlation measurements are an experimental procedure based on spike-triggered averaging (de Boer and Kuyper, 1968; Chichilnisky, 2001). Stimuli $\boldsymbol{x}$ are drawn from some statistical ensemble and presented one after the other. Each time the neuron elicits a spike, the stimulus $\boldsymbol{x}$ presented just before the firing is recorded. The reverse correlation filter is the mean of all inputs that have triggered a spike

$$\boldsymbol{x}_{\text{RevCorr}} = \langle \boldsymbol{x} \rangle_{\text{spike}} = \frac{\sum_t n_t \boldsymbol{x}_t}{\sum_t n_t}, \tag{11.6}$$

https://doi.org/10.1017/CBO9781107447615.013 Published online by Cambridge University Press

where n_t is the spike count in trial t. Loosely speaking, the reverse correlation technique finds the typical stimulus that causes a spike. In order to make our intuitions more precise, we proceed in two steps.

First, let us consider an ensemble $p(\boldsymbol{x})$ of stimuli $\boldsymbol{x}$ with a "power" constraint $|\boldsymbol{x}|^2 < c$. Intuitively, the power constraint means that the maximal light intensity across the whole screen is limited. In this case, the stimulus that is most likely to generate a spike under the linear receptive field model (11.10) is the one which is aligned with the receptive field

$$\boldsymbol{x}_{\text{opt}} \propto \boldsymbol{k}\,; \tag{11.7}$$

see Exercises. Thus the receptive field vector $\boldsymbol{k}$ can be interpreted as the optimal stimulus to cause a spike.

Second, let us consider an ensemble of stimuli $\boldsymbol{x}$ with a radially symmetric distribution, where the probability of a possibly multi-dimensional $\boldsymbol{x}$ is equal to the probability of observing its norm $|\boldsymbol{x}| : p(\boldsymbol{x}) = p_c(|\boldsymbol{x}|)$. Examples include the standard Gaussian distribution, or the uniform distribution with power constraint $p(\boldsymbol{x}) = p_0$ for $|\boldsymbol{x}|^2 < c$ and zero otherwise. We assume that spikes are generated with the LNP model of Eq. (11.5). An important result is that the experimental reverse correlation technique yields an unbiased estimator of the filter $\boldsymbol{k}$, i.e.,

$$\langle \boldsymbol{x}_{\text{RevCorr}} \rangle = \boldsymbol{k}\,. \tag{11.8}$$

The proof (Bussgang, 1952; Simoncelli *et al.*, 2004) follows from the fact that each arbitrary input vector $\boldsymbol{x}_t$ can be separated into a component parallel to $\boldsymbol{k}$ and one orthogonal to it. Since we are free to choose the scale of the filter $\boldsymbol{k}$ we can impose $|\boldsymbol{k}| = 1$ and write

$$\boldsymbol{x}_t = (\boldsymbol{k} \cdot \boldsymbol{x}_t)\boldsymbol{k} + (\mathbf{e} \cdot \boldsymbol{x}_t)\,\mathbf{e} \tag{11.9}$$

where $\mathbf{e}$ is a unit vector in the subspace orthogonal to $\boldsymbol{k}$. For firing, only the component parallel to $\boldsymbol{k}$ matters. The symmetry of the distribution $p(\boldsymbol{x})$ guarantees that spike-triggered averaging is insensitive to the component orthogonal to $\boldsymbol{k}$; see Exercises.

In summary, reverse correlations are an experimental technique to determine the receptive field properties of a sensory neuron under an LNP model. The success of the reverse correlation technique as an experimental approach is intimately linked to its interpretability in terms of the LNP model.

Reverse correlations in the LNP model can also be analyzed in a statistical framework. To keep the arguments simple, we focus on the *linear* case and set

$$f(\boldsymbol{k} \cdot \boldsymbol{x}_t) = \rho_0 + \boldsymbol{k} \cdot \boldsymbol{x}_t\,. \tag{11.10}$$

The parameters minimizing the squared error between the model firing rate $\rho_0 + \boldsymbol{k} \cdot \mathbf{x}$ and the observed firing rate n_t are then

$$\boldsymbol{k}_{\text{opt}} = (X^{\mathrm{T}}X)^{-1} \left(\sum_t n_t \boldsymbol{x}_t \right) / \mathrm{d}t\,. \tag{11.11}$$

https://doi.org/10.1017/CBO9781107447615.013 Published online by Cambridge University Press

We note that, for short observation intervals Δ, the spike count n_t is either zero or 1. Therefore the term in parentheses on the right-hand side is proportional to the classical spike-triggered average; see Eq. (11.6). The factor in front of the parentheses, $X^{\mathrm{T}}X$, is a scaled estimate of the covariance of the inputs $\boldsymbol{x}$. For stimuli consisting of uncorrelated white noise or light dots at random positions, the covariance structure is particularly simple.

See Paninski (2004) for further connections between reverse correlation and likelihood-based estimates of the parameters in the LNP model.

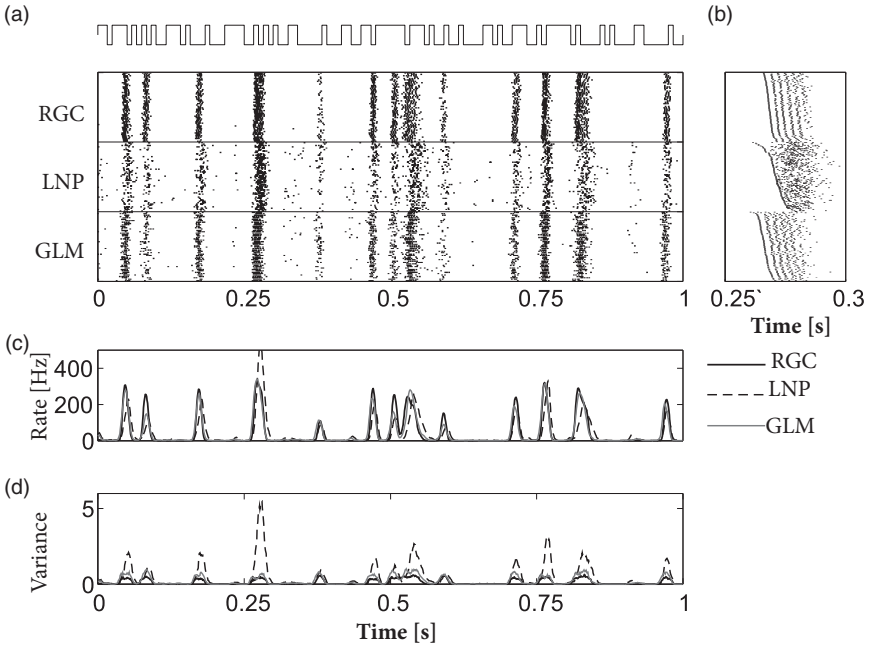

Fig. 11.7 Example predictions of retinal ganglion ON-cell (RGC) activity using the generalized linear encoding model with and without spike-history terms. (a) Recorded responses to repeated full-field light stimulus (top) of true ON cell ("RGC"), simulated LNP model (no spike-history terms; "LNP"), and Generalized Linear Model including spike-history terms ("GLM"). Each row corresponds to the response during a single stimulus presentation. (b) Magnified sections of rasters, with rows sorted in order of first spike time within the window in order to show spike-timing details. Note that the predictions of the model including spike-history terms are in each case more accurate than those of the Poisson (LNP) model. (PSTH variance accounted for: 91%, compared to 39% for the LNP model). (c) Time-dependent firing rate plotted as a PSTH. (d) Variance of the time-dependent firing rate. All data shown here are cross-validated "test" data (i.e., the estimated model parameters were in each case computed based on a non-overlapping "training" dataset not shown here). From Paninski *et al*. (2007) based on data from Uzzell and Chichilnisky (2004).

https://doi.org/10.1017/CBO9781107447615.013 Published online by Cambridge University Press

11.2.2 Multiple neurons

Using multi-electrode arrays, Pillow *et al.* (2008) recorded from multiple ganglion cells in the retina provided with spatio-temporal white noise stimuli. This stimulation reaches the ganglion cells after being transducted by photoreceptors and interneurons of the retina. It is assumed that the effect of light stimulation can be taken into account by a linear filter of the spatio-temporally structured stimulus. An SRM-like model of the membrane potential of a neuron i surrounded by n other neurons is

$$u_i(t) = \sum_f \eta_i(t - t_i^f) + \boldsymbol{k}_i \cdot \boldsymbol{x}(t) + \sum_{j \neq i} \sum_f \varepsilon_{ij}(t - t_j^f) + u_{\text{rest}}. \tag{11.12}$$

The light stimulus $\boldsymbol{x}(t)$ filtered by the receptive field of neuron i, $\boldsymbol{k}_i$, replaces the artificially injected external current in Eq. (11.1). The spike-afterpotential $\eta_i(t)$ affects the membrane potential as a function of the neuron's own spikes. Spikes from a neighboring neuron j modify the membrane potential of neuron i according to the coupling function $\varepsilon_{ij}(t)$.

The extracellular electrodes used by Pillow *et al*. (2008) did not probe the membrane potential. Nonetheless, by comparing the spike times with the conditional firing intensity $\rho(t|\{S\}) = \frac{1}{\tau_0} \exp(u(t))$ we can maximize the likelihood of observing the set of spike trains $\{S\}$ (Chapter 10). This way we can identify the spatio-temporal receptive field $\boldsymbol{k}_i$, the spike-afterpotential $\eta(t)$ and the coupling functions $\varepsilon_{ij}(t)$.

The fitted functions $\boldsymbol{k}_i$ showed two types of receptive fields (Pillow *et al*., 2008). The ON cells were sensitive to recent increase in luminosity while the OFF cells were sensitive to recent decrease. The coupling functions also reflect the two different neuron types. The coupling from ON cells to ON cells is excitatory and the coupling from ON cells to OFF cells is inhibitory, and conversely for couplings from OFF cells.

How accurate are the predictions of the multi-neuron model? Figure 11.8 describes the prediction performance. The spike-trains and PSTHs of the real and modeled neurons are similar. The spike-train likelihood reaches 2 bits per spike and the PSTH is predicted with 80–93% accuracy. Overall, the coupled model appears as a valid description of neurons embedded in a network.

Pillow *et al*. (2008) also asked about the relevance of coupling between neurons. Are the coupling functions an essential part of the model or can the activity be accurately predicted without them? Optimizing the model with and without the coupling function independently, they found that the prediction of PSTH variance was unaffected. The spike prediction performance, however, showed a consistent improvement for the coupled model. (See (Vidne *et al*., 2012) for further analysis using a model incorporating unobserved common noise effects.) Interneuron coupling played a greater role in decoding, as we shall see in the next section.

11.3 Decoding

"Decoding" refers to the problem of how to "read out" the information contained in a set of neural spike trains (Fig. 11.9) and has both theoretical and practical implications for the

https://doi.org/10.1017/CBO9781107447615.013 Published online by Cambridge University Press

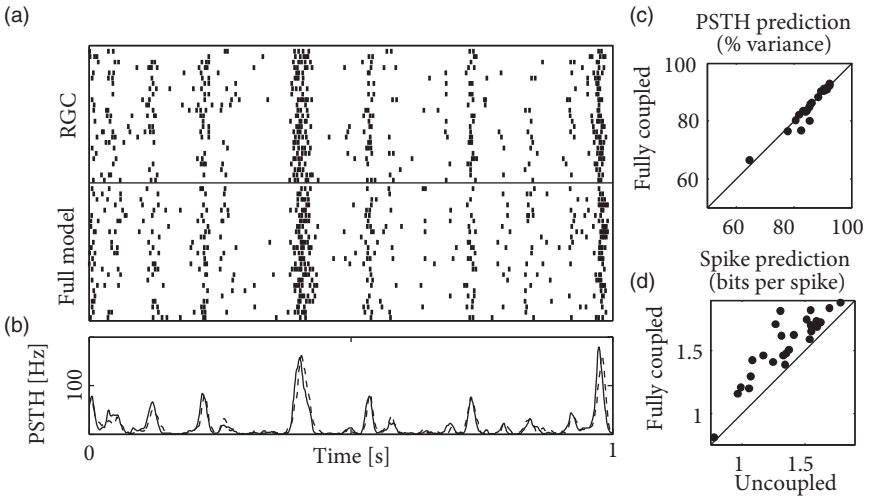

Fig. 11.8 Spike-train prediction of a retinal ganglion cell within its network. (a) Raster of responses of a retinal ganglion cell (RGC; top) to 25 repetitions of 1 s stimulus, and responses of the fully coupled model (Full model; bottom) to the same stimulus. (b) PSTH of the RGC (full black line) and the fully coupled model (dashed black line). (c) PSTH prediction for the fully coupled model of different cells plotted against the the PSTH prediction of a model fitted without interneuron coupling. (d) Log-likelihood (Eq. (10.41)) for the fully coupled model of different cells plotted against the log-likelihood of a model fitted without interneuron coupling. Modified from Pillow *et al.* (2008).

study of neural coding (Rieke *et al.*, 1997; Donoghue, 2002). A variety of statistical techniques have been applied to this problem (Rieke *et al.*, 1997; E. Brown *et al.*, 1998; Pillow *et al.*, 2011; Ahmadian *et al.*, 2011b); in this section, we focus specifically on decoding methods that rely on Bayesian "inversion" of the generalized linear encoding model discussed above and in Chapter 10. That is, we apply Bayes' rule to obtain the posterior probability of the stimulus, conditional on the observed response:

$$p(\boldsymbol{x}|D) \propto p(D|\boldsymbol{x})p(\boldsymbol{x}), \tag{11.13}$$

where $p(\boldsymbol{x})$ is the prior stimulus probability. As an aside we note that a similar idea was used above when we incorporated prior knowledge to regularize our estimates of the encoding model parameter θ; here we are assuming that θ, or equivalently $p(D|\boldsymbol{x})$, has already been estimated to a reasonable degree of precision, and now we want to incorporate our prior knowledge of the stimulus $\boldsymbol{x}$.

The primary appeal of such Bayesian decoding methods is that they are optimal if we assume that the encoding model $p(D|\boldsymbol{x})$ is correct. Decoding therefore serves as a means for probing which aspects of the stimulus are preserved by the response, and also as a tool for comparing different encoding models. For example, we can decode a spike train using different models (e.g., including vs. ignoring spike-history effects) and examine which

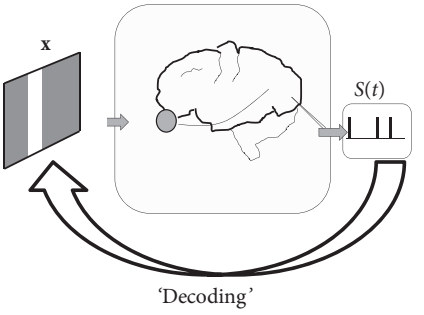

Fig. 11.9 The decoding problem in visual neuroscience. How much can we learn about a stimulus, given the spike trains of a group of neurons in the visual pathway?

encoding model allows us to best decode the true stimulus (Pillow *et al.*, 2005). Such a test may in principle give a different outcome than a comparison which focuses on the encoding model's ability to predict spike-train statistics. In what follows, we illustrate how to decode using the stimulus which maximizes the posterior distribution $p(\boldsymbol{x}|D)$, and show how a simple approximation to this posterior allows us to estimate how much information the spike-train response carries about the stimulus.

11.3.1 Maximum a posteriori *decoding*

The *maximum a posteriori* (MAP) estimate is the stimulus $\boldsymbol{x}$ that is most probable given the observed spike response D, i.e., the $\boldsymbol{x}$ that maximizes $p(\boldsymbol{x}|D)$. Computing the MAP estimate for $\boldsymbol{x}$ once again requires that we search in a high-dimensional space (the space of all possible stimuli $\boldsymbol{x}$) to find the maximizer of a nonlinear function, $p(\boldsymbol{x}|D)$. Luckily, in the GLM, the stimulus $\boldsymbol{x}$ interacts linearly with the model parameters θ, implying that concavity of the log-likelihood with respect to $\boldsymbol{x}$ holds under exactly the same conditions as does concavity in θ (Paninski, 2004). Moreover, the sum of two concave functions is concave, so the log-posterior,

$$\log p(\boldsymbol{x}|D) = \log p(D|\boldsymbol{x}) + \log p(\boldsymbol{x}) + c, \tag{11.14}$$

is concave as long as the stimulus log-prior $\log p(\boldsymbol{x})$ is itself a concave function of $\boldsymbol{x}$ (e.g., p is Gaussian). In this case, again, we may easily compute $\hat{x}_{\mathrm{MAP}}$ by numerically ascending the function $\log p(\boldsymbol{x}|D)$.

We emphasize that the MAP estimate of the stimulus is, in general, a *nonlinear* function of the observed spiking data D. As an empirical test of the MAP estimate, we can compare its performance with that of the optimal *linear* estimate (OLE, see example below), the best linear estimate of the stimulus as a function of the observed spiking data D (Rieke *et al.*, 1997).

Figure 11.10 shows a comparison of the two decoding techniques, given responses D generated by a GLM encoding model with known parameters, as a function of stimulus

https://doi.org/10.1017/CBO9781107447615.013 Published online by Cambridge University Press

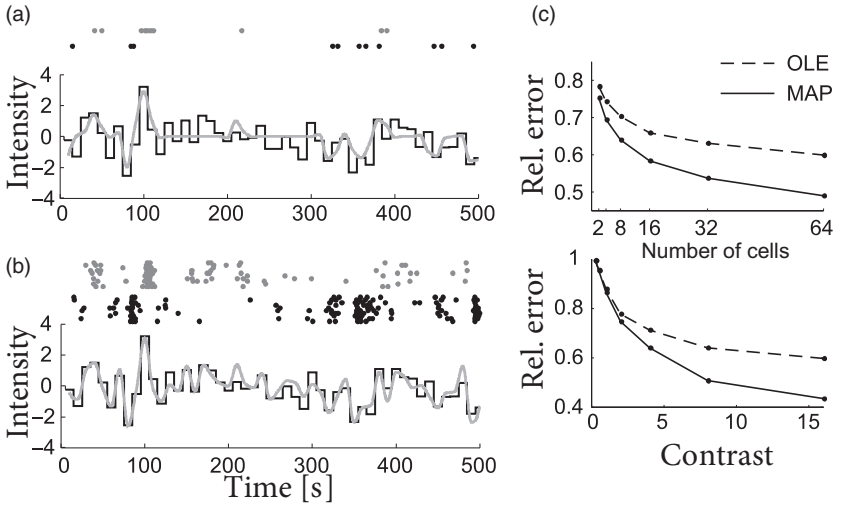

Fig. 11.10 Illustration of MAP (*maximum a posteriori*) decoding. (a) Simulated spike trains from a single pair of simulated ON and OFF retinal ganglion cells (above, gray and block dots) were used to compute the MAP estimate (gray) of a 500 ms Gaussian white noise stimulus (black), sampled at 100 Hz. (b) Spike trains from 10 identical, independent ON and OFF cells in response to the same stimulus, with the associated MAP estimate of the stimulus, illustrating convergence to the true stimulus as the responses of more cells are observed. (c) Comparison of the optimal linear estimate (OLE) and MAP estimate on simulated data, as a function of the number of observed cells (top) and stimulus contrast (variance; bottom). For each data point, the parameters of the OLE were estimated using a long run of simulated data. "Relative error" denotes the average RMS error between the true and estimated stimulus, averaged over 100 trials, divided by the RMS amplitude of the true stimulus.

contrast (variance) and size of the neuronal population. The MAP clearly outperforms the OLE at high contrasts or large population sizes. More importantly, the MAP approach provides us with a great deal of flexibility in considering different encoding models or prior distributions: we can simply substitute in a new $p(D|\boldsymbol{x})$ or $p(\boldsymbol{x})$ and recompute the MAP estimator, without having to obtain new estimates of the regression parameters as required by the OLE; see (Ramirez *et al.*, 2011) for an example of this type of analysis. Finally, there are close connections between MAP decoding and the optimal control of neural spiking; see Ahmadian *et al.* (2011a) for further discussion.

Example: Linear stimulus reconstruction

We predict the stimulus $\boldsymbol{x}_t$ by linear filtering of the observed spike times $t^1, t^2, \ldots, t^F < t$,

$$x(t) = x_0 + \sum_f k(t - t^f) \tag{11.15}$$

https://doi.org/10.1017/CBO9781107447615.013 Published online by Cambridge University Press

where the sum runs over all spike times. The aim is to find the shape of the filter k, i.e., the optimal linear estimator (OLE) of the stimulus (Rieke *et al.*, 1997).

Parameters of the OLE can be obtained using standard least-squares regression of the spiking data onto the stimulus $\boldsymbol{x}$. To do so, we discretize time and the temporal filter k. Mathematically, the optimization problem is then essentially the same as above where we aimed at predicting spikes by a linear model of the stimulus (Section 10.2). The only difference is that here we are regressing the spikes onto the stimulus, whereas previously we were regressing the stimulus onto the spike response.

11.3.2 Assessing decoding uncertainty (*)

In addition to providing a reliable estimate of the stimulus underlying a set of spike responses, computing the MAP estimate $\hat{x}_{\mathrm{MAP}}$ gives us easy access to several important quantities for analyzing the neural code. In particular, the variance of the posterior distribution around $\hat{x}_{\mathrm{MAP}}$ tells us something about which stimulus features are best encoded by the response D. For example, along stimulus axes where the posterior has small variance (i.e., the posterior declines rapidly as we move away from $\hat{x}_{\mathrm{MAP}}$), we have relatively high certainty that the true $\boldsymbol{x}$ is close to $\hat{x}_{\mathrm{MAP}}$. Conversely, we have relatively low certainty about any feature axis along which the posterior variance is large.

We can measure the scale of the posterior distribution along an arbitrary axis in a fairly simple manner: since we know (by the above concavity arguments) that the posterior is characterized by a single "bump," and the position of the peak of this bump is already characterized by $\hat{x}_{\mathrm{MAP}}$, it is enough to measure the curvature of this bump at the peak $\hat{x}_{\mathrm{MAP}}$. Mathematically, we measure this curvature by computing the "Hessian" matrix A of second derivatives of the log-posterior,

$$A_{ij} = -\frac{\partial^2}{\partial x_i \partial x_j} \log p(x_i|D)\,. \tag{11.16}$$

Moreover, the eigendecomposition of this matrix A tells us exactly which axes of stimulus space correspond to the "best" and "worst" encoded features of the neural response: small eigenvalues of A correspond to directions of small curvature, where the observed data D poorly constrains the posterior distribution $p(\boldsymbol{x}|D)$ (and therefore the posterior variance will be relatively large in this direction), while conversely large eigenvalues in A imply relatively precise knowledge of $\boldsymbol{x}$, i.e., small posterior variance (Huys *et al.*, 2006) (for this reason the Hessian of the log-likelihood $p(D|x)$ is referred to as the "observed Fisher information matrix" in the statistics literature). In principle, this posterior uncertainty analysis can potentially clarify what features of the stimulus a "downstream" neuron might care most about.

We can furthermore use this Hessian to construct a useful approximation to the posterior $p(\boldsymbol{x}|D)$. The idea is simply to approximate this log-concave bump with a Gaussian function, where the parameters of the Gaussian are chosen to exactly match the peak and curvature

https://doi.org/10.1017/CBO9781107447615.013 Published online by Cambridge University Press

of the true posterior. This approximation is quite common in the physics and statistics literature (Kass and Raftery, 1995; E. Brown *et al*., 1998; Rieke *et al*., 1997). Specifically,

$$p(\boldsymbol{x}|D) \approx (2\pi)^{-d/2}|A|^{1/2}\mathrm{e}^{-(\boldsymbol{x}-\hat{x}_{\mathrm{MAP}})^{\mathrm{T}}A(\boldsymbol{x}-\hat{x}_{\mathrm{MAP}})/2}, \tag{11.17}$$

with $d = \dim(\boldsymbol{x})$. We can then read off the approximate posterior entropy or variance of x_i: e.g., $\mathrm{var}(x_i|D) \approx [A^{-1}]_{ii}$. As discussed further in Ahmadian *et al*. (2011b) and Pillow *et al*. (2011), the approximation by the Gaussian of Eq. (11.17) is often quite accurate in the context of decoding. See Rieke *et al*. (1997) and Pillow *et al*. (2011) for discussion of a related bound on the posterior entropy, which can be used to bound the mutual information between the stimulus and response.

11.3.3 Decoding in vision and neuroprosthetics

We have established that generalized integrate-and-fire models can predict with good accuracy the activity of real neurons. Is it sensible to assert that we have understood how the neural system translates stimulus into patterns of action potentials? If so we should be able to read the neuronal activity to reconstruct its tangible meaning. As was briefly discussed in the introduction to Section 11.3, reading the neural code has practical applications; a common example of such applications is to help tetraplegic patients to control artificial limbs. In this section, we illustrate decoding in two distinct scenarios. In the first scenario (Fig. 11.11), a monochrome movie is reconstructed from the activity of neurons in the visual pathway. In the second example (Fig. 11.12), it is the time-dependent velocities of hand movements that are decoded from activity in the area MI of the cortex.

Using the methods described in the introduction to Section 11.3, Pillow *et al*. (2008) reconstructed the time-dependent light stimulus from 27 ganglion cells recorded in the retina. First, coupled integrate-and-fire models were optimized on training data (see Section 11.2.2). Once the appropriate set of parameter was determined, spike trains from the data reserved for testing were used to decode the stimulus. Decoding was performed with the methods discussed in the introduction to Section 11.3.

The stimulus was a spatio-temporal binary white noise. The decoding performance can be quantified by evaluating the signal-to-noise ratio for different frequencies (Fig. 11.11). For most of the frequencies, the signal-to-noise ratio of the decoded signal was greater than 1, meaning that the decoded signal was greater than the error. For the fully coupled model discussed in Section 11.2.2, the signal-to-noise ratio can be higher than 3.5 for some frequencies. The decoding performance is expected to grow with the number of recorded neurons, as can be seen in Fig. 11.11c.

We now consider a second example which has applications for neuroprosthetics. The ultimate aim of neuroprosthetics is to help human patients who have lost a limb. Prosthetic limbs are often available for these patients. While prosthesis works from a mechanical point of view, the intuitive control of the prosthetic device poses big challenges. One possible route of research is to read out, directly from the brain, the intentions of the user of the prosthetic device.

https://doi.org/10.1017/CBO9781107447615.013 Published online by Cambridge University Press

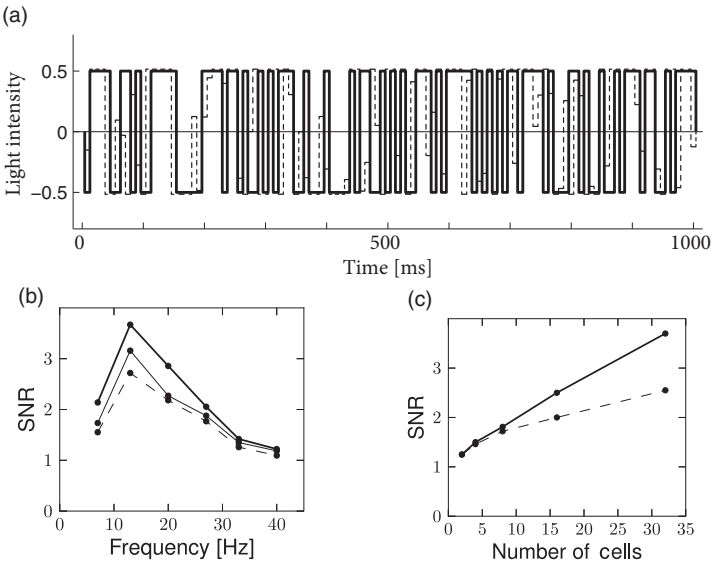

Fig. 11.11 Decoding of light stimulus from recordings of neurons in the retina. (a) Binary light stimulus (thick black) is compared with the decoded stimulus using Bayesian MAP (dashed line, Section 11.3.1). (b) The signal-to-noise ratio (SNR) as function of frequency for decoding using the fully coupled model (thick line), the uncoupled model (thin line) or using an optimal linear decoder (dashed lines). (c) Increasing the number of cells improves the decoding performance of both the coupled model (thick line) and the optimal linear decoder (dashed lines). (a) and (c) are redrawn from Pillow *et al*. (2011), (b) follows a similar figure to one in Pillow *et al*. (2008).

In preliminary experiments, a monkey moves a tracking device with his hand while an electrode records from neurons of its cortex. The electrode is placed in an area associated with planning movements (Donoghue *et al*., 1998). Truccolo *et al*. (2005) used generalized integrate-and-fire models to decode the hand movements from the recorded activity.

Again here, the first step was to fit the model parameters. The model itself was very similar to the one seen in Section 11.2.2 but without coupling terms and with a different nonlinear relation between model membrane potential and firing intensity. A more noteworthy difference is the input x which consisted of hand velocity such that the receptive field k mapped how the x- and y-components of the velocity influenced the driving potential.

Instead of the method described in Section 11.3.1, Truccolo *et al*. (2005) used a point-process filter (Eden *et al*., 2004). The decoding algorithm is a recursive algorithm for calculating the Bayesian estimate of the stimulus at time t in term the past activity. This recursive approach is necessary in this real-time application. The decoding performance is illustrated in Fig. 11.12. Signal-to-noise ratio for this decoding was between 1.0 and 2.5, which is rather impressive given the typical variability of cortical neurons and the small number of cells used for decoding (between 5 and 20). This exemplifies that generalized integrate-and-fire models can help in building a brain–machine interface for controlling

https://doi.org/10.1017/CBO9781107447615.013 Published online by Cambridge University Press

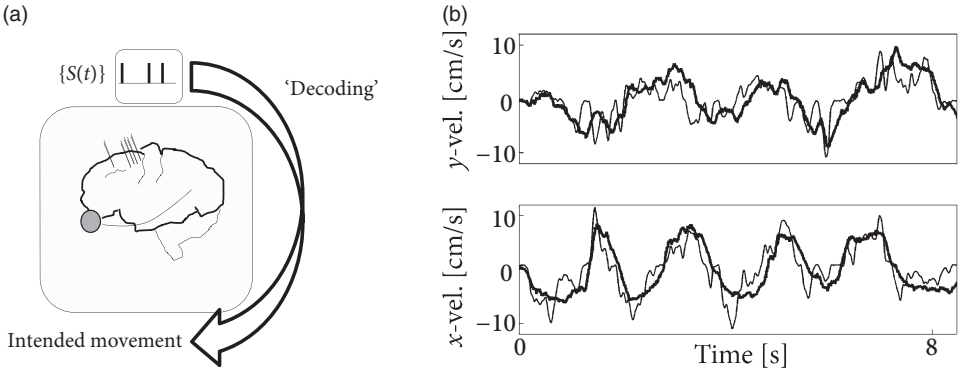

Fig. 11.12 Decoding hand velocity from spiking activity in area MI of cortex. (a) Schematics. (b) The real hand velocity (thin black line) is compared to the decoded velocity (thick black line) for the *y*-(top) and the *x*-components (bottom). Modified from Truccolo *et al*. (2005).

prosthetic limbs by "reading" the activity of cortical neurons. A number of groups are now working to further improve these methods for use in prosthetic systems.

11.4 Summary

Generalized integrate-and-fire models can predict the spiking activity of cortical cells such as the main inhibitory and excitatory cell type in layer 2–3 of the cortex. For excitatory neurons, more than 80% of spike timings can be predicted by these models, while for inhibitory neurons the percentage is close to 100%. Similar model performance is seen in the retina, where the activity of up to 250 neurons can be predicted simultaneously (Vidne *et al*., 2012).

The same models can also be used to decode the activity of neurons. For instance, the spike trains of retinal neurons can be decoded so as to reconstruct a slightly blurred version of the original image movie shown to the retina. Also, the activity of motor cortical neurons can be decoded to reconstruct the intended hand movement in two (or more) dimensions. Thus, the abstract mathematical framework of generalized integrate-and-fire models might ultimately contribute to technical solutions that help human patients.

Literature

The influential book by Rieke *et al*. (1997) gives a broad introduction to the field of neural coding with a special focus on decoding. The LNP model, reverse correlation techniques, and application to receptive field measurements are reviewed in Simoncelli *et al*. (2004).

Predictions of spike timings for a time-dependent input with models including spike-history effects were performed by, for example, Keat *et al*. (2001) and Jolivet *et al*. (2006),

and different methods and approaches were compared in a series of international competitions (Jolivet *et al*., 2008a,b).

The first decoding attempts used time-averaged firing rates to decode information from a diverse population of neurons (Georgopoulos *et al*., 1986). Then the methods were made more precise in an effort to understand the temporal structure of the neural code (Optican and Richmond, 1987; Bialek *et al*., 1991). In particular linear stimulus reconstruction from measured spike trains (Rieke *et al*., 1997) has been widely applied.

Efficient decoding methods are a necessary requirement if a prosthetic arm is controlled by the spikes recorded from cortical neurons. Introducing spike history effects (Truccolo *et al*., 2005) or interneuron coupling (Pillow *et al*., 2008) helped to improve decoding accuracy, but the improvement of decoding techniques went in parallel with other technical achievements (Shoham, 2001; Brockwell *et al*., 2004, 2007; Eden *et al*., 2004; Truccolo *et al*., 2005; Srinivasan and Brown, 2007; Kulkarni and Paninski, 2007; Koyama *et al*., 2010; Paninski *et al*., 2010).

The discussion of the statistical principles of encoding and decoding in the present and the previous chapter is partly based on the treatment in Paninski *et al*. (2007).

Exercises

1. **Linear filter as optimal stimulus**. *Consider an ensemble of stimuli* $\boldsymbol{x}$ *with a "power" constraint* $|\boldsymbol{x}|^2 < c$.

   (a) Show that, under the linear rate model of Eq.(11.10)*, the stimulus that maximizes the instantaneous rate is* $\boldsymbol{x} = \boldsymbol{k}$.

   Hint: Use Lagrange multipliers to implement the constraint $|\boldsymbol{x}|^2 = c$.

   (b) Assume that the a spatially localized time-dependent stimulus $x(t)$ *is presented in the center of the positive lobe of the neurons receptive field. Describe the neuronal response as*

$$\rho(t) = \rho_0 + \int_0^S \kappa(s)\, x(t-s)\mathrm{d}s, \tag{11.18}$$

   where ρ_0 *is the spontaneous firing rate in the presence of a gray screen and S the temporal extent of the filter* κ*. What stimulus is most likely to cause a spike under the constraint* $\int_0^S [x(t-s)]^2 \mathrm{d}s < c$*? Interpret your result.*

2. **LNP model and reverse correlations**. *Show that, if an experimenter uses stimuli* $\boldsymbol{x}$ *with a radially symmetric distribution* $p(\boldsymbol{x}) = q(|\boldsymbol{x}|)$*, then reverse correlation measurements provide an unbiased estimate linear filter* $\boldsymbol{k}$ *under an LNP model*

$$\rho(t) = f(\boldsymbol{k} \cdot \boldsymbol{x}_t); \tag{11.19}$$

   i.e., the expectation of the reverse correlation is parallel to $\boldsymbol{k}$.

   Hint: Write the stimulus as

$$\boldsymbol{x} = (\boldsymbol{k} \cdot \boldsymbol{x})\boldsymbol{k} + (\mathbf{e} \cdot \boldsymbol{x})\,\mathbf{e} \tag{11.20}$$

   and determine the reverse correlation measurement by averaging over all stimuli weighted with their probability to cause a spike.

https://doi.org/10.1017/CBO9781107447615.013 Published online by Cambridge University Press