NOTE Communicated by Carl van Vreeswijk

Analytical Model for the Effects of Learning on Spike Count Distributions

Gianni Settanni
Alessandro Treves
SISSA, Programme in Neuroscience, Trieste, Italy

The spike count distribution observed when recording from a variety of neurons in many different conditions has a fairly stereotypical shape, with a single mode at zero or close to a low average count, and a long, quasi-exponential tail to high counts. Such a distribution has been suggested to be the direct result of three simple facts: the firing frequency of a typical cortical neuron is close to linear in the summed input current entering the soma, above a threshold; the input current varies on several timescales, both faster and slower than the window used to count spikes; and the input distribution at any timescale can be taken to be approximately normal. The third assumption is violated by associative learning, which generates correlations between the synaptic weight vector on the dendritic tree of a neuron, and the input activity vectors it is repeatedly subject to. We show analytically that for a simple feedforward model, the normal distribution of the slow components of the input current becomes the sum of two quasi-normal terms. The term *important below threshold* shifts with learning, while the term *important above threshold* does not shift but grows in width. These deviations from the standard distribution may be observable in appropriate recording experiments.

1 Spike Counts and the S + F model

The variability in the emission of action potentials by nerve cells may be characterized by several measures, one of which is the distribution of spike counts in a time window of fixed length. While other measures, such as the distribution of consecutive interspike intervals, are more direct descriptions of the irregularity in the firing, spike count distributions provide estimates of the entropy of neural codes, inasmuch as the code used by a particular neuron is thought to be expressed simply by its firing frequency. Because of this interpretation, the observation that spike count distributions often have quasi-exponential tails (Abeles, Vaadia, & Bergman, 1990; Barnes, McNaughton, Mizumori, Leonald, & Lin, 1990) has been linked to an optimal coding principle (Levy & Baxter, 1996). If the spike count is taken to be the symbol coding for the message represented

by the input arriving at the cell at any given time, and the mean count is constrained by a fixed metabolic budget, then optimal usage of symbols occurs, in the noiseless case, when their probabilities are exponentially distributed—that is, have maximal entropy under the constraint (Shannon, 1948). It is not clear, however, how this coding principle could apply to the more meaningful situation of noisy coding and, more important, how it could be compatible with deviations from the pure exponential shape often prominent at the low-count end of the distribution, that is, the nonzero mode.

An alternative suggestion (Panzeri, Booth, Wakeman, Rolls, & Treves, 1996; Treves, Panzeri, Rolls, Booth, & Wakeman, 1999) is that the stereotypical shape does not reflect any design or optimization, but rather reflects precisely the lack of any principle capable of imparting significant structure to the distribution; that is, it represents a sort of null hypothesis against which any organizing principle could be tested. The suggestion is embodied in a crude model of the variability of the input current into the soma of a typical neuron, which assumes that (1) the input current translates linearly into a spike count—of course, above a firing threshold; (2) its variability has frequency components at several different time scales, both slower and faster (hence the name *S + F model*) than the time window used to count spikes; and (3) each component is normally distributed. The S + F model generates a formula for the spike count distribution with three free parameters: the position of the mean current relative to firing threshold and the standard deviations of slow and fast components of its variability. The formula was found to fit adequately spike counts recorded from monkey inferior temporal cortex neurons responding to quasi-ecological stimulation with a video of natural images, counts that could not instead be fitted by the exponential or other simple models (Treves et al., 1999).

The S + F model was then found to provide satisfactory fits in other experiments, such as with primate hippocampal neurons responding to continuously changing views of the monkey's environment (Panzeri, Rolls, Treves, Robertson, & Georges-François, 1997) or with a class of rat somatosensory neurons spontaneously active under anesthesia (Irina Erchova, pers. comm.). This raises the issue of whether it is at all possible to identify situations in which a well-defined principle or process does affect firing statistics, and its effects can be quantitatively demonstrated in the spike count distributions, as deviations from the null hypothesis expectation. We consider here a model of associative learning of discrete input vectors by a (feedforward) network comprising a single output neuron. While the S + F model neglects correlations between afferent input patterns and synaptic weights, the learning process considered here produces precisely such correlations, and we have calculated analytically the resulting changes in spike count distributions. The possibility of observing these changes in experiments is discussed briefly at the end.

2 An S + F Model Refined by Learning

The correlations induced by associative learning could affect any frequency component in the variability of the input current. To stick to a simple model, we consider only the case in which they affect solely the low-frequency components. This corresponds to the simplified scenario in which the slow variability in the input (and output) of our neuron codes for meaningful, slow-varying, stimuli, while fast fluctuations, reflect only noise. We thus take fast fluctuations to be normally distributed, as in the standard S + F model, and in fact do not include them, leaving them to be added up at the end, after deriving the distribution of the slow components. This should not be taken to imply that fast variability is negligible (it is not; see Treves et al., 1999), but the analytical convenience of concentrating on slow variability warrants the minor imprecision in the transparent expression obtained at the end.

The single output unit in the model receives N inputs, through synaptic weights $\mathbf{J}$ that modify with a learning rule that models associative plasticity. Learning is one shot, in that the weights are taken to have been modified by a single presentation of each of p input patterns. The learning rule includes balanced potentiation and depression components, so that the net average change of each weight is zero. The variance in the value of each weight is also taken to remain constant. The input patterns are uncorrelated among themselves and with the preexisting synaptic weight vector. Under these conditions, the distribution of the summed input current over all novel input pattern vectors remains the same (normal in the $N \to \infty$ limit) after learning. What changes, and what we are going to compute, is the distribution of the input current over the p familiar input vectors—those that have been learned.

The output distribution is calculated with the mean-field analysis detailed in the appendix. For the sake of clarity, we try to keep the notation consistent with Treves (1995), where a similar calculation was reported. Storage and retrieval (S and R) here refer to the first presentation of one of the p input patterns, and to the subsequent presentation that generates the output distribution we aim for.

η is the input vector during storage. It represents firing rates computed over a time window of, say, a few hundred milliseconds, and may be measured in Hz. We take each of its N components to be distributed independently,

$$P(\eta) = \prod_i P_\eta(\eta_i), \tag{2.1}$$

according to some distribution P_η, which for consistency should itself be of the stereotypical form mentioned above. One result we find, though, is that the precise form of P_η is not critical.

$\mathbf{V}$ is the input vector during retrieval. The stimulus that has been previously learned is taken to have been reproduced with some added gaussian

noise δ (with zero mean and variance σ_δ^2), followed by rectification:[1]

$$V_i = [\eta_i + \delta_i]^+ . \tag{2.2}$$

Therefore

$$P(\mathbf{V}|\eta) = \prod_i \left[\delta(V_i)\Phi\left(-\eta_i/\sigma_\delta\right) + \frac{\Theta(V_i)}{\sqrt{2\pi}\,\sigma_\delta} \exp -\frac{(V_i - \eta_i)^2}{2\sigma_\delta^2} \right], \tag{2.3}$$

where $\delta(x)$ is the Dirac delta function, $\Theta(x)$ is the Heaviside function, and $\Phi(x)$ is the normalized probability integral, that is, the integral of the normal distribution of unit variance: $\Phi(x) = \int_{-\infty}^{x} (dx/\sqrt{2\pi}) \exp -x^2/2$. The first term in each factor of the product represents the probability that the input unit i be below threshold, and the second term that it be above threshold. The noise level σ_δ parameterizes the variability in the firing frequency of input units, measured over a trial of, say, a few hundred milliseconds, among trials with the same stimulus. It is thus a measure of slow components in the noise, which could again be taken to be similar, like the frequency distribution, between input and output units, and it could in principle be evaluated experimentally. Note that the distributions $P(\eta)$ and $P(V)$ need not be assumed to be identical, even if both take the general stereotypical form; the addition of the noise term δ, which induces a difference between the two, could be thought to reflect the altered attentional state, for example, of successive with respect to the first presentation of a novel stimulus.

Z is the output during storage resulting from the product between input and synaptic vectors followed by thresholding and rectification. Gaussian noise ϵ^S with zero mean and variance σ_ϵ^2 is also added to the output,

$$Z = \left[\mathbf{J}^S \cdot \eta + Z_\circ + \epsilon^S \right]^+ . \tag{2.4}$$

The threshold $(-Z_\circ)$ may lump together a bona-fide current threshold, inhibitory terms due to nonselective effects of interneurons and competition among output cells, and the baseline mean value of the product $\mathbf{J}^L \cdot \eta$, taken to be constant. The gain of the threshold-linear transfer function has been set to 1 by rescaling synaptic weights to pure numbers of order $1/N$, so that Z, ϵ^S, and Z_0, like η, may be measured in Hz. The output distribution during storage is

$$P(Z|\eta) = \delta(Z)\Phi\left(-\frac{\mathbf{J}^S \cdot \eta + Z_\circ}{\sigma_\epsilon} \right) + \frac{\Theta(Z)}{\sqrt{2\pi}\,\sigma_\epsilon} \exp -\frac{(Z - \mathbf{J}^S \cdot \eta - Z_\circ)^2}{2\sigma_\epsilon^2}, \tag{2.5}$$

[1] $[x]^+ = x$ for $x > 0$ and 0 otherwise.

where the rectification has been applied directly, without first adding fast fluctuations. This omission, which is carried over in the synaptic modification terms below, is deliberate; it simplifies analytically what was already a rather crude model.

u is the steady component of the summed input current to the neuron during retrieval. It is convenient to calculate its distribution before adding fast noise and rectification, after which operations one would have the actual output during retrieval, U. If we take u to include slow noise with the same variance σ_ϵ^2, it will follow the conditional probability density,

$$P(u|Z, \mathbf{V}, \eta) = \frac{1}{\sqrt{2\pi}\,\sigma_\epsilon} \exp -\frac{(u - \mathbf{J}^R \cdot \mathbf{V} - U_\circ)^2}{2\sigma_\epsilon^2}, \tag{2.6}$$

which we have to average appropriately in order to find the target distribution.

$\mathbf{J}^S$ and $\mathbf{J}^R$ are the weight vectors during storage and retrieval. With respect to the pattern being considered, they are assumed to have random components, except for a term, present in $\mathbf{J}^R$ but absent in $\mathbf{J}^S$, that reflects the storage of the pattern and produces the effect on the spike count distribution, during retrieval, which is the aim of the calculation. Each component J_i^S is then taken to have zero mean (any baseline value can be incorporated into the constant threshold $-Z_\circ$) and fixed variance $\sigma_J^2 = 1/N$ (so that the scalar product $\mathbf{J}^S \cdot \eta$ is of order 1). Apart from the relevant pattern, the components of the new vector J_i^R reflect also the storage of many other patterns, intervened between storage and retrieval. A stable regime, in which storing new patterns and gradually forgetting old ones does not alter the mean strength and variance of the vector, can be described (after Treves, 1995) by a pseudo-rotation from the random direction $\mathbf{J}^S$ into a new random direction $\mathbf{J}^N$,

$$J_i^R = \cos(\theta)\, J_i^S + \sqrt{\gamma}\,\frac{h}{\sqrt{N}}(\eta_i - \bar{\eta})(Z - Z_\circ) + \sin(\theta)\, J_i^N, \tag{2.7}$$

where:

- J_i^S is multiplied by a factor $\cos(\theta)$, that reduces its relative importance with time, parameterized by θ. J_i^N, again of zero mean and variance $\sigma_J^2 = 1/N$, is multiplied by $\sin(\theta)$, which grows with the successive learning of new patterns. The relation of θ to real time need not be detailed, except that obviously $\theta = 0$ implies that no other pattern has modified the weights between the storage and retrieval of the one being considered, while $\theta = \pi/2$ implies complete oblivion of the original weights.
- The associative modification term is multiplied by a normalizing factor $h/\sqrt{N}$, designed to ensure that the variance of the modification term

is set to $\gamma\,\sigma_J^2$. Since $\sigma_J^2 = 1/N$, one needs to adjust $1/h^2$ to the value of the variance of the η and Z factors. In this way the plasticity γ can be taken to be the average proportion of the synaptic weight variance accounted for by a single learned pattern[2] if all that J^R encodes are p patterns, with equal strength, $\gamma = 1/p$.

The quantity calculated, in the large-N limit, is the distribution $\langle P(u)\rangle$ of the static or low-frequency component of the current into the output neuron. The brackets indicate averaging over all possible values of the weight vectors. To convert u into an output spike count, one would need to consider additional high-frequency components of the noise (as explained by Treves et al., 1999), to multiply by a gain (which, if different among patterns, cannot be taken to equal 1) and discretize the resulting output frequency U into a number of spikes emitted. These steps are needed when analyzing experimental results, but since they could be handled in a number of alternative ways, they are beyond the scope of this article, which focuses on how $\langle P(u)\rangle$ differs from a normal distribution.

3 Result and Parameters

The calculation reported in the appendix yields the expression

$$\begin{aligned}\langle P(u)\rangle = \frac{1}{\sqrt{2\pi\,\det\mathbf{T}}}\Bigg\{&\Phi\left[-\frac{\beta_1(u-U_\circ+Z_\circ g)+\beta_0 Z_\circ}{\sqrt{\beta_0}}\right]\\ &\times\frac{1}{\sqrt{\beta_0}}\exp\left[-\frac{(u-U_\circ+Z_\circ g)^2}{2\beta_0\,\det\mathbf{T}}\right]\\ &+\Phi\left[\frac{(\beta_1+\beta_2 g)(u-U_\circ)+(\beta_0+2\beta_1 g+\beta_2 g^2)Z_\circ}{\sqrt{\beta_0+2\beta_1 g+\beta_2 g^2}}\right]\\ &\times\frac{1}{\sqrt{\beta_0+2\beta_1 g+\beta_2 g^2}}\\ &\times\exp\left[-\frac{(u-U_\circ)^2}{2(\beta_0+2\beta_1 g+\beta_2 g^2)\,\det\mathbf{T}}\right]\Bigg\},\end{aligned}\tag{3.1}$$

where the matrix $\mathbf{T}$ and the matrix elements $\beta_{0,1,2}$ are

$$\mathbf{T} = \begin{pmatrix}\sigma_\epsilon^2 + z_\circ & \cos(\theta)w_\circ\\ \cos(\theta)w_\circ & \sigma_\epsilon^2 + y_\circ\end{pmatrix}\tag{3.2}$$

$$\mathbf{T}^{-1} = \begin{pmatrix}\beta_0 & -\beta_1\\ -\beta_1 & \beta_2\end{pmatrix},\tag{3.3}$$

[2] For consistency, a decay factor $\cos(\theta)$ should express the gradual forgetting of this learned pattern, along with the others. We assume such factor to be incorporated into the $\sqrt{\gamma}$ factor, purely to simplify the resulting formulas.

and the averages $x_\circ$, $y_\circ$, $w_\circ$, and $z_\circ$ are defined in the appendix. Learning is now parameterized by $g = hx_\circ\sqrt{N\gamma}$, which, apart from the number $hx_\circ$ of order unity, can be seen to measure roughly the inverse square root of the storage load, that is, of the effective number of learned patterns $1/\gamma$ divided by the number of inputs N.

The expression is a sum of two terms: the first originating from cases in which the original unthresholded response ζ was below threshold and the second from those with $\zeta > 0$.[3] Each term is a gaussian modulated by a $\Phi(x)$ factor, which suppresses the gaussian for values of u not matching the corresponding ζ. Thus, the first term is more important for u values below threshold, and once the current is converted into a spike count, its detailed form will be largely uninfluential. The second term is more important above threshold and is more directly observable. The partition in two term therefore stems from the thresholding (rectification) of the response ζ, which is the crucial assumption, while the detailed form of each term reflects all other ingredients of the model.

In the limit of zero plasticity, $\gamma \to 0$ so that also $g \to 0$, both terms reduce to normal distributions of mean $U_\circ$ and variance $\beta_0 \det \mathbf{T} \equiv \sigma_\epsilon^2 + y_\circ$, with the two prefactors adding up to unity. The model then reduces to a normal distribution of slow fluctuations, as in the standard S + F model, with their variance given by the sum of that of "slow noise", σ_ϵ^2, and "signal" ($y_\circ$ can be seen as the product of the variance of synaptic weights and that of the activity of input units).

The two terms depend on learning (i.e., on plasticity) in two simple but different ways. One should first realize what range of values is accessible for the parameter g. The parameter γ in principle can range from 0 (no plasticity) to 1 (the entire synaptic variance is due to the storage of the single memory pattern being examined); a meaningful set of values, though, is around $1/N$, since $p \simeq \mathcal{O}(N)$ is the memory capacity of associative nets (Rolls & Treves, 1998). Given that, besides $\sqrt{\gamma N}$, the other factors that determine g reduce to a number of order unity, g itself is nonnegative, and can be considered to range from 0 to values of order 1.

As g increases from 0, the first gaussian has its mean shifted by an amount $-Z_\circ g$, that is, toward negative values if the mean output during the learning phase, $Z_\circ$, is positive, and to positive values in the (also common) case in which the distribution of responses to novel stimuli—ideally derived from the $\langle P(Z)\rangle$ distribution—corresponds to a negative mean of the slow component of the S + F current. The width of this first gaussian does not change. The modulating factor Φ contributes to make the former effect difficult to detect, as it suppresses further, for $Z_\circ > 0$, the gaussian peak.

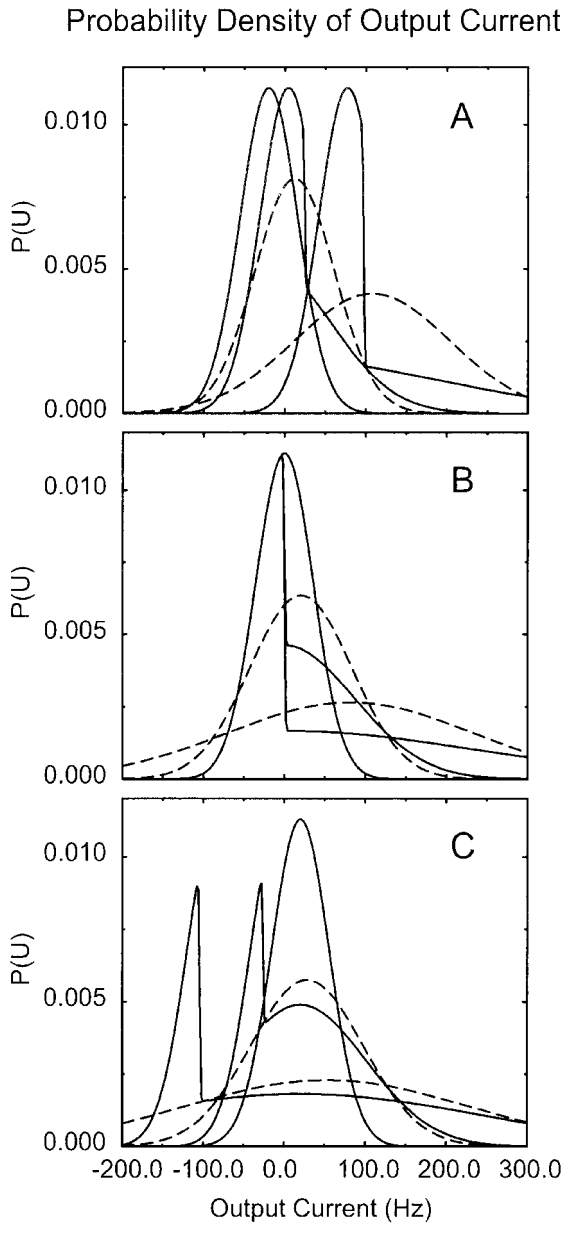
[3] Integrating $\langle P(\zeta, u)\rangle$ first over u and then over ζ, one can check that the normalization of equation 3.1 is correct.

The second gaussian has the mean unaffected by learning, while its width grows, roughly doubling in value when g reaches values of order 1 (if the β factors, which are all positive, are also of similar magnitude). The modulating prefactor has a complex dependence on g, but is in any case larger for $Z_\circ > 0$.

The distribution depends on two additional parameters: the noise on input variables, σ_δ, and the input distribution $P_\eta(\eta)$, which theoretically has infinite degrees of freedom. However, it depends on both in a rather mild way, only through the averages $x_\circ$, $y_\circ$, $w_\circ$, and $z_\circ$. In fact, it can be seen that for $\sigma_\delta \to 0$, that is, negligible input noise, the specific form of P_η becomes irrelevant, and only its first two momenta matter. For the sake of simplicity, we use a binary P_η in the graphs, but any other choice gives similar results. The binary P_η reduces to a probability $1 - a$ for the input activity to be 0, and a probability a to be at the arbitrary level of 50 Hz; a parameterizes the sparseness of the activity (Treves & Rolls, 1991) in that it gives, for such a binary distribution, the fraction of active units.

The graphs show the effect of increasing values of the learning parameter g for three cases that correspond to negative, zero, and positive mean levels of the output current, and thus reproduce three typical regimes of firing statistics. In each case, we set $Z_\circ = U_\circ$ and equal, and very low, noise levels in the input and output $\sigma_\epsilon = \sigma_\delta = 0.5$ Hz. Three values of the learning parameter are included, which correspond to $\sqrt{N\gamma} = 0.0$, $\sqrt{N\gamma} = 1.45$, and $\sqrt{N\gamma} = 5.8$. The resulting g values differ in each figure as they depend on the value of the hx_0 factor and are reported in the legend. In any case the three values correspond to no plasticity, intense plasticity, and very strong plasticity ($\sqrt{N\gamma} = 5.8$ implies that even taking $N \simeq 10^4$, a single pattern accounts for about 1/300 of the variance of synaptic weights, on

Figure 1: *Facing page.* The effects of learning on spike count distributions that fall largely (a) below threshold, (b) around threshold, or (c) above threshold. Each graph shows as solid curves the normal distribution $P(u)$ for $g = 0$, and the distributions obtained with two different values of the learning parameter, while the dashed lines indicate the gaussian curves that best fit each distribution. (a) The mean value of the original distribution is $U_\circ = Z_\circ = -20$ Hz and the two nonzero values of the plasticity correspond to $g = 1.21$ and $g = 4.86$. (b) The mean value is $U_\circ = Z_\circ = 0$ Hz, while the same plasticity values result in $g = 1.44$ and $g = 5.77$. (c) $U_\circ = Z_\circ = 20$ Hz, $g = 1.30$ and $g = 5.21$. The g values used in each of the three panels are somewhat different because of the different h factor, resulting from a different $\langle (Z - Z_\circ)^2 \rangle$ average. Other parameters: $a = 0.5$, $\theta = 0$, $\sigma_\delta = \sigma_\epsilon = 0.5$ Hz. Since differences in the portion of the distribution below threshold are difficult to detect in practice, the most favorable situation to experimentally observing learning effects is that exemplified in panel a, with the original distribution largely below threshold, and the modified distribution substantially shifted to higher spike counts.



average). The sparseness is set in each case at $a = 0.5$; different sparseness values change the distributions quantitatively but not the qualitative effects of learning. In the first case, Figure 1a, the original output distribution is

above threshold only with its upper tail, $Z_\circ = -20$ Hz. The main effect is a shift of the bulk of the distribution to higher values, with its width remaing constant, while the upper tail is broadened. This situation (mode of the original distribution below threshold, that is quasi-exponential spike count) corresponds to a fraction of cortical recordings and may be more typical of hippocampal cells (Panzeri et al., 1997). In the second case, Figure 1b, the mean output in the absence of learning is exactly at threshold, $Z_\circ = 0$ Hz. The half distribution below threshold then stays unchanged, while the half distribution above is broadened by learning. The latter effect is the dominant one in the third case, Figure 1c, in which the original output spike count has been taken to have a peak above threshold, $Z_\circ = 20$ Hz. In this case, only the lower tail of the distribution shifts, and in the opposite direction, to lower values below threshold (an effect difficult to detect in practice).

The effects on the distribution are somewhat more complex if $U_\circ \neq Z_\circ$, and it would take several figures to describe the detailed dependence on θ, σ_δ, σ_ϵ, and so on. However, Figure 1 provides a useful indication of the main effects—those that could be hoped to be observed in experiments—and the low noise levels used make such effects particularly salient.

4 Can Deviations from Normality Be Observed?

Figure 1 also shows, with dashed lines, the normal distributions in u that most closely match (in the least mean square sense) the actual distribution for the two nonzero values of the learning parameter. Visual inspection of the figure clarifies the relative likelihood of observing the effects of learning in experiments. In experiments in which the only data are the firing statistics to presumably well-learned visual stimuli, the effects of learning have to be demonstrated as mismatches between the observed spike count and the closest underlying normal distribution of (slow) fluctuations. Figure 1 indicates that such a mismatch will be substantial only below threshold, and then more so for distributions concentrated below threshold (as the one in the example of Figure 1a). The shape of the distribution of slow fluctuations below threshold is difficult to extract, particularly with the limited data available in practice, from the observed spike count. It is therefore likely that deviations from normality, if they indeed occur as an effect of learning, will be observable only in experiments designed for that purpose, in particular, involving extensive sampling (recording sessions). A thorough analysis will be needed to disentangle among deviations from normality, if observed, those due to learning from those due to any of the many simplifying assumptions of our model.

The situation is different if the effects of learning can also be gauged by the deviations of the observed spike count from a "control" spike count obtained in equivalent conditions, but with the cell stimulated with novel (nonlearned) stimuli. In that case the parameters of the normal distribution

assumed by the S + F model for $g = \gamma = 0$ can in principle be extracted from the data and used to try to fit the spike count obtained with the familiar (presumably learned) stimuli. Therefore, not only the shape of the distribution of slow fluctuations but also its estimated mean and variance will offer indications about the effects of learning, whether they are consistent with those predicted by the analysis and, if so, yield estimates of the learning parameters g and γ. Now deviations from normality due to other effects can hopefully be subtracted out.

Experiments that in principle allow for such a comparison have been carried out in the laboratory of James Ringo (Ringo & Nowicka, 1996). The idea is to train a monkey to discriminate among a set of 12 to 40 visual images, which are generated as simple combinations of elementary shapes and colors, with a given algorithm. These images are used throughout training and become familiar to the monkey. During testing, the statistics of single cell responses to such images are contrasted with the statistics of responses to a much larger set of images, each of which is novel but generated by the same algorithm. Under such conditions, any difference in the statistics can be used, in particular the mean and variance of the spike count (Ringo & Nowicka, 1996) and, after analysis with the S + F model, the estimated mean and variance of the distribution of slow fluctuations. An analysis along these lines is in progress (facing subtleties posed, as usual, by limited sampling) and will be reported elsewhere.

In addition to limited sampling, the analysis of experimental recording has to confront effects inherent in the behavior of real neurons, which have not been considered in this simple model. For example, for experiments in which steady stimuli are presented in successive trials, each for a fixed time, in principle one has to take into account the variability in the latency of the response, adaptation in the firing, across-trials trends in the response to the same stimulus, and so on. The quantification of these effects requires a study of its own, beyond the scope of this article, but their presence should clearly be borne in mind.

In conclusion, a simple (threshold-linear) model predicts simple effects of learning on the statistics of trial-to-trial fluctuations in the input current to a neuron. These effects can be summarized by the next-order-cumulant rule: below threshold, where the output, fast noise aside, is zero (i.e., constant with respect to the input current), the effect of learning is a linear increase in the first-order cumulant of the input distribution (i.e., its mean value); above threshold, where the output is roughly linear in the input, the effect is on the second-order cumulant (i.e., the variance). Experimental constraints make the validation of these model predictions not quite straightforward; but if the effects turn out to be observable, they will allow an estimate, at least as an order of magnitude, of the plasticity parameter, that is, a measure of the amount of learning stored in the synapses to a neuron.

Appendix

The quantity we have evaluated, $\langle P(u)\rangle$, is the average across all possible values of the synaptic weight vector of the distribution $P(u)$, which is itself already integrated over all values of η, $\mathbf{V}$, and Z. It can thus be written

$$\langle P(u)\rangle = \int \prod_i \left(\frac{dJ_i^S}{\sqrt{2\pi}\,\sigma_J} e^{-\frac{(J_i^S)^2}{2\sigma_J^2}} \right) \prod_i \left(\frac{dJ_i^N}{\sqrt{2\pi}\,\sigma_J} e^{-\frac{(J_i^N)^2}{2\sigma_J^2}} \right) \times \int d\mathbf{V}\, dZ\, d\eta P(u|Z, \mathbf{V}, \eta) P(\mathbf{V}|\eta) P(Z|\eta) P(\eta). \tag{A.1}$$

To proceed, one has just to insert the conditional probabilities from equations 2.3, 2.5, and 2.6 and carry out a succession of integrals. It is convenient to write $P(\mathbf{V}|\eta)$ and $P(Z|\eta)$ as pure gaussian integrals over the dummy variables v_i and ζ, of which V_i and Z are the real parts (e.g., $Z = \zeta\Theta(\zeta)$). Moreover, the normal distributions of the variables ζ and u around their mean values can be written as gaussian integrals in the noise terms ϵ^S and ϵ^R. One can then integrate over the synaptic weights, obtaining

$$\begin{aligned}\langle P(u)\rangle = & \int_{-\infty}^{+\infty} \frac{d\epsilon^R}{2\pi\sigma_\epsilon^2} \exp\left[\frac{i\epsilon^R(u - U_\circ)}{\sigma_\epsilon^2} - \frac{(\epsilon^R)^2}{2\sigma_\epsilon^2} \right] \\ & \times \int_{-\infty}^{+\infty} \frac{d\epsilon^S d\zeta}{2\pi\sigma_\epsilon^2} \exp\left[\frac{i\epsilon^S(\zeta - Z_\circ)}{\sigma_\epsilon^2} - \frac{(\epsilon^S)^2}{2\sigma_\epsilon^2} \right] \int \prod_i \frac{d\delta_i dv_i}{2\pi\sigma_\delta^2} d\eta_i P_\eta(\eta_i) \\ & \times \exp i \sum_i \left\{ \frac{\epsilon^R h}{\sigma_\epsilon^2} \sqrt{\frac{\gamma}{N}} (\eta_i - \bar{\eta})[\zeta\Theta(\zeta) - Z_\circ] v_i \Theta(v_i) + \frac{\delta_i(v_i - \eta_i)}{\sigma_\delta^2} \right\} \\ & \times \exp - \sum_i \left\{ \frac{\delta_i^2}{2\sigma_\delta^2} + \frac{\sigma_J^2}{2} \frac{[\eta_i\epsilon^S + \cos(\theta) v_i \Theta(v_i)\epsilon^R]^2}{\sigma_\epsilon^4} \right. \\ & \left. + \frac{\sigma_J^2}{2} \frac{[\sin(\theta) v_i \Theta(v_i)\epsilon^R]^2}{\sigma_\epsilon^4} \right\}. \end{aligned} \tag{A.2}$$

One may then introduce, in order to separate integration variables, the average input parameters,

$$x = \frac{1}{N} \sum_i (\eta_i - \bar{\eta}) V_i \tag{A.3}$$

$$y = \frac{1}{N} \sum_i V_i^2 \tag{A.4}$$

$$w = \frac{1}{N}\sum_i \eta_i V_i \tag{A.5}$$

$$z = \frac{1}{N}\sum_i \eta_i^2, \tag{A.6}$$

which are also integrated over, and constrained to take the above values by delta functions defined in terms of the conjugated parameters $\tilde{x}$, $\tilde{y}$, $\tilde{w}$, and $\tilde{z}$. The intermediate formula becomes

$$\begin{aligned}\langle P(u)\rangle = &\int \frac{N dx d\tilde{x}}{2\pi}\int \frac{N dy d\tilde{y}}{2\pi}\int \frac{N dw d\tilde{w}}{2\pi}\int \frac{N dz d\tilde{z}}{2\pi}\\ &\times \int_{-\infty}^{+\infty}\frac{d\epsilon^{\mathrm{R}}}{2\pi\sigma_\epsilon^2}\int_{-\infty}^{+\infty}\frac{d\epsilon^{\mathrm{S}} d\zeta}{2\pi\sigma_\epsilon^2}\exp NF\\ &\times \exp -\left\{\frac{(\epsilon^{\mathrm{R}})^2+(\epsilon^{\mathrm{S}})^2}{2\sigma_\epsilon^2}\right.\\ &\qquad\left.+\frac{\sigma_f^2 N}{2\sigma_\epsilon^4}[z(\epsilon^{\mathrm{S}})^2+2w\cos(\theta)\epsilon^{\mathrm{S}}\epsilon^{\mathrm{R}}+y(\epsilon^{\mathrm{R}})^2]\right\}\\ &\times \exp i\left\{\frac{\epsilon^{\mathrm{R}}(u-U_\circ)}{\sigma_\epsilon^2}+\frac{\epsilon^{\mathrm{R}}h}{\sigma_\epsilon^2}\sqrt{\gamma N}[\zeta\Theta(\zeta)-Z_\circ]x\right.\\ &\qquad\left.+\frac{\epsilon^{\mathrm{S}}(\zeta-Z_\circ)}{\sigma_\epsilon^2}\right\},\end{aligned} \tag{A.7}$$

where

$$\begin{aligned}\exp F = &\int \frac{d\delta dv}{2\pi\sigma_\delta^2}d\eta P_\eta(\eta)\exp\left[-\frac{\delta^2}{2\sigma_\delta^2}+i(x\tilde{x}+y\tilde{y}+w\tilde{w}+z\tilde{z})\right]\\ &\times\exp i\left[\frac{\delta(v-\eta)}{\sigma_\delta^2}-\tilde{x}(\eta-\bar{\eta})v\Theta(v)\right.\\ &\qquad\left.-\tilde{y}v^2\Theta(v)-\tilde{w}\eta v\Theta(v)-\tilde{z}\eta^2\right].\end{aligned} \tag{A.8}$$

Having set in the model $\sigma_f^2 N = 1$, it is possible to calculate the integrals in the $N\to\infty$ limit with the saddle point method, which amounts to using the formula,

$$\begin{aligned}&\lim_{N\to\infty}\int d^n x g(\mathbf{x})\exp(NF(\mathbf{x}))\approx\\ &\qquad\approx g(\mathbf{x}_{\max})\exp[NF(\mathbf{x}_{max})]\sqrt{\left(\frac{2\pi}{N}\right)^n}\sqrt{\frac{1}{-\det H[F(\mathbf{x}_{\max})]}},\end{aligned} \tag{A.9}$$

where $\mathbf{x}$ is a short hand for the $n = 8$-dimensional vector $(x, y, w, z, \tilde{x}, \tilde{y}, \tilde{w}, \tilde{z})$, which maximizes the exponent F in $\mathbf{x}_{\max}$, and $H[F(\mathbf{x}_{\max})]$ is the Hessian of F at the maximum. One finds the maximum of F at

$$\tilde{x}_\circ = \tilde{y}_\circ = \tilde{w}_\circ = \tilde{z}_\circ = 0 \tag{A.10}$$

$$x_\circ = \int d\eta P_\eta(\eta)(\eta - \bar{\eta}) \left[\eta \Phi\left(\frac{\eta}{\sigma_\delta}\right) + \frac{\sigma_\delta}{\sqrt{2\pi}} \exp\left(-\frac{\eta^2}{2\sigma_\delta^2}\right)\right] \tag{A.11}$$

$$y_\circ = \int d\eta P_\eta(\eta) \left[(\eta^2 + \sigma_\delta^2) \Phi\left(\frac{\eta}{\sigma_\delta}\right) + \frac{\eta\sigma_\delta}{\sqrt{2\pi}} \exp\left(-\frac{\eta^2}{2\sigma_\delta^2}\right)\right] \tag{A.12}$$

$$w_\circ = \int d\eta P_\eta(\eta)\eta \left[\eta \Phi\left(\frac{\eta}{\sigma_\delta}\right) + \frac{\sigma_\delta}{\sqrt{2\pi}} \exp\left(-\frac{\eta^2}{2\sigma_\delta^2}\right)\right] \tag{A.13}$$

$$z_\circ = \int d\eta P_\eta(\eta)\eta^2, \tag{A.14}$$

in which Φ is, as above, the normal distribution function. At the maximum,

$$F\ (x_\circ, y_\circ, w_\circ, z_\circ, \tilde{x}_\circ, \tilde{y}_\circ, \tilde{w}_\circ, \tilde{z}_\circ) = 0 \tag{A.15}$$

$$\det H\ [F(x_\circ, y_\circ, w_\circ, w_\circ, \tilde{x}_\circ, \tilde{y}_\circ, \tilde{w}_\circ, \tilde{z}_\circ) = -1, \tag{A.16}$$

so we are left with

$$\begin{aligned}
\langle P(u)\rangle &\approx g(x_\circ, y_\circ, w_\circ, z_\circ, \tilde{x}_\circ = 0, \tilde{y}_\circ = 0, \tilde{w}_\circ = 0, \tilde{z}_\circ = 0) \\
&= \int_{-\infty}^{+\infty} \frac{d\epsilon^{\mathrm{R}}}{2\pi\sigma_\epsilon^2} \int_{-\infty}^{+\infty} \frac{d\epsilon^{\mathrm{S}} d\zeta}{2\pi\sigma_\epsilon^2} \\
&\quad \times \exp i \left\{ \frac{\epsilon^{\mathrm{R}}(u - U_\circ)}{\sigma_\epsilon^2} + \frac{\epsilon^{\mathrm{R}} h}{\sigma_\epsilon^2} \sqrt{\gamma N} [\zeta\Theta(\zeta) - Z_\circ] x_\circ + \frac{\epsilon^{\mathrm{S}}(\zeta - Z_\circ)}{\sigma_\epsilon^2} \right\} \\
&\quad \times \exp - \left\{ \frac{(\epsilon^{\mathrm{R}})^2 + (\epsilon^{\mathrm{S}})^2}{2\sigma_\epsilon^2} \right. \\
&\qquad \left. + \frac{\sigma_J^2 N}{2\sigma_\epsilon^4} [z_\circ (\epsilon^{\mathrm{S}})^2 + 2w_\circ \cos(\theta) \epsilon^{\mathrm{S}} \epsilon^{\mathrm{R}} + y_\circ (\epsilon^{\mathrm{R}})^2] \right\}.
\end{aligned} \tag{A.17}$$

The above expression is but a gaussian integral in $\mathbf{R}^2$ in the variables ϵ^{S} and ϵ^{R}; one has to carry out the final integral in $d\zeta$ to obtain the expression in the text.

Acknowledgments

We are grateful to James Ringo, Stefano Panzeri, and Andrea Benucci for discussions of our results. This work was in partial fulfillment of require-

ments for an M.Sc. thesis (G.S.) and was supported in part by CNR, INFM, and HFSP.

References

Abeles, M., Vaadia, E., & Bergman, H. (1990). Firing patterns of single units in the prefrontal cortex and neural network models. *Network, 1*, 13–25.

Barnes, C. A., McNaughton, B. L., Mizumori, S. J. Y., Leonard, B. W., & Lin, L.-H. (1990). Comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing. In J. Storm-Mathisen, J. Zimmer, & O. P. Ottersen (Eds.), *Understanding the brain through the hippocampus* (pp. 287–300). Amsterdam: Elsevier.

Levy, W. B., & Baxter, R. A. (1996). Energy efficient neural codes. *Neural Comp., 8*, 531–543.

Panzeri, S., Booth, M., Wakeman, E. A., Rolls, E. T., & Treves, A. (1996). Do firing rate distributions reflect anything beyond just chance? *Society for Neuroscience Abstracts, 22*, 1124.

Panzeri, S., Rolls, E. T., Treves, A., Robertson, R. G., & Georges-François, P. (1997). Efficient encoding by the firing of hippocampal spatial view cells. *Society for Neuroscience Abstracts, 23*, 195.4.

Ringo, J. L., & Nowicka, A. (1996). Long term memory for visual images in the hippocampus and inferotemporal cortex. *Society for Neuroscience Abstracts, 22*, 281.

Rolls, E. T., & Treves, A. (1998). *Neural networks and brain function*. Oxford: Oxford University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *AT&T Bell Labs. Tech. J., 27*, 379–423.

Treves, A. (1995). Quantitative estimate of the information relayed by the Schaffer collaterals. *J. Comp. Neurosci., 2*, 259–272.

Treves, A., Panzeri, S., Rolls, E. T., Booth, M. C. A., & Wakeman, E. A. (1999). Firing rate distributions and efficiency of information transmission of inferior temporal cortex neurons to natural visual stimuli. *Neural Comp., 11*, 611–641.

Treves, A., & Rolls, E. T. What determines the capacity of autoassociative memories in the brain. *Network*, 2:371–397, 1991.

Received November 2, 1998; accepted June 8, 1999.