LETTER Communicated by Alexandre Pouget

Mutual Information, Fisher Information, and Population Coding

Nicolas Brunel
Jean-Pierre Nadal
Laboratoire de Physique Statistique de l'E.N.S.,[†] Ecole Normale Supérieure, 75231 Paris Cedex 05, France

In the context of parameter estimation and model selection, it is only quite recently that a direct link between the Fisher information and information-theoretic quantities has been exhibited. We give an interpretation of this link within the standard framework of information theory. We show that in the context of population coding, the mutual information between the activity of a large array of neurons and a stimulus to which the neurons are tuned is naturally related to the Fisher information. In the light of this result, we consider the optimization of the tuning curves parameters in the case of neurons responding to a stimulus represented by an angular variable.

1 Introduction

A natural framework to study how neurons communicate, or transmit information, in the nervous system is information theory (see, e.g., Blahut, 1988; Cover & Thomas, 1991). In recent years the use of information theory in neuroscience has motivated a large amount of work (e.g., Laughlin, 1981; Linsker, 1988; Barlow, Kaushal, & Mitchison, 1989; Bialek, Rieke, de Ruyter van Steveninck, & Warland, 1991; Van Hateren, 1992; Atick, 1992; Nadal & Parga, 1994). A neurophysiologist often asks in an informal sense how much information the spike train of a single neuron, or of a population of neurons, provides about an external stimulus. For example, a high activity of a CA3 hippocampal neuron may tell with good precision where a rat is in an environment. Information theory provides mathematical tools for measuring this "information" or "selectivity": signals are characterized by a probability distribution, and the spike train of a neuron, or of a population, is characterized by a probability distribution conditioned by the signal. The mutual information between the signal and the neural representation is then a measure of the statistical dependency between the signal and the spike train(s).

[†] Laboratory associated with CNRS, ENS, and universities Paris 6 and Paris 7.

Neural Computation **10, 1731–1757** (1998)

A related domain, which also belongs to information theory, is the field of statistical parameter estimation. Here one typically has a sample of observations drawn from a distribution that depends on a parameter, or a set of parameters, that one wishes to estimate. The Cramer-Rao inequality then tells us that the mean squared error of any unbiased estimator of the underlying parameter(s) is lower bounded by the inverse of a quantity, which is defined as the Fisher information (Blahut, 1988). This means that the Fisher information is a measure of how well one can estimate a parameter from an observation with a given probability law. Thus in this sense it is also an "information" quantity.

In spite of the similar intuitive meanings of these two quantities, an explicit relationship between the Fisher information and information-theoretic quantities has been derived only recently (Clarke & Barron, 1990; Rissanen, 1996), in the limit of a large number of observations. This link was exhibited first in the context of parameter estimation (Clarke & Barron, 1990) for the case of statistically independent and identically distributed observations. Then it was generalized to a broader context within the framework of stochastic complexity, with, as a result, a refined "minimum description length" criterion for model selection (Rissanen, 1996).

The first goal of this article is to show that within the framework of information theory, this link manifests itself very naturally in the context of neural coding:

- In the limit of a large number of neurons coding for a low-dimensional stimulus (population coding), the mutual information between the activities of the neuronal population and the stimulus becomes equal to the mutual information between the stimulus and an efficient gaussian estimator, under appropriate conditions, detailed in section 3. Here "efficient" means that the variance of this estimator reaches the Cramer-Rao bound. Since this variance is related to the Fisher information, the equality provides a quantitative link between mutual and Fisher informations.
- This equality is also shown to hold for a single cell in the case of a gaussian noise with vanishing variance.
- The mutual information between the stimulus and an efficient gaussian estimator reaches the mutual information between stimulus and the neuronal activities asymptotically from below.

In the light of this relationship between Fisher and mutual information, we examine in section 5 several issues related to population codes, using neurons coding for an angular variable with a triangular or bell-shaped tuning curve. Such neurons are common in many neural structures. Cells of the postsubiculum (Taube, Muller, & Ranck, 1990) and anterior thalamic nuclei (Taube, 1995) of the rat are tuned to its head direction. Cells in MT cortex (Maunsell & Van Essen, 1983) of the monkey are tuned to the direction of perceived motion. Cells in motor cortex of the monkey (Georgopoulos, Kalaska, Caminiti, & Massey, 1982) are tuned to the direction of the arm. We study the case of an array of N neurons, firing as a Poisson process in response to an angular stimulus with a frequency defined by the tuning curve of the neuron, in an interval of duration t. In many cases, Poisson processes are considered to be reasonable approximations of the firing process of cortical neurons (see, e.g., Softky & Koch, 1993).

We calculate the Fisher information with an arbitrary density of preferred angles. Next we address the question of the optimization of the tuning curves, making use of the link between mutual information and Fisher information. The optimal density of preferred angles (i.e., the one that maximizes the mutual information) is calculated as a function of the distribution of angles, in section 5.2. As shown by Seung and Sompolinsky (1993), the Fisher information, in the large N limit, diverges when the tuning width of the neurons goes to zero. We show in section 5.3 that a finite tuning width stems from optimization criteria, which consider a finite system in which only a small number of spikes has been emitted by the whole population. We illustrate our results using triangular tuning curves in section 5.4.

2 General Framework

2.1 Parameter Estimation and Population Coding. In the general context of parameter estimation, one wishes to estimate a parameter θ from a set of N observations $\{x_i, i = 1, \ldots, N\} \equiv \vec{x}$ (where the x_i's might be discrete or continuous). θ may characterize a model $P(\vec{x}|\theta)$, which is expected to be a good description of the stochastic process generating the observations $\{x_i\}$. In the simplest case, the x_i's are independent realizations of the same random variable, and

$$P(\vec{x}|\theta) = \prod_{i=1}^{N} p(x_i|\theta). \tag{2.1}$$

It may be the case—but this is not necessary—that the *true* process $p^*(x)$ belongs to the family under consideration, so that $p^*(x) = p(x|\theta_t)$ where θ_t is the *true* value of the parameter.

In the context of sensory coding, and more specifically population coding (see, e.g., Seung & Sompolinsky, 1993, Snippe, 1996), θ is a stimulus (e.g,. an angle), and the information about this stimulus is contained in the activities $\{x_i, i = 1, \ldots, N\}$ of a population of a large number N of neurons. In the simplest case x_i represents the activity of the ith neuron of the output layer of a feedforward network with no lateral connection, so that the probability density function (p.d.f.) $P(\vec{x}|\theta)$ is factorized:

$$P(\vec{x}|\theta) = \prod_{i=1}^{N} p_i(x_i|\theta). \tag{2.2}$$

Here $p_i(x_i|\theta)$ is the (neuron-dependent) p.d.f. of the activity x_i at neuron i when the input stimulus takes the value θ.

If the task of the neural system is to obtain a good estimate of the stimulus value, the problem is a particular case of parameter estimation where there exists a true value—the one that generated the observed activity $\vec{x}$.

2.2 The Cramer-Rao Bound. In general one can find different algorithms for computing an estimate $\hat{\theta}(\vec{x})$ of θ from the observation of $\vec{x}$. If the chosen estimator $\hat{\theta}$ (algorithm) is unbiased, that is, if

$$\int d^N x P(\vec{x}|\theta)\hat{\theta}(\vec{x}) \ = \ \theta,$$

the variance of the estimator,

$$\sigma_\theta^2 = \left\langle (\hat{\theta} - \theta)^2 \right\rangle_\theta ,$$

in which $\langle \ . \ \rangle_\theta$ denotes the integration over $\vec{x}$ *given* θ (a sum in the case of a discrete state vector) with the p.d.f. $P(\vec{x}|\theta)$, is bounded below according to (Cramer-Rao bound; see, e.g., Blahut, 1988):

$$\sigma_\theta^2 \geq \frac{1}{\mathcal{J}(\theta)} \tag{2.3}$$

where $\mathcal{J}(\theta)$ is the Fisher information:

$$\mathcal{J}(\theta) = \left\langle - \frac{\partial^2 \ \ln P(\vec{x}|\theta)}{\partial \ \theta^2} \right\rangle_\theta . \tag{2.4}$$

For a multidimensional parameter, equation (2.3) is replaced by an inequality for the covariance matrix, with $\mathcal{J}(\theta)$, the Fisher information matrix, being then expressed in terms of the second derivatives of $\ln P(\vec{x}|\theta)$ (Blahut, 1988). For simplicity we will restrict the discussion to the case of a scalar parameter, and consider the straightforward extension to the multidimensional case in section 3.2.

An *efficient* estimator is one that saturates the bound. The maximum likelihood (ML) estimator is known to be efficient in the large N limit.

3 Mutual Information and Fisher Information

3.1 Main Result. We now give the interpretation of the Cramer-Rao bound in terms of information content. First, note that the Fisher information (see equation (2.4)) is not itself an information quantity. The terminology comes from an intuitive interpretation of the bound: our knowledge ("information") about a stimulus θ is limited according to this bound. This qualitative statement has been turned into a quantitative statement in Clarke and Barron (1990) and Rissanen (1996). Here we give a different presentation based on a standard information-theoretic point of view, which is relevant for sensory coding, rather than from the point of view of parameter estimation and model selection.

We consider the mutual information between the observable $\vec{x}$ and the stimulus θ. It can be defined very naturally in the context of sensory coding because θ is itself a random quantity, generated with some p.d.f. $\rho(\theta)$, which characterizes the environment. The mutual information is defined by (Blahut, 1988):

$$I[\theta, \vec{x}] \;=\; \int d\theta d^N x \rho(\theta)\; P(\vec{x}|\theta)\; \log \frac{P(\vec{x}|\theta)}{Q(\vec{x})}, \tag{3.1}$$

where $Q(\vec{x})$ is the p.d.f. of $\vec{x}$:

$$Q(\vec{x}) = \int d\theta \rho(\theta) P(\vec{x}|\theta). \tag{3.2}$$

Other measures of the statistical dependency between input and output could be considered, but the mutual information is the only one (up to a multiplicative constant) satisfying a set of fundamental requirements (Shannon & Weaver, 1949).

Suppose there exists an unbiased efficient estimator $\hat{\theta} = T(\vec{x})$. It has mean θ and variance $1/\mathcal{J}(\theta)$. The amount of information gained about θ in the computation of that estimator is

$$I[\theta, \hat{\theta}] = \mathcal{H}[\hat{\theta}] - \int d\theta \rho(\theta) \mathcal{H}[\hat{\theta}|\theta], \tag{3.3}$$

where $\mathcal{H}[\hat{\theta}]$ is the entropy of the estimator,

$$\mathcal{H}[\hat{\theta}] = -\int d\hat{\theta}\, \Pr(\hat{\theta}) \ln \Pr(\hat{\theta}),$$

and $\mathcal{H}[\hat{\theta}|\theta]$ its entropy given θ. The latter, for each θ, is smaller than the entropy of a gaussian distribution with the same variance $1/\mathcal{J}(\theta)$. This implies

$$I[\theta, \hat{\theta}] \geq \mathcal{H}[\hat{\theta}] - \int d\theta \rho(\theta) \frac{1}{2} \ln \left(\frac{2\pi e}{\mathcal{J}(\theta)} \right). \tag{3.4}$$

Since processing cannot increase information (see, e.g., Blahut, 1988, pp. 158–159), the information $I[\theta, \vec{x}]$ conveyed by $\vec{x}$ about θ is at least equal to the one conveyed by the estimator: $I[\theta, \vec{x}] \geq I[\theta, \hat{\theta}]$. For the efficient estimator, this means

$$I[\theta, \vec{x}] \;\geq\; \mathcal{H}[\hat{\theta}] - \int d\theta \rho(\theta) \frac{1}{2} \ln \left(\frac{2\pi e}{\mathcal{J}(\theta)} \right). \tag{3.5}$$

In the limit in which the distribution of the estimator is sharply peaked around its mean value (in particular, this implies $\mathcal{J}(\theta) \gg 1$), the entropy of the estimator becomes identical to the entropy of the stimulus. The right-hand side (r.h.s.) in the above inequality then becomes equal to I_{Fisher} plus terms of order $1/\mathcal{J}(\theta)$, with I_{Fisher} defined as

$$I_{Fisher} = \mathcal{H}(\Theta) - \int d\theta \rho(\theta) \frac{1}{2} \ln \left(\frac{2\pi e}{\mathcal{J}(\theta)} \right). \tag{3.6}$$

In the above expression, the first term is the entropy of the stimulus,

$$\mathcal{H}(\theta) = -\int d\theta \rho(\theta) \ln \rho(\theta). \tag{3.7}$$

For a discrete distribution, this would be the information gain resulting from a perfect knowledge of θ. The second term is the equivocation due to the gaussian fluctuations of the estimator around its mean value. We thus have, in this limit of a good estimator,

$$I[\theta, \vec{x}] \ \geq \ I_{Fisher}. \tag{3.8}$$

The inequality (see equation 3.8), with I_{Fisher} given by equation 3.6, gives the essence of the link between mutual information and Fisher information. It results from an elementary application of the simple but fundamental theorem on information processing, and of the Cramer-Rao bound.

If the Cramer-Rao bound was to be understood as a statement on information content, $I[\theta, \vec{x}]$ could not be strictly larger than I_{Fisher}. If not, there would be a way to extract from $\vec{x}$ more information than I_{Fisher}. Hence the above inequality would be in fact an equality, that is:

$$I[\theta, \vec{x}] = -\int d\theta \rho(\theta) \ln \rho(\theta) - \int d\theta \rho(\theta) \frac{1}{2} \ln \left(\frac{2\pi e}{\mathcal{J}(\theta)} \right). \tag{3.9}$$

However, the fact that the equality should hold is not obvious. The Cramer-Rao bound does not tell us whether knowledge on cumulants other than the variance could be obtained. Indeed, if the estimator has a nongaussian distribution, the inequality will be strict; we will give an example in section 4 where we discuss the case of a single output cell ($N = 1$). In the large N limit, however, there exists an efficient estimator (the maximum likelihood), and relevant probability distributions become close to gaussian distributions, so that one can expect equation 3.9 to be true in that limit. This is indeed the case, and what is proved in Rissanen (1996) within the framework of stochastic complexity, under suitable but not very restrictive hypotheses.

In the appendix, we show, using completely different techniques, that equation 3.9 holds provided the following conditions are satisfied:

1. All derivatives of $G(\vec{x}|\theta) \equiv \ln P(\vec{x}|\theta)/N$ with respect to the stimulus θ are of order one.
2. The cumulants (with respect to the distribution $P(\vec{x}|\theta)$) of order n of $aG'_\theta + bG''_\theta$ are of order $1/N^{n-1}$ for all a, b, n.

The meaning of the second condition is that at a given value of N, the cumulants should decrease sufficiently rapidly with n. This is in particular true when x_i given θ are independent, as for model 2.2, but holds also in the more general case when the x_i are correlated, provided the above conditions hold, as we show explicitly in the appendix using an example of correlated x_i.

3.2 Extensions and Remarks.

Multiparameter Case and Model Selection. It is straightforward to extend equation 3.8 to the case of a K-dimensional stimulus $\vec{\theta}$ with p.d.f. $\rho(\vec{\theta}\ )$, and to derive the equality equation 3.9 for $K \ll N$. The Fisher information matrix is defined as (Blahut, 1988)

$$\mathcal{J}_{ij}\left(\vec{\theta}\ \right) = \left\langle -\frac{\partial^2 \ln P\left(\vec{x}|\vec{\theta}\ \right)}{\partial\theta_i \partial\theta_j} \right\rangle_{\vec{\theta}}.$$

The quantity I_{Fisher} for the multidimensional case is then

$$I_{Fisher} = -\int d^K\theta\ \rho(\vec{\theta}) \ln \rho(\vec{\theta}) - \int d^K\theta\ \rho(\vec{\theta}) \frac{1}{2} \ln\left(\frac{(2\pi e)^K}{\det \mathcal{J}(\vec{\theta})}\right). \qquad (3.10)$$

The second term is now equal to the entropy of a gaussian with covariance matrix $\mathcal{J}^{-1}(\vec{\theta}\ )$, averaged over $\vec{\theta}$ with p.d.f. $\rho(\vec{\theta})$. In the large N limit ($K << N$), one gets as for $K = 1$ the equality $I = I_{Fisher}$.

One can note that formulas 3.9 and 3.10 are also meaningful in the more general context of parameter estimation, even when θ is not a priori a random variable. Within the Bayesian framework (Clarke & Barron, 1990), it is natural to introduce a prior distribution on the parameter space, $\rho(\theta)$. Typically, this distribution is chosen as the flattest possible one that takes into account any prior knowledge or constraint on the parameter space. Then I tells us how well θ can be localized within the parameter space from the observation of the data $\vec{x}$.

Within the framework of MDL (minimum description length) (Rissanen, 1996) the natural prior is the one that maximizes the mutual information—that is, the one realizing the Shannon capacity. Maximizing $I = I_{Fisher}$ with respect to ρ, one finds that this optimal input distribution is given by the square root of the Fisher information:

$$\rho(\theta) = \frac{\sqrt{\mathcal{J}(\theta)}}{\int d\theta' \sqrt{\mathcal{J}(\theta')}}$$

(for the multidimensional case, $\mathcal{J}$ in the above expression has to be replaced by det $\mathcal{J}$). This corresponds to the stimulus distribution for which the neural system is best adapted.

Biased Estimators. The preceding discussion can be easily extended to the case of biased estimators, that is, for estimators $\hat{\theta}$ with $< \hat{\theta} >_\theta = m(\theta) \neq \theta$. The Cramer-Rao bound in such a case reads

$$\frac{\sigma_\theta^2}{\left(\frac{dm}{d\theta}\right)^2} \geq \frac{1}{\mathcal{J}(\theta)}. \tag{3.11}$$

This is a form of the bias-variance compromise. One can thus write an inequality similar to equation 3.4, replacing $\mathcal{J}$ by $\mathcal{J}/(dm/d\theta)^2$. In the limit where the estimator is sharply peaked around its mean value $m(\theta)$, one has $\rho(\theta)d\theta \sim P(\hat{\theta})d\hat{\theta}$, and $\hat{\theta} \sim m(\theta)$, so that

$$\mathcal{H}[\hat{\theta}] = \mathcal{H}[\theta] + \int d\theta \rho(\theta) \log |\frac{dm}{d\theta}|.$$

Upon inserting $\mathcal{H}[\hat{\theta}]$ in the r.h.s. of the inequality 3.4, the terms $\frac{dm}{d\theta}$ cancel. The bound, equation 3.8, is thus also valid even when the known efficient estimator is biased.

The Cramer-Rao bound can also be understood as a bound for the discriminability d' used in psychophysics for characterizing performance in a discrimination task between θ and $\theta + \delta\theta$ (see, e.g., Green & Swets, 1966). As discussed in Seung and Sompolinsky (1993),

$$d' \leq \delta\theta \sqrt{\mathcal{J}(\theta)}, \tag{3.12}$$

with equality for an efficient estimator, and with d' properly normalized with respect to the bias:

$$d'^2 = \frac{\left(\delta\theta \frac{dm}{d\theta}\right)^2}{\sigma_\theta^2}. \tag{3.13}$$

4 The Case of a Single Neuron

4.1 A Continuous Neuron with Vanishing Output Noise. We consider the case of a single neuron characterized by a scalar output V with a deterministic function of the input (stimulus) θ plus some noise, with a possibly stimulus-dependent variance,

$$V = f(\theta) \; + \; z\,\sigma\sqrt{g(\theta)}, \tag{4.1}$$

where f and g are deterministic functions, and σ is a parameter giving the scale of the variance of the noise, and z is a random variable with an arbitrary (that is, not necessarily gaussian) distribution $Q(z)$ with zero mean and unit variance. We are interested in the low noise limit, $\sigma \to 0$. It is not difficult to write the Fisher information $\mathcal{J}(\theta)$ and the mutual information $I[\theta, V]$ in the limit of vanishing σ. One gets, for sufficiently regular $Q(.)$,

$$I[\theta, V] = \mathcal{H}(\Theta) + \int d\theta\rho(\theta)\frac{1}{2}\log\frac{f'^2(\theta)}{\sigma^2 g(\theta)} \; - \; \mathcal{H}(Z), \tag{4.2}$$

where $\mathcal{H}(Z)$ is the entropy of the z-distribution Q:

$$\mathcal{H}(Z) = -\int dz Q(z)\log Q(z). \tag{4.3}$$

For the Fisher information one finds

$$\mathcal{J}(\theta) = \frac{f'^2(\theta)}{\sigma^2 g(\theta)} \int dz\frac{Q'^2(z)}{Q(z)}, \tag{4.4}$$

so that

$$I_{Fisher}[\theta, V] = \mathcal{H}(\Theta) + \int d\theta\rho(\theta)\frac{1}{2}\log\frac{f'^2(\theta)}{\sigma^2 g(\theta)} \; + \; \frac{1}{2}\log\int dz\frac{Q'^2(z)}{Q(z)}. \tag{4.5}$$

If the noise distribution Q is the normal distribution, one has $\mathcal{H}(Z) = \frac{1}{2}\log 2\pi e$, and the integral in equation 4.4 is equal to 1, so that one has $I = I_{Fisher}$. Otherwise one can easily check that $I > I_{Fisher}$, in agreement with the general result (see equation 3.8).

4.2 Optimization of the Transfer Function. The maximization of the mutual information with respect to the choice of the transfer function f has been studied in the case of a stimulus-independent additive noise, that is, $g \equiv 1$, by Laughlin (1981) and Nadal and Parga (1994). The expression for the mutual information, equation 4.2, with $g = 1$, has been computed by Nadal and Parga (1994). What is new here is the link with the Fisher information.

The mutual information is maximized when f is chosen according to the "equalization rule," that is, when the (absolute value of) the derivative of f is equal to the p.d.f. ρ: the activity V is then uniformly distributed between its min and max values. In the more general case in which g depends on the stimulus, the maximum of I is reached when $\hat{f}$, defined by

$$\hat{f}' \equiv f'/\sqrt{g},$$

satisfies the equalization rule,

$$\hat{f} = A \int^{\theta} dx\rho(x) \ + \ B, \tag{4.6}$$

where A and B are arbitrary given parameters (for $g = 1$, they define the min and max values of f). An interesting case is $g = f$, which is relevant for the analysis of a Poisson neuron in the large time limit (see the next subsection). In this case $f'/\sqrt{g} = 2\sqrt{f}'$, and the maximum of I is reached when the square root of f satisfies the equalization rule.

The fact that the mutual information is related to the Fisher information in the case of a single neuron with vanishing noise means that maximizing information transfer is identical to minimizing the variance of reconstruction error. In fact, two different qualitative lines of reasoning were known to lead to the equalization rule: one related to information transfer (the output V should have a uniform distribution; see, e.g., Laughlin, 1981) and one related to reconstruction error. (The slope of the transfer function should be as large as possible in order to minimize this error, and this, with the constraint that f is bounded, leads to the compromise $|f'| = \rho$. A large error can be tolerated for rare events.) We have shown here the formal link between these two approaches, using the link between mutual and Fisher information.

4.3 A Poisson Neuron. A related case is the one of a single neuron emitting spikes according to a Poisson process (in the next section we will consider a population of such neurons). The probability for observing k spikes in the interval $[0, t]$ while the stimulus θ is perceived, is

$$p(k|\theta) = \frac{(\nu(\theta)t)^k}{k!} \exp(-\nu(\theta)t), \tag{4.7}$$

where the frequency ν is assumed to be a deterministic function $\nu(\theta)$ (the tuning curve) of the stimulus θ:

$$\theta \ \to \ \nu = \nu(\theta). \tag{4.8}$$

If the stimulus is drawn randomly from a distribution $\rho(\theta)$, the frequency distribution $\mathcal{P}(\nu)$ is given by

$$\mathcal{P}(\nu) = \int d\theta\rho(\theta)\ \delta(\ \nu \ - \ \nu(\theta)). \tag{4.9}$$

The information-processing ability of such model neuron has been studied in great detail by Stein (1967). The results of interest here are as follows.

At short times, the mutual information between the stimulus and the cell activity is, at first order in t (Stein 1967),

$$I(t) \sim t \int d\nu \mathcal{P}(\nu)\nu \;\log \frac{\nu}{\mu} \equiv I_1(t), \tag{4.10}$$

where μ is the mean frequency. One can easily check that $I_1(t) \geq I(t)$ for any duration t. In fact at long times, information increases only as $\log t$: in the large time limit, one gets (Stein 1967)

$$I(t) \;=\; \int d\nu \mathcal{P}(\nu) \log \left(\mathcal{P}(\nu) \sqrt{\frac{2\pi e \nu}{t}} \right). \tag{4.11}$$

From this expression, one gets that the optimal tuning curve is such that $\sqrt{\nu}$ is uniformly distributed between its extreme values $\nu_{\min}$ and $\nu_{\max}$. We can now analyze this result in view of the relationship between Fisher and mutual information. Making the change of variable $\nu \to \theta$, with

$$\rho(\theta) d\theta = \mathcal{P}(\nu) d\nu,$$

together with equation 4.8, one can rewrite the mutual information at large times precisely as

$$I(t) \;=\; I_{Fisher}, \tag{4.12}$$

where I_{Fisher} is defined as in equation 3.6 with $\mathcal{J}(\theta)$ the Fisher information associated with this single neuron:

$$\mathcal{J}(\theta) = t \, \frac{\nu^{'2}(\theta)}{\nu(\theta)}. \tag{4.13}$$

This result can be understood in the following way. In the limit of large t, the distribution of the number of emitted spikes divided by t, $V \equiv k/t$ tends to be a gaussian, with mean $\nu(\theta)$ and variance $\nu(\theta)/t$, so that the properties of the spiking neuron become similar to those of a neuron having a continuous activity V, given by

$$\theta \to V = \nu(\theta) \; + \; z \sqrt{\nu(\theta)/t},$$

where z is a gaussian random variable with zero mean and unit variance. This is a particular case of equation 4.1, with $\sigma = 1/\sqrt{t}$, $f(.) = g(.) = \nu(.)$.

5 Population of Direction-Selective Spiking Neurons

5.1 Fisher Information. We now illustrate the main statement of section 3 in the context of population coding. We consider a large number N of neurons coding for a scalar stimulus, (e.g., an angle). Equation 3.9 tells us that to compute the mutual information, we first have to calculate the Fisher information.

When the activities $\{x_i\}$ of the neurons given θ are independent, $P(\vec{x}|\theta) = \Pi p_i(x_i|\theta)$, the Fisher information can be written

$$\mathcal{J}(\theta) = \sum_{i=1}^{N} \left\langle \frac{1}{p_i^2(x_i|\theta)} \left(\frac{\partial p_i(x_i|\theta)}{\partial \theta} \right)^2 \right\rangle_{i,\theta}, \tag{5.1}$$

where $\langle . \rangle_{i,\theta}$ is the integration over x_i with the p.d.f. $p_i(x_i|\theta)$.

We restrict ourselves to the case of neurons firing as a Poisson process with rate $\nu_i(\theta)$ in response to a stimulus $\theta \in [-\pi, \pi]$. $\nu_i(\theta)$ therefore represent the tuning curve of neuron i. We make the following assumptions: $\nu_i(\theta)$ has a single maximum at the preferred stimulus θ_i; the tuning curve depends on only the distance between the current stimulus and the preferred one and is a periodic function of this distance,

$$\nu_i(\theta) = \phi(\theta - \theta_i), \tag{5.2}$$

through the same function ϕ. The locations of the preferred stimuli of the neurons are independently and identically distributed (i.i.d.) variables in the interval $\theta \in [-\pi, \pi]$ with density $r(\theta)$.

Since our model neurons fire as a Poisson process, the information contained in their spike trains in an interval of duration t is fully contained in the number of spikes x_i emitted by each neuron in this interval. For a Poisson process we have the law

$$p_i(x_i|\theta) = \frac{(\nu_i(\theta)t)^{x_i}}{x_i!} \exp(-\nu_i(\theta)t). \tag{5.3}$$

From equations 5.1 and 5.3 we can easily calculate the Fisher information:

$$\mathcal{J}(\theta) = t \sum_{i=1}^{N} \frac{\nu_i'(\theta)^2}{\nu_i(\theta)}.$$

For N large we can replace the sum by the average over the distribution of preferred stimuli, that is,

$$\overline{\mathcal{J}(\theta)} = tN \int_{-\pi}^{\pi} d\theta' r(\theta') \frac{\phi'(\theta - \theta')^2}{\phi(\theta - \theta')}.$$

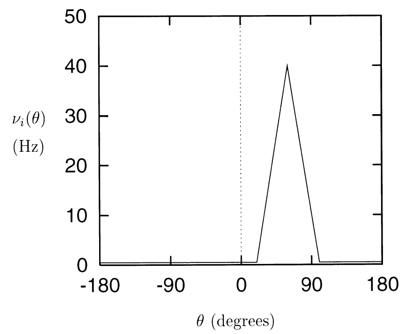

Figure 1: Triangular tuning curve corresponding to a minimal frequency $\nu_{\min} =$ 0.5 Hz, $\nu_{\max} = 40$ Hz, a receptive field half-width $a = 40$ degrees, a preferred angle $\theta_i = 60$ degrees.

For an isotropic distribution $r(\theta) = 1/(2\pi)$ we recover the result of Seung and Sompolinsky (1993).

To understand how the Fisher information depends on other parameters of the tuning curve ϕ, we redefine

$$\phi(\theta - \theta_i) = \{\nu_{\min} + (\nu_{\max} - \nu_{\min})\Phi\left(\frac{|\theta - \theta_i|}{a}\right),$$

where $\nu_{\min}$ and $\nu_{\max}$ are the minimal and maximal frequency, a is the width of the tuning curve, and Φ is a decreasing function of $|\theta - \theta_i|/a$ such that $\Phi = 1$ for the preferred stimulus $\theta = \theta_i$, and $\Phi = 0$ for stimuli far from the preferred stimulus, $|\theta - \theta_i| \gg a$. In terms of these parameters we have

$$\overline{\mathcal{J}(\theta)} = tN\frac{(\nu_{\max} - \nu_{\min})}{a}\int dzr(\theta + az)\frac{\Phi'(z)^2}{\frac{\nu_{\min}}{\nu_{\max}-\nu_{\min}} + \Phi(z)}.$$

The particular case of a triangular tuning curve,

$$\Phi(x) = \begin{cases} (1 - |x|) & x \in [-1, 1] \\ 0. & |x| > 1, \end{cases} \tag{5.4}$$

is shown in Figure 1. It will be considered in more detail below. For this tuning curve, and for a uniform distribution of preferred stimuli, the Fisher information has the simple form,

$$\overline{\mathcal{J}(\theta)} = tN\frac{(\nu_{\max} - \nu_{\min})}{\pi a}\ln\frac{\nu_{\max}}{\nu_{\min}}. \tag{5.5}$$

Thus, as already noted by Seung and Sompolinsky (1993), the Fisher information diverges in different extreme cases: when the maximal frequency $\nu_{\max}$ goes to infinity and when the tuning width a goes to zero. Moreover, functions Φ can be found such that the Fisher information diverges (e.g., $\Phi(x) = \sqrt{1-x^2}$) for any value of $\nu_{\min}$, $\nu_{\max}$, and a. Thus, the optimization of the Fisher information with respect to these parameters is an ill-defined problem without additional constraints. Note that in these cases, the equation relating the Fisher information to the mutual information is no longer valid.

There is, however, a well-defined optimization problem, which is the optimization with respect to the distribution of preferred orientations. It is considered in section 5.2. Then we show how finite size effects transform the problem of the optimization of both Fisher and mutual information with respect to the tuning width a into a well-defined problem. Last, we present some numerical estimates of these quantities, inserting some real data (Taube et al., 1990) in equation 3.9.

5.2 Optimization over the Distribution of Preferred Orientations. We ask which distribution of preferred orientations $r(\theta)$ optimizes the mutual information I. Obviously the optimal r will depend on the distribution of orientations $\rho(\theta)$. Optimizing equation 3.9 with respect to $r(\theta')$ subject to the normalization constraint $\int r(\theta')d\theta' = 1$ gives

$$\int d\theta \frac{\rho(\theta)}{\int d\theta'' r(\theta'')\psi(\theta - \theta'')}\psi(\theta - \theta') = ct \quad \text{for all } \theta',$$

in which we have defined

$$\psi(x) = \frac{\phi'(x)^2}{\phi(x)}. \tag{5.6}$$

This condition is satisfied when

$$\rho(\theta) = \frac{\int d\theta' r(\theta')\psi(\theta - \theta')}{\int d\theta' \psi(\theta')}. \tag{5.7}$$

Thus, the optimal distribution of preferred stimuli is the one that, convolved with ψ (i.e., a quantity proportional to the Fisher information), matches the distribution of stimuli. Of course in the particular case of $\rho(\theta) = 1/(2\pi)$, we obtain $r_{opt}(\theta) = 1/(2\pi)$. Note that equation 5.7 is also valid for unbounded stimulus values.

This result (equation 5.7) is specific to the optimization of the mutual information. Different results would be obtained for, say, the maximization of the average of the Fisher information or the minimization of the average of its inverse. In fact, there is no optimum for the mean Fisher information, since it is linear in $r(.)$.

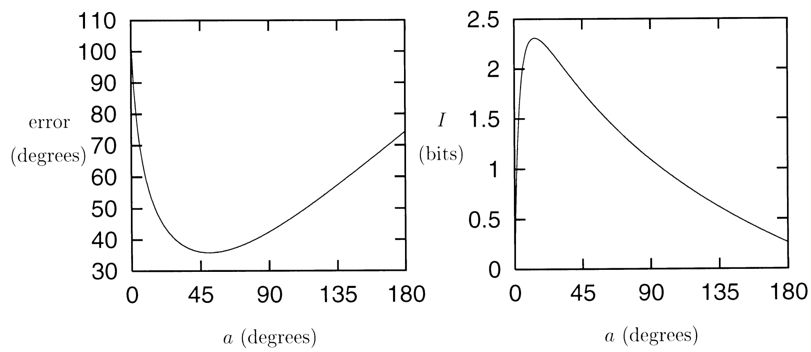

Figure 2: (Left) SD of the reconstruction error after a single spike, as a function of a. (Right) Mutual information between the spike and the stimulus as a function of a. Note that minimizing the SD of the reconstruction error is in this case different from maximizing the mutual information.

5.3 Finite Size Effects: The Case of a Single Spike. We have seen that the Fisher information, in the large N limit, diverges when the tuning width a goes to zero. To investigate whether this property is specific to the large N limit, we study the case of a finite number of neurons in a very short time interval in which a single spike has been emitted by the whole population in response to the stimulus θ. In this situation, it is clear that the optimal estimator of the stimulus (the ML estimate in that case) is given by the preferred stimulus of the neuron that emitted the spike. For finite N, the Cramer-Rao bound in general is not saturated, and we have to calculate directly the performance of the estimator. It is a simple exercise to calculate the standard deviation (SD) of the error made by such an estimate for a triangular tuning curve given in equation 5.4,

$$\mathrm{SD(error)} = \sqrt{\frac{4\pi^3 \nu_{\min} + a^3(\nu_{\max} - \nu_{\min})}{6(2\pi \nu_{\min} + a(\nu_{\max} - \nu_{\min}))}}$$

which always has a minimum for $0 < a < \pi$. We show in Figure 2 the SD of the reconstruction error after a single spike as a function of a, for $\nu_{\max}/\nu_{\min} = 80$.

It has a minimum for a about 50 degrees, for which the SD of the error is about 35 degrees.

The mutual information, on the other hand, is

$$I = \frac{1}{\pi\bar{\nu}}\left[\frac{a}{\nu_{\max}-\nu_{\min}}\left(\frac{\nu_{\max}^2}{2}\log\left(\frac{\nu_{\max}}{\bar{\nu}}\right)\right.\right.$$
$$\left.-\frac{\nu_{\min}^2}{2}\log\left(\frac{\nu_{\min}}{\bar{\nu}}\right)-\frac{1}{4}\left(\nu_{\max}^2-\nu_{\min}^2\right)\right)+$$
$$\left.+(\pi-a)\nu_{\min}\log\left(\frac{\nu_{\min}}{\bar{\nu}}\right)\right]$$

where

$$\bar{\nu} = \nu_{\min} + \frac{a}{2\pi}(\nu_{\max}-\nu_{\min})$$

It also has a maximum for positive a. The width that maximizes I is different from the width that minimizes the SD of the reconstruction error, as shown in Figure 2. This is the case in general for nongaussian tuning curves. In this case, the half-width maximizing the mutual information is around 20 degrees. Note that in a wide range of a, the first spike brings about 2 bits of information about the stimulus.

Thus, a finite optimal a stems from the constraint of already minimizing the error when only a small number of spikes have been emitted by the whole neuronal array. It implies that the largest receptive fields are most useful at very short times when only a rough estimate is possible, while smaller receptive fields will be most useful at larger times, when a more accurate estimate can be obtained.

5.4 Application to the Analysis of Empirical Data. In this section we use the experimental data of Taube et al. (1990) to show how equation 3.9 can be used to estimate both Fisher and mutual information conveyed by large populations of neurons on an angular stimulus (in this case the head direction of a rat). Taube et al. (1990) have shown that in the postsubiculum of rats, tuning curves can be well fitted by triangular tuning curves and that the distribution of preferred orientations is consistent with a uniform distribution. They also determined the distribution of the parameters of the tuning curve, $\nu_{\max}$, a and the signal-to-noise ratio (SNR) $\alpha = \nu_{\max}/\nu_{\min}$ over the recorded neurons. These data indicate that these parameters have an important variability from neuron to neuron. Equation 5.5, in the case of such inhomogeneities, has to be replaced by

$$\overline{\mathcal{J}(\theta)} = \frac{tN}{\pi}\int d\nu_{\max}\,da\,d\alpha\,\Pr(\nu_{\max}, a, \alpha)\frac{\nu_{\max}}{a}\left(1-\frac{1}{\alpha}\right)\ln\alpha. \tag{5.8}$$

in which $\Pr(\nu_{\max}, a, \alpha)$ is the joint probability of parameters $\nu_{\max}$, a and α.

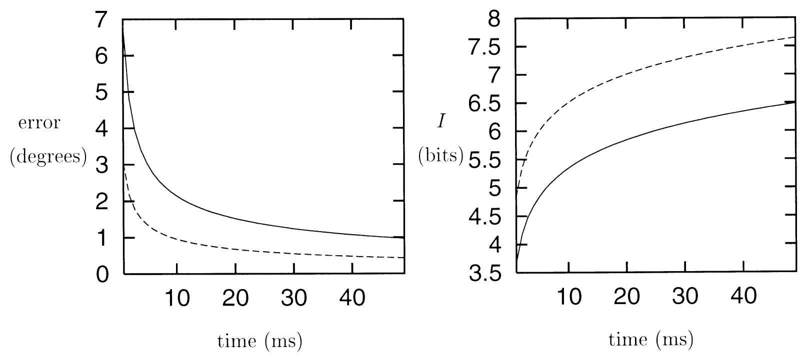

Figure 3: (Left) Minimal reconstruction error as given by the Cramer-Rao bound for $N = 1000$ (full curve), $N = 5000$ (dashed curve) postsubiculum neurons, using data from Taube et al. (1990) as a function of time. (Right) Mutual information for $N = 1000$ (full curve), $N = 5000$ (dashed curve), using the same data and equation 3.10.

Under global constraints, one may expect each neuron to contribute in the same way to the information, that is, $(\nu_{\max}/a)(1 - 1/\alpha)\ln\alpha$ is constant. This would imply that the width a increases with $\nu_{\max}$. Figure 9 of Taube et al. (1990) shows that there is indeed a trend for higher firing rate cells to have wider directional firing ranges.

We can now insert the distributions of parameters measured in Taube et al. (1990) in equation 5.8 to estimate the minimal reconstruction error that can be done on the head direction using the output of N postsubiculum neurons during an interval of duration t. It is shown in the left part of Figure 3. Since we assume that the number of neurons is large, the mutual information conveyed by this population can be estimated using equation 3.9. It is shown in the right part of the same figure. In the case of $N = 5000$ neurons, the error is as small as one degree even at $t = 10$ ms, an interval during which only a small proportion of selective neurons has emitted a spike. Note that one degree is the order of magnitude of the error made typically in perceptual discrimination tasks (see, e.g., Pouget & Thorpe 1991). During the same interval, the activity of the population of neurons carries about 6.5 bits about the stimulus. Doubling the number of neurons or the duration of the interval divides the minimal reconstruction error by $\sqrt{2}$ and increases the mutual information by 0.5 bit.

6 Conclusion

In this article we have exhibited the link between Fisher information and mutual information in the context of neural coding. This link was first derived in the context of Bayesian parameter estimation by Clarke and Barron (1990) and then in the context of stochastic complexity by Rissanen (1996). We have shown that the result of Rissanen applies to population coding—that is, when the number of neurons is very large compared to the dimension of the stimulus. Our derivation of the link uses completely different techniques. The result is that the mutual information between the neural activities and the stimulus is equal to the one between the stimulus and an ideal gaussian unbiased estimator whose variance is equal to the inverse of the Fisher information. The result is true not only for independent observations, but also for correlated activities (see Rissanen, 1996, and the appendix). This is important in the context of neural coding since noise in different cells might in some cases be correlated due to common inputs or to lateral connections.

This result implies that in the limit of a large number of neurons, maximization of the mutual information leads to optimal performance in the estimation of the stimulus. We have thus considered the problem of optimizing the tuning curves by maximizing the mutual information over the parameters defining the tuning curves: optimization of the choice of preferred orientations, widths of the tuning curves. In the simple model we have considered, the optimal value for the width is zero, as in Seung and Sompolinsky (1993). However, we have shown that finite size effects necessarily lead to a nonzero optimal value, independent of the decoding scheme.

We have discussed in detail the case of a one-dimensional stimulus (an angle). A similar relationship between mutual information and the Fisher information matrix holds for any dimensionality of the stimulus, as long as it remains small compared to the number of neurons. It would be straightforward to consider in the more general case the optimization of the tuning curves. Zhang, Ginzburg, McNaughton, and Sejnowski (1998) have computed the Fisher information matrix for two- and three- dimensional stimuli. Their results imply that optimal tuning curve parameters will depend strongly on the dimensionality of the stimulus.

We have briefly discussed the cases of a finite number of neurons and the short time limit. In this case maximization of the mutual information leads in general to different results than does minimization of the variance of reconstruction error, as found also in networks with the same number of input and output continuous neurons (Ruderman, 1994). We are currently working on these limits for which many aspects remain to be clarified.

We have not addressed the problem of decoding. In the asymptotic limit, the maximum likelihood (ML) decoding is optimal. Recently Pouget and Zhang (1997) showed that a simple recurrent network can perform the computation of the ML estimate. This suggests that the optimal performance, from the point of view of both information content and decoding, can be reached by a simple cortical architecture.

Appendix

Our goal is to derive equation 3.10, that is, to compute the mutual information $I = I[P, \rho]$ between the random variables $\vec{x}$ and θ, working in the large N limit. We recall that $\vec{x}$ can be seen as either a set of N observations related to the measurement of an unknown parameter θ or the set of responses of N neurons to a stimulus θ. The mutual information I is defined by

$$I = \int d\theta \rho(\theta) \left\langle \ln \frac{P(\vec{x}|\theta)}{Q(\vec{x})} \right\rangle_\theta , \tag{A.1}$$

where $Q(\vec{x})$ is the p.d.f. of $\vec{x}$:

$$Q(\vec{x}) = \int d\theta \rho(\theta) P(\vec{x}|\theta). \tag{A.2}$$

In equation A.1, $\langle\ .\ \rangle_\theta$ denotes the integration over $\vec{x}$ *given* θ with the p.d.f. $P(\vec{x}|\theta)$. We define

$$G(\vec{x}|\theta) \equiv \frac{1}{N} \ln P(\vec{x}|\theta). \tag{A.3}$$

We will make the following hypothesis:

1. All derivatives of G with respect to the stimulus θ are of order 1 in the large N limit.
2. The cumulants of order n of $xG'_\theta + yG''_\theta$ are of order $1/N^{n-1}$ in the large N limit.

Both properties are verified for the factorized models (see equations 2.1 and 2.2), but also in some cases in which x_i given θ are correlated variables, as we show at the end of the appendix.

The large N limit allows us to use the saddle-point method (Bhattacharya & Rao, 1976; Parisi, 1988) for the computation of integrals over θ, in particular for the computation of the p.d.f. $Q(\vec{x})$, using the fact that $P(\vec{x}|\theta)$ will appear to be sharply peaked around its most probable value, the maximum likelihood (ML) estimator of θ. We will use standard cumulant expansions for the integration over $\vec{x}$ in the equivocation part of I, and this will eventually lead to the announced result, equation 3.10.

Distribution of $\vec{x}$. The p.d.f. $Q(\vec{x})$ can be written

$$Q(\vec{x}) = \int d\theta \rho(\theta) \exp NG(\vec{x}|\theta). \tag{A.4}$$

For large N, the integral is dominated by the maxima of the integrand. These are defined by the solutions of

$$G'_\theta(\vec{x}|\theta) = 0, \tag{A.5}$$

which satisfy $G''_\theta(\vec{x}|\theta) < 0$. Above we have denoted by G'_θ (resp. G''_θ) the first (resp. second) partial derivative of G with respect to θ. Let us assume that $G(\vec{x}|\theta)$ has a single global maximum at $\theta_m(\vec{x})$. The Taylor expansion around $\theta_m(\vec{x})$ is

$$G(\vec{x}|\theta) = G(\vec{x}|\theta_m(\vec{x})) + \frac{1}{2}G''_\theta(\vec{x}|\theta_m(\vec{x}))(\theta - \theta_m(\vec{x}))^2 + \ldots$$

Using standard saddle-point techniques we find,

$$Q(\vec{x}) = Q_m(\vec{x})\left(1 + O\left(\frac{1}{N}\right)\right), \tag{A.6}$$

with

$$Q_m(\vec{x}) \equiv \rho_m(\vec{x})\sqrt{\frac{2\pi}{N|\Gamma(\vec{x})|}}, \exp\left[NG_m(\vec{x})\right], \tag{A.7}$$

where

$$\rho_m(\vec{x}) \equiv \rho(\theta_m(\vec{x})), \tag{A.8}$$

$$G_m(\vec{x}) \equiv G(\vec{x}|\theta_m(\vec{x})), \tag{A.9}$$

and

$$\Gamma(\vec{x}) \equiv G''_\theta(\vec{x}|\theta_m(\vec{x})). \tag{A.10}$$

Note that $\theta_m(\vec{x})$ is the ML estimator of θ.

The Mutual Information: Integration over θ. Let us start with the following expression of the mutual information:

$$I = -\int d\theta\rho(\theta)\ln\rho(\theta) + \int d^N x\, Q(\vec{x}) \int d\theta\; Q(\theta|\vec{x})\ln Q(\theta|\vec{x}),$$

with $Q(\theta|\vec{x}) = \frac{P(\vec{x}|\theta)\rho(\theta)}{Q(\vec{x})}$. The first term is the entropy of the input distribution. The second term can be written

$$-\int d^N x\, Q(\vec{x})\ln Q(\vec{x}) + \int d^N x \int d\theta\; P(\vec{x}|\theta)\rho(\theta)\ln P(\vec{x}|\theta)\rho(\theta). \tag{A.11}$$

In the above expression, the first part is the entropy of $\vec{x}$ in which we can replace $Q(\vec{x})$ by $Q_m(\vec{x})$ as given in equation A.7, leading to

$$-\int d^N x\, Q_m(\vec{x}) \left[NG_m + \ln \rho_m - \frac{1}{2} \ln \frac{N|\Gamma(\vec{x})|}{2\pi} \right].$$

The last term in equation A.11 can be written as

$$\int d^N x \int d\theta\, A(\vec{x}|\theta)\, \exp A(\vec{x}|\theta),$$

with

$$A(\vec{x}|\theta) \equiv \ln P(\vec{x}|\theta)\rho(\theta).$$

Now

$$\int d\theta A(\vec{x}|\theta)\, \exp A(\vec{x}|\theta) = \partial_\lambda \int d\theta \exp \lambda A|_{\lambda=1},$$

which is again computed with the saddle-point method,

$$\begin{aligned} \int d\theta A(\vec{x}|\theta)\, \exp A(\vec{x}|\theta) &= \partial_\lambda \sqrt{\frac{2\pi}{\lambda N |\Gamma(\vec{x})|}} \exp \lambda \left[NG_m + \ln \rho_m \right] \Bigg|_{\lambda=1} \\ &= Q_m \left[NG_m + \ln \rho_m - \frac{1}{2} \right]. \end{aligned}$$

Finally, putting everything together, the mutual information can be written as

$$\begin{aligned} I = &-\int d\theta \rho(\theta) \ln \rho(\theta) \\ &+ \int d^N x\, \rho(\theta_m(\vec{x})) \sqrt{\frac{2\pi}{N|\Gamma(\vec{x})|}} \exp\left[NG_m(\vec{x})\right] \left(\frac{1}{2} \ln \frac{N|\Gamma(\vec{x})|}{2\pi e} \right). \end{aligned} \quad \text{(A.12)}$$

It is interesting to compare equations 3.10 and A.12. As in equation 3.10, the first term above is the entropy $\mathcal{H}[\theta] = -\int d\theta \rho(\theta) \ln \rho(\theta)$ of the stimulus distribution; the second term, the equivocation, is given in equation A.12 by the average over the p.d.f. of $\vec{u}$ of the logarithm of the variance of the estimator.

The Mutual Information: Integration over $\vec{x}$. The last difficulty is to perform in equation A.12 the trace on $\vec{x}$. One cannot apply the saddle-point method directly because the number of integration variables is precisely equal to the number N that makes the exponential large. However, the difficulty is circumvented by the introduction of a small (compared to N) auxiliary integration variables, in such a way that the integration over the

x_i's can be done exactly. Then we again use the fact that N is large to perform the integration over the auxilary variables to leading order in N.

First we use the relation

$$F(\theta_m(\vec{x})) = \int d\theta F(\theta) |G''_\theta(\vec{x}|\theta)| \delta\left(G'_\theta(\vec{x}|\theta)\right)$$

in order to deal with $\theta_m(\vec{x})$, which is valid for an arbitrary function F. We then use an integral representation of the delta function:

$$\delta\left(G'_\theta(\vec{x}|\theta)\right) = \int \frac{dy}{2\pi} \exp\left(iyG'_\theta(\vec{x}|\theta)\right).$$

Similarly, in order to deal with $G''_\theta(\vec{x}|\theta)$, we introduce conjugate variables τ, $\hat{\tau}$. For any function F we can write

$$F(G''_\theta(\vec{x}|\theta)) = \int d\tau d\hat{\tau} \frac{1}{2\pi} F(\tau) \exp\left(i\hat{\tau}(\tau - G''_\theta(\vec{x}|\theta))\right).$$

Putting everything together, we get

$$I = \mathcal{H}[\theta] + \int d\theta dy d\tau d\hat{\tau} \frac{\sqrt{|\tau|}}{\sqrt{N}(2\pi)^{\frac{3}{2}}} \rho(\theta) \times \left(\frac{1}{2}\ln\left(\frac{N|\tau|}{2\pi e}\right)\right) \exp\left(i\hat{\tau}\tau + K(\theta, y, \hat{\tau})\right), \tag{A.13}$$

in which

$$K(\theta, y, \hat{\tau}) = \ln \left\langle \exp\left(-i\hat{\tau}\frac{\partial^2 G(\vec{x}|\theta)}{\partial\theta^2} + iy\frac{\partial G(\vec{x}|\theta)}{\partial\theta}\right)\right\rangle_\theta$$

(recall that $\langle \ldots \rangle_\theta = \int d^N x \exp[NG(\vec{x}|\theta)] \ldots$). We now make the cumulant expansion

$$\left\langle \exp A \right\rangle_\theta = \exp\left(\langle A \rangle_\theta + \frac{1}{2}(\left\langle A^2 \right\rangle_\theta - \langle A \rangle_\theta^2) + \cdots\right)$$

for

$$A \equiv -i\hat{\tau}G''_\theta + iyG'_\theta. \tag{A.14}$$

The cumulant expansion will be valid if the cumulants of order n of A with the law $\exp[NG(\vec{x}|\theta)]$ decrease sufficiently rapidly with n. A sufficient condition is

assumption: the cumulants of order n of A $(n = 1, 2, \ldots)$ *are of order* $1/N^{n-1}$. (A.15)

Using the following identities obtained by deriving twice $1 = \langle 1 \rangle_\theta$ with respect to θ,

$$0 = \left\langle G'_\theta \right\rangle_\theta$$
$$0 = \left\langle G''_\theta \right\rangle_\theta + N \left\langle \left(G'_\theta\right)^2 \right\rangle_\theta ,$$

one gets

$$K \; = \; i\hat{\tau} J - \frac{\hat{\tau}^2 \Delta^2}{2N} - \frac{y^2 J}{2N} + \frac{y \hat{\tau} Z}{N} + O\left(\frac{1}{N^2}\right) \tag{A.16}$$

where J, Δ, Z are given by

$$J \equiv - \left\langle G''_\theta \right\rangle_\theta = N \left\langle \left(G'_\theta\right)^2 \right\rangle_\theta$$
$$\Delta^2 \equiv N \left(\left\langle \left(G''_\theta\right)^2 \right\rangle_\theta - \left\langle G''_\theta \right\rangle_\theta^2 \right)$$
$$Z \equiv N \left\langle G'_\theta G''_\theta \right\rangle_\theta .$$

Note that the Fisher information $\mathcal{J}(\theta)$ is equal to $N\,J$, and that Δ^2 and Z are of order 1 because of the assumption A.15.

In these terms we have

$$I = \mathcal{H}[\theta] + \int d\theta dy d\tau d\hat{\tau} \frac{\sqrt{|\tau|}}{\sqrt{N} (2\pi)^{\frac{3}{2}}} \rho(\theta) \left(\frac{1}{2} \ln \left(\frac{N |\tau|}{2\pi e} \right) \right)$$
$$\exp\left(i\hat{\tau}(\tau + J) - \frac{\hat{\tau}^2 \Delta^2}{2N} - \frac{y^2 J}{2N} + \frac{y\hat{\tau} Z}{N} + O\left(\frac{1}{N^2}\right) \right).$$

Our last task is to integrate over the remaining auxiliary variables τ, $\hat{\tau}$, y. Using the fact that $\Delta^2 - \frac{Z^2}{J} > 0$, deduced from the Schwartz inequality,

$$< G'_\theta \; (G''_\theta - < G''_\theta >) >^2 \; \leq \; < G'^2_\theta >< (G''_\theta - < G''_\theta >)^2 >,$$

the integrations over y and $\hat{\tau}$ are simple gaussian integrations, leading to:

$$I = \mathcal{H}[\theta] + \int d\theta \rho(\theta)$$
$$\times \int \frac{d\tau}{\sqrt{2\pi}} \sqrt{\frac{N}{\Delta^2 - \frac{Z^2}{J}}} \sqrt{\frac{|\tau|}{J}} \frac{1}{2} \ln\left(\frac{N|\tau|}{2\pi e} \right) \exp\left(- \frac{N}{2} \frac{(\tau + J)^2}{\Delta^2 - \frac{Z^2}{J}} \right).$$

The integration over τ is with a gaussian weight centered at $\tau = -J$ and with a width going to zero as N goes to infinity:

$$\lim_{N \to \infty} \frac{1}{\sqrt{2\pi}} \sqrt{\frac{N}{\Delta^2 - \frac{Z^2}{J}}} \exp - \frac{N}{2} \frac{(\tau + J)^2}{\Delta^2 - \frac{Z^2}{J}} \; = \; \delta(\tau + J).$$

Using the fact that the Fisher information is $\mathcal{J}(\theta) = NJ$, we obtain

$$I = -\int d\theta \rho(\theta) \ln \rho(\theta) - \int d\theta \rho(\theta) \frac{1}{2} \ln \left(\frac{2\pi e}{\mathcal{J}(\theta)} \right) (1 + O(1/N)), \quad \text{(A.17)}$$

which is the announced result (equation 3.10).

The conditions (see A.15) of validity of the calculation are satisfied when x_i given θ are independent, as in equations 2.1 and 2.2, but can also be satisfied when they are correlated. We discuss below these two cases.

Conditional Independence of Activities. In the case of independent neurons, the model in equation 2.2, one can easily check that the cumulant expansion at order n gives terms of order $1/N^{n-1}$. Indeed, in that case, one has

$$G(\vec{x}|\theta) = \frac{1}{N} \sum_i g_i(x_i|\theta), \quad \text{(A.18)}$$

so that

$$A = \frac{1}{N} \sum_i A_i, \text{ with } A_i = -i\hat{\tau} \frac{\partial^2 g_i(x_i|\theta)}{\partial \theta^2} + iy \frac{\partial g_i(x_i|\theta)}{\partial \theta}. \quad \text{(A.19)}$$

The cumulant expansion then reads

$$\begin{aligned} \langle \exp A \rangle &= \exp \sum_i \log \left\langle \exp \frac{A_i}{N} \right\rangle \\ &= \exp \sum_i \left(\frac{1}{N} \langle A_i \rangle + \frac{1}{N^2} (\langle A_i^2 \rangle - \langle A_i \rangle^2) + O(1/N^3) \right). \end{aligned} \quad \text{(A.20)}$$

Thus equation A.16 holds, with J, Δ, Z given by

$$\begin{aligned} J &= -\frac{1}{N} \sum_i \left\langle \frac{\partial^2 g_i}{\partial \theta^2} \right\rangle_\theta = \frac{1}{N} \sum_i \left\langle \left(\frac{\partial g_i}{\partial \theta} \right)^2 \right\rangle_\theta \\ \Delta^2 &= \frac{1}{N} \sum_i \left(\left\langle \left(\frac{\partial^2 g_i}{\partial \theta^2} \right)^2 \right\rangle_\theta - \left\langle \frac{\partial^2 g_i}{\partial \theta^2} \right\rangle_\theta^2 \right) \\ Z &= \frac{1}{N} \sum_i \left\langle \frac{\partial g_i}{\partial \theta} \frac{\partial^2 g_i}{\partial \theta^2} \right\rangle_\theta . \end{aligned} \quad \text{(A.21)}$$

Correlated Neurons. The conditions on the cumulants of A, A.15, do not imply that the x_i are independent, but they do have the qualitative meaning that they convey of order N independent observations. To see this, we give an example of correlated activities for which the conditions are satisfied.

We consider the following simple model. Each x_i can be expressed in terms of the same N independent random variables, ξ_a, $a = 1, \ldots, N$, as

$$x_i = \sum_a M_{i,a}\xi_a. \tag{A.22}$$

where M is a θ-independent invertible matrix, and the ξ's are, given θ, statistically independent variables of arbitrary p.d.f. $\rho_a(\xi|\theta)$, $a = 1, \ldots, N$. The factorized case is recovered for M diagonal. In the case where the ρ's are gaussian and M is orthogonal, equation A.22 is the principal component decomposition of the x's. We show now that the case M invertible with arbitrary ρ's satisfies the conditions A.15.

First, it is obvious that the result (see equation 3.10) holds: with the change of variables $\vec{x} \to M^{-1}\vec{x} = \vec{\xi}$, one recovers the case of independent (given θ) activities. One can then apply equation 3.10 to $I(\theta, \vec{\xi})$. Since $P(\vec{\xi}|\theta) = P(\vec{x}|\theta)|\det M|$, with M independent of θ, $I(\theta, \vec{x}) = I(\theta, \vec{\xi})$ and the Fisher information associated to $P(\vec{\xi}|\theta)$ is equal to the one associated to $P(\vec{x}|\theta)$, so that equation 3.10 holds for $I(\theta, \vec{x})$. Second, one can check directly that the conditions A.15 hold. For our model, G is

$$G(\vec{x}|\theta) = -\frac{1}{N}\ln|\det M| \; + \; \frac{1}{N}\sum_a \ln\rho_a\left(\sum_i M^{-1}_{a,i}x_i\big|\theta\right), \tag{A.23}$$

so that the cumulants of $G'_\theta(\vec{x}|\theta)$ and $G''_\theta(\vec{x}|\theta)$ with respect to the pdf $P(\vec{x}|\theta)$ are equal to the cumulants of $G'_\theta(\vec{\xi}|\theta)$ and $G''_\theta(\vec{\xi}|\theta)$ with respect to the factorized pdf $P(\vec{\xi}|\theta) = \prod_a \rho_a(\xi|\theta)$ for which A.15 holds.

Acknowledgments

We thank Alexandre Pouget and Sophie Deneve for an interesting discussion, and Sid Wiener for drawing the data of Taube et al. (1990) to our attention. We are grateful to Alexandre Pouget and Peter Latham for pointing out a mistake in an earlier version of the article, and to the referees for comments that helped us to improve the article significantly.

References

Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing? *Network*, *3*, 213–251.

Barlow, H. B., Kaushal, T. P., & Mitchison G. J. (1989). Finding minimum entropy codes. *Neural Comp.*, *1*, 412–423.

Bhattacharya, R. N., & Rao, R. R. (1976). *Normal approximation and asymptotic expansions*. New York: Wiley.

Bialek, W., Rieke, F., de Ruyter van Steveninck, R., & Warland, D. (1991). Reading a neural code. *Science*, *252*, 1854–1857.

Blahut, R.E. (1988). *Principles and practice of information theory*. Reading, MA: Addison-Wesley.

Clarke, B. S., & Barron, A. R. (1990) Information theoretic asymptotics of Bayes methods. *IEEE Trans. on Information Theory*, *36*, 453–471.

Cover, T. M., & Thomas, J. A. (1991). *Information theory*. New York: Wiley.

Georgopoulos, A. P., Kalaska, J. F., Caminiti, R., & Massey, J. T. (1982). On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *J. Neurosci.*, *2*, 1527–1537.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Laughlin, S. B. (1981). A simple coding procedure enhances a neuron's information capacity. *Z. Naturf.*, *C36*, 910–912.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, *21*, 105–117.

Maunsell, J. H. R., & Van Essen, D. C. (1983). Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation. *J. Neurophysiol.*, *49*, 1127–1147.

Nadal, J.-P., & Parga, N. (1994). Nonlinear neurons in the low noise limit: A factorial code maximizes information transfer. *Network*, *5*, 565–581.

Parisi, G. (1988). *Statistical field theory*, Reading, MA: Addison-Wesley.

Pouget, A. & Thorpe, S. J. (1991). Connexionist models of orientation identification. *Connection Science*, *3*, 127–142.

Pouget, A. & Zhang, K. (1997). Statistically efficient estimations using cortical lateral connections. In M. C. Moza, M. I. Jordan, & T. Petsche (Eds.), *Advances in neural information processing systems*, 9 (pp. 97–103). Cambridge, MA: MIT press.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Trans. on Information Theory*, *42*, 40–47.

Ruderman, D. (1994). Designing receptive fields for highest fidelity. *Network*, *5*, 147–155.

Seung, H. S., & Sompolinsky, H. (1993). Simple models for reading neural population codes. *P.N.A.S. USA*, *90*, 10749–10753.

Shannon, S. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Snippe, H. P. (1996). Parameter extraction from population codes: A critical assesment. *Neural Comp.*, *8*, 511–529.

Softky, W. R., & Koch, C. (1993). The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *J. Neurosci.*, *13*, 334.

Stein, R. (1967). The information capacity of nerve cells using a frequency code. *Biophys. J.*, *7*, 797–826.

Taube, J. S. (1995). Head direction cells recorded in the anterior thalamic nuclei of freely moving rats. *J. Neurosci.*, *15*, 70–86.

Taube, J. S., Muller, R. U., & Ranck, J. B. (1990). Head direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *J. Neurosci.*, *10*, 420–435.

Van Hateren, J. H. (1992). Theoretical predictions of spatiotemporal receptive fields of fly LMCS, and experimental validation. *J. Comp. Physiology A*, *171*, 157–170.

Zhang, K., Ginzburg, I., McNaughton, B. L., & Sejnowski, T. J. (1998). Interpreting neuronal population activity by reconstruction: A unified framework with application to hippocampal place cells *J. Neurophysiol.*, *79*, 1017–1044.

Received November 3, 1997; accepted February 20, 1998.