LETTER Communicated by Kechen Zhang

Optimal Short-Term Population Coding: When Fisher Information Fails

M. Bethge
mbethge@physik.uni-bremen.de
D. Rotermund
davrot@physik.uni-bremen.de
K. Pawelzik
pawelzik@physik.uni-bremen.de
Institute of Theoretical Physics, University of Bremen, Bremen, D-28334 Germany

Efficient coding has been proposed as a first principle explaining neuronal response properties in the central nervous system. The shape of optimal codes, however, strongly depends on the natural limitations of the particular physical system. Here we investigate how optimal neuronal encoding strategies are influenced by the finite number of neurons N (place constraint), the limited decoding time window length T (time constraint), the maximum neuronal firing rate $f_{\max}$ (power constraint), and the maximal average rate $\langle f \rangle_{\max}$ (energy constraint). While Fisher information provides a general lower bound for the mean squared error of unbiased signal reconstruction, its use to characterize the coding precision is limited. Analyzing simple examples, we illustrate some typical pitfalls and thereby show that Fisher information provides a valid measure for the precision of a code only if the dynamic range $(f_{\min}T, f_{\max}T)$ is sufficiently large. In particular, we demonstrate that the optimal width of gaussian tuning curves depends on the available decoding time T. Within the broader class of unimodal tuning functions, it turns out that the shape of a Fisher-optimal coding scheme is not unique. We solve this ambiguity by taking the minimum mean square error into account, which leads to flat tuning curves. The tuning width, however, remains to be determined by energy constraints rather than by the principle of efficient coding.

1 Introduction

Since Attneave (1954) and Barlow (1959) proposed redundancy reduction or coding efficiency as a major principle for neuronal representations governing the formation of the central nervous system, much theoretical and experimental work has attempted to identify this principle within various structures in the brain. Neurons in cortical circuits are typically coupled to many other neurons ($10^3 - 10^4$) (Braitenberg & Schüz, 1991), and their activity is propagated and converted in a highly distributed manner. This

raises the question of how signal processing can in principle be performed by large populations of neurons. Technically, this question corresponds to characterizing the class of computations (i.e., filters or functions) that can be realized by the combination of elementary units (model neurons). Most theoretical studies, however, deal only with noise-free units, which cannot account for the fact that the reliable communication of a signal (that is, the realization of the identity, which may be seen as the simplest case of noisy computation) may require a large number of neurons.

Here, we address the issue of optimal signal representation in populations of neurons. Population codes (Rolls & Cowey, 1970; Georgopoulos, Schwartz, & Kettner, 1986; Paradiso, 1988) constitute an important class of neural coding schemes, in which the signal is represented by the number of spikes emitted from each of the neurons within a given counting time window. In other words, a population code is completely determined by the spike counting statistics and may be considered as a simplified snapshot description of the time-continuous neuronal filtering process of temporal integration over synaptic inputs.

In this article, we seek to identify optimal encoding strategies (i.e., a set of tuning functions) under the assumption of a Poisson noise model, given a limited number of neurons N and a finite decoding time window of length T. Similar to previous studies by Panzeri et al. (Panzeri, Bielle, Rolls, Skaggs, & Treves, 1996; Panzeri, Treves, Schultz, & Rolls, 1999), we are particularly interested in short time windows, since psychophysical experiments have shown that efficient computations can be performed in cortex at a rate where each neuron has fired on average only once (Rolls & Tovee, 1994; Thorpe, Fitz, & Marlot, 1996). In order to make the optimization problem well posed, we specify both the objective function (see section 2) and the set of tuning function arrays within which the optimum is determined. With respect to the latter, we intend to make the class of candidate encoding strategies as large as possible in order to learn something about the arrangement and shape of tuning functions that correspond to the ultimate limit of precision. In particular, this means that we do not restrict the class of encoding strategies a priori to those tuning functions that are considered biologically plausible; rather, we intend to check the explanatory power of the coding efficiency principle.

Previous studies of optimal population coding have focused on whether broad or small tuning widths are advantageous for the representation of stimulus features (Hinton, McClelland, & Rumelhart, 1986; Baldi & Heiligenberg, 1988; Snippe & Koenderink, 1992). In particular, an optimal scale for populations of neurons with arbitrary radially symmetric tuning functions and arbitrary dimensions of the encoded parameter (Zhang & Sejnowski, 1999) has been determined with respect to the average Fisher information. Subsequently, this analysis has been further generalized in Eurich and Wilke (2000) by allowing for different scales for each stimulus dimension. In contrast to the apparent generality of this approach, its validity is

substantially restricted by two problems that make further analysis necessary.

First, the optimization of the scale or width was based on a comparison between tuning functions with exactly the same shape, neglecting the fact that the precision may depend much less on the scale or width than on other aspects of the shape of the tuning functions (see section 6). In contrast, here we seek to find population codes that are optimal with respect to the entire class of tuning functions specified by very basic constraints only, such as a limited maximum firing rate or unimodality.

The other problem results from the limited validity of Fisher information J (see equation 2.7) as a measure for the precision of a population code. The precision of a population code is usually defined independently from Fisher information on the basis of the conditional mean squared error (c.m.s.e.) of an ideal observer as a function of the presented stimulus (Baldi & Heiligenberg, 1988). While the Cramér-Rao bound (see equation 2.8) provides a general lower bound on the c.m.s.e. of any estimator (Aitken & Silverstone, 1942; Frechet, 1943; Rao, 1946; Cramér, 1946), this inequality does not justify the use of Fisher information as a general measure for the precision of population codes, because the Cramér-Rao bound is not unique for biased estimators, and it often cannot be attained. Furthermore, the notion of an ideal observer is not unique, so that no general definition of the resolution exists that fits all problems, but any definition necessarily depends on further specifications.

Our analysis investigates the question of optimal encoding taking these two aspects into account. Another goal of this article is to strengthen intuition in how far Fisher information can be used to judge on the precision of different encodings. In the next section, we describe the methods and notations used throughout the article. In particular, a thorough justification of Fisher information as a measure for the precision of population codes on the basis of asymptotic efficiency is presented. In section 3, we determine the optimal scale for the example of gaussian tuning curves with respect to Fisher information, on the one hand, and with respect to the minimum mean squared error (MMSE) in the case of a small counting time window, on the other hand. This example indicates that Fisher information does not account for the MMSE in the case of short-term population coding. Subsequently, we show that this problem becomes especially relevant for Fisher-optimal codes if one drops the a priori restriction to gaussian-shaped tuning curves by presenting an example in section 4, where two neurons are sufficient to achieve arbitrary large Fisher information. The conditions under which Fisher information can be used to determine the MMSE are discussed in section 5. In section 6, we investigate the case where the tuning functions are constrained to have a single maximum only (i.e., unimodal tuning functions) and the crucial role of energy constraints for the tuning width is demonstrated in section 7. Finally, we discuss the prerequisites and implications of our results in section 8.

2 Methods and Notations

We give a brief overview of the relevant quantities and methods used throughout this article. The encoded random variable will be denoted by x and the observable spike count vector by $\mathbf{k}$, whose N components are the numbers of spikes of the N neurons (see Figure 1). Because all observable quantities take values within a limited range only, we set the range of x to the open unit interval $x \in (0, 1)$ without loss of generality. For convenience, we assume x to be uniformly distributed with density $p(x) = \Theta(x)\Theta(1-x)$, where $\Theta(y)$ denotes the Heaviside function, which is one if $y > 0$ and zero otherwise. The distribution specified by $p(x)$ is also called the a priori distribution, because it determines general properties of the signal x that are independent of the observed $\mathbf{k}$.

The encoding of x is specified by the set of tuning functions $\{f_j(x)\}_{j=1}^N$ that give the mean number of spikes for each neuron j divided by the length T of the counting time window. Together with the assumption of independent Poisson noise, the response statistics of the entire population is then described by

$$p(\mathbf{k} \mid x) = \prod_{j=1}^{N} p_{\mu_j}(k_j) = \prod_{j=1}^{N} \frac{(Tf_j(x))^{k_j}}{k_j!} \exp\{-Tf_j(x)\}, \tag{2.1}$$

where $p_{\mu_j}(k_j)$ denotes the probability mass function of the Poisson distribution with parameter $\mu_j = Tf_j(x)$, which gives the mean and the variance of the spike count. Apart from the asymptotic cases $T \to 0$ and $T \to \infty$, we will frequently consider the case $f_{\max}T = 1$, that is, each neuron does not fire more than one spike on average. As we will discuss in section 8, we suspect that $f_{\max}T = 1$ is of the relevant order for signal transmission in cortex.

Throughout the article, the mean of a random variable is denoted by E[.]. In particular, we will have

$$\mathrm{E}[A(x, \mathbf{k})] = \int p(x) \sum_{\mathbf{k}} p(\mathbf{k} \mid x)\, A(x, \mathbf{k})\, dx \tag{2.2}$$

$$= \int p(x) \sum_{k_1=0}^{\infty}, \ldots, \sum_{k_N=0}^{\infty} p(\mathbf{k} \mid x)\, A(x, \mathbf{k})\, dx, \tag{2.3}$$

which reduces to $\mathrm{E}[A(x)] = \int p(x)A(x)\, dx$, if A is a function of x only. $\mathrm{E}[A \mid x] = \sum_{\mathbf{k}} p(\mathbf{k} \mid x)A(\mathbf{k}, x)$ is called the conditional mean given x. We will also consider the conditional mean $\mathrm{E}[A \mid \mathbf{k}] = \int p(x \mid \mathbf{k})A(\mathbf{k}, x)\, dx$, where the so-called a posteriori distribution $p(x \mid \mathbf{k})$ can be obtained by using the Bayes formula:

$$p(x \mid \mathbf{k}) = \frac{p(\mathbf{k} \mid x)p(x)}{\int p(\mathbf{k} \mid \tilde{x})p(\tilde{x})d\tilde{x}}. \tag{2.4}$$

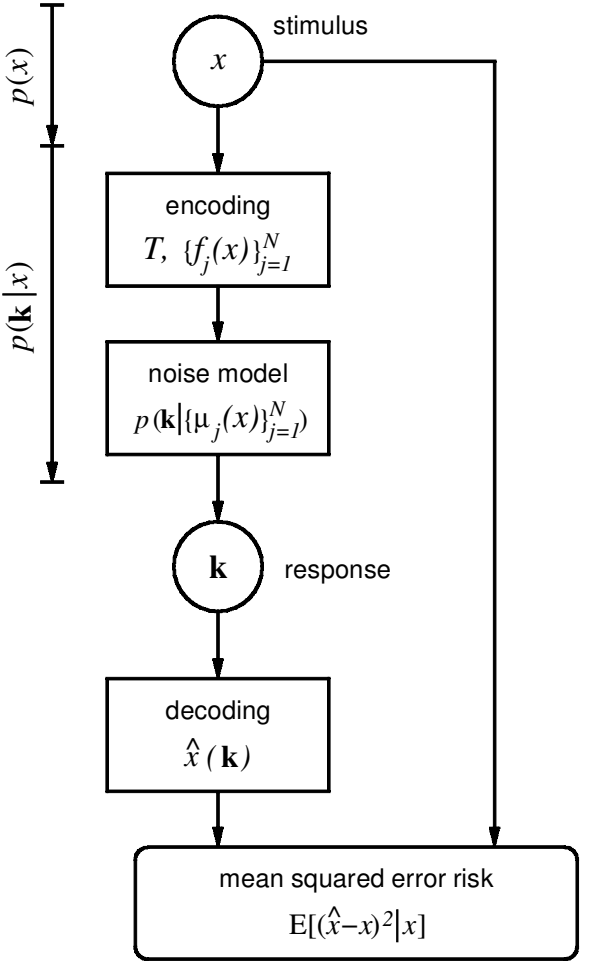

Figure 1: General scheme of encoding and decoding. The relationship between a stimulus signal x and the neuronal response $\mathbf{k}$ is determined by the likelihood $p(\mathbf{k} \mid x)$, which can be decomposed into the encoding and the noise model. The mean spike counts $\mu_j(x) \equiv \mathrm{E}[k_j \mid x] = Tf_j(x)$ as functions of x specify the encoding, while for the noise model, we almost always choose the product of Poisson distributions in this article. Any function $\hat{x}$: $\mathbf{k} \mapsto \hat{x}(\mathbf{k})$ may be considered as a candidate estimator of x. The performance of an estimator $\hat{x}$ with respect to a given x is judged by its mean squared error risk.

2.1 Defining an Ideal Observer. The precision of a neural code is usually defined on the basis of the mean squared error risk $r(x) = \mathrm{E}[(x - \hat{x})^2 \mid x]$, with which a certain estimator $\hat{x}$ reconstructs the presented stimulus x from the response of the neuronal population. While the estimator is intended to represent an ideal observer, it is actually not possible to determine an estimator that is preferable among all possible estimators independently of the problem at hand. This is because no estimator exists that minimizes the risk $r(x)$ for all x, apart from trivial cases, where error-free estimation $r(x) \equiv 0$ is possible (Lehmann & Casella, 1999).

However, if one considers sequences of estimators $(\hat{x}_m)_{m=1}^{\infty}$, where each element $\hat{x}_m(\mathbf{k}(1), \ldots, \mathbf{k}(m))$ refers to m independent and identically distributed (i.i.d.) spike count vectors $(\mathbf{k}(1), \ldots, \mathbf{k}(m))$, the corresponding sequence of risk functions asymptotically decreases proportional to $1/m$ for many estimators.[1] This can essentially be explained by the central limit theorem. More precisely, it can be shown under some rather weak assumptions about $p(\mathbf{k} \mid x)$ and $p(x)$ (Lehmann & Casella, 1999) that the rescaled error $\sqrt{m}(\hat{x}_m - x)$ converges in law to a normal distribution with zero mean and a variance, which is the reciprocal value of Fisher information:

$$J[p(\mathbf{k} \mid x)] \equiv \mathrm{E}[(\partial_x \log p(\mathbf{k} \mid x))^2 \mid x]. \tag{2.5}$$

Such (sequences of) estimators are called asymptotically efficient.

In the case of a homogeneous Poisson process as considered in this article, it is equivalent to consider sequences of increasing decoding time windows $T_m = mT_0$. While the corresponding estimators are functions of a single spike count vector $\mathbf{k}$ only, this spike count vector can be interpreted as the sum $\sum_{t=1}^{m} \mathbf{k}(t)$ of m spike count vectors that are independently drawn from a Poisson distribution corresponding to the time window T_0. Thereby, no information gets lost because $\sum_{t=1}^{m} k_j(t)$ is a sufficient statistics for the parameter of the Poisson distribution (Lehmann & Casella, 1999). Accordingly, for asymptotically efficient estimators holds

$$\lim_{T \to \infty} r(x, T) \cdot J[\{f_j(x)\}_{j=1}^{N}] = 1, \tag{2.6}$$

[1] Note that the term *estimator* is used with different notions. While basically any arbitrary function of the observable random variables is called an estimator, if it is intended to predict some quantity, we here have a somewhat different meaning: If one talks about *the* maximum likelihood estimator or *the* mean square estimator, one does not refer to a unique function, but to a certain set of estimators defined by a unique construction rule. If the latter is specified, a sequence of observations naturally leads to a unique sequence of estimators.

where $r(x, T)$ denotes the risk depending on T and the Fisher information is determined by

$$J[\{f_j(x)\}_{j=1}^N] = T \sum_{j=1}^{N} \frac{f_j'^2(x)}{f_j(x)}, \tag{2.7}$$

which is obtained by inserting equation 2.1 into equation 2.5 (Paradiso, 1988; Seung & Sompolinsky, 1993).

While Fisher information also shows up in the Cramér-Rao bound for unbiased estimators in a similar way, the latter by itself is not sufficient to justify the use of Fisher information as a general measure for the coding precision. In its general version, the Cramér-Rao bound,

$$r(x) \geq \frac{(\partial_x \mathrm{E}[\hat{x}(\mathbf{k}) \mid x])^2}{J[p(\mathbf{k} \mid x)]} + (\mathrm{E}[\hat{x}(\mathbf{k}) \mid x] - x)^2, \tag{2.8}$$

is not unique for different estimators. Furthermore, even if uniqueness is given as, for example, in the case of uniformly unbiased estimators, a comparison of different encodings cannot necessarily be traced back to a comparison of some lower bounds on the decoding risks. Instead, it is indispensable to determine a sufficiently close approximation of the actual values. Since the exact equality $r(x) = J(x)^{-1}$ holds true only in very rare cases (see section A.2), the notion of asymptotic efficiency presented above is crucial for the use of Fisher information (for a more detailed discussion of differences and relationships between asymptotic theory and the Cramér-Rao bound, see, e.g., Lehmann & Casella, 1999). While some previous publications on population coding relate their results to the asymptotic limit (Paradiso, 1988; Seung & Sompolinsky, 1993; Dayan & Abbott, 2001), it still lacks a thorough discussion of the conditions, when this limit can be used to estimate the risk of an optimal estimator in cases where the number of neurons is arbitrary large but finite.

Before we go deeper into the discussion of asymptotic efficiency, we should first determine a notion of an ideal observer that holds beyond the scope of asymptotic efficiency. In principle, there are two approaches that enable a unique selection of an optimal estimator on the basis of the risk: either the class of allowed estimators is restricted a priori such that the order between the risks $r_1(x), r_2(x)$ of any two estimators becomes completely independent from x, or the estimators are compared on the basis of a loss functional $\mathcal{F}[r(x)]$ over the whole range of x. In other words, the unique selection of an estimator is possible only on the basis of further specifications.

Here we consider the average risk,

$$\chi^2 = \mathrm{E}[(\hat{x} - x)^2] = \mathrm{E}[\mathrm{E}[(\hat{x} - x)^2 \mid x]] = \mathrm{E}[\mathrm{E}[(\hat{x} - x)^2 \mid \mathbf{k}]] \tag{2.9}$$

which is well established in information theory (Cover & Thomas, 1991) and neuroscience (Dayan & Abbott, 2001; Salinas & Abbott, 1994; Johnson,

1996; Roddey, Girish, & Miller, 2000). According to χ^2, the best estimator $\hat{x}_{MS}$ is given by

$$\hat{x}_{MS}(\mathbf{k}) \equiv \underset{\hat{x}}{\operatorname{argmin}} \, \mathrm{E}[(\hat{x} - x)^2 \mid \mathbf{k}] = \underset{\hat{x}}{\operatorname{argmin}} \int (\hat{x}(\mathbf{k}) - x)^2 \, p(x \mid \mathbf{k}) \, dx, \quad (2.10)$$

which is termed the mean square estimator (MS estimator). Equation 2.10 can formally be solved so that the best estimator with respect to the mean squared error loss is known to be the conditional mean (Lehmann & Casella, 1999):

$$\hat{x}_{MS}(\mathbf{k}) = \mathrm{E}[x \mid \mathbf{k}] = \int x \, p(x \mid \mathbf{k}) \, dx. \quad (2.11)$$

The conditional mean squared error $\mathrm{E}[(\hat{x}_{MS} - x)^2 \mid \mathbf{k}]$ of $\hat{x}_{MS}$ given an observation $\mathbf{k}$ equals the variance $\mathrm{E}[x^2 \mid \mathbf{k}] - \mathrm{E}[x \mid \mathbf{k}]^2$ of the a posteriori distribution. Hence, the MMSE is generally determined by

$$\chi^2_{MS} = \mathrm{E}[(\hat{x}_{MS} - x)^2] = \mathrm{E}[x^2] - \mathrm{E}[\hat{x}^2_{MS}]. \quad (2.12)$$

2.2 MS Estimator and Fisher Information. Unfortunately, it is often not possible to derive the moments of the a posteriori distribution analytically, and numerical efforts can be immense. Provided a set of regularity conditions holds, however, the MS estimator is known to be asymptotically efficient (Lehmann & Casella, 1999). This means that with an increasing number of observations, the risk asymptotically approaches $1/J(x)$ for all x. Because this implies that the mean values of both converge, the MMSE is then asymptotically equal to the mean asymptotic error (MASE):

$$\chi^2_{AS} \equiv E\left[\frac{1}{J[\{f_j(x)\}_{j=1}^N]}\right] = \frac{1}{T}\int_0^1 \left(\sum_{j=1}^N \frac{f_j'^2(x)}{f_j(x)}\right)^{-1} dx. \quad (2.13)$$

Note, however that even if one can prove asymptotic normality for a certain sequence of estimators $(\hat{x}_m)_{m=1}^\infty$, it is still necessary to know how large m has to be so that the MASE becomes a good approximation of the MMSE (i.e., for the relative difference between both it holds $|\chi^2_{MS} - \chi^2_{AS}|/\chi^2_{MS} < \epsilon \ll 1$). We will demonstrate that T typically has a large effect on the shape of the MMSE-optimal code, although we have already shown that the large T limit can be considered as the limit of an asymptotically normal sequence of estimators. In contrast, the shape of Fisher-optimal codes is necessarily independent of the available decoding time because T appears only as a constant factor in equation 2.13.

Previous articles on population coding using Fisher information referred to the limit of large N rather than to the limit of large T considered above.

There is an important difference between these two kinds of limiting processes: As long as the tuning functions are taken to be static, the integration of spikes over time corresponds to a sum over i.i.d. spike count vectors. In contrast, for the spike counts of different neurons, we typically have $p(k_i \mid x) \neq p(k_j \mid x)$ for all $i \neq j$. This diversity of the tuning functions can crucially slow the convergence of the MMSE to the MASE, and it may even destroy the property of asymptotic efficiency. In fact, it is possible to construct sequences of tuning functions so that the difference between the MASE and MMSE becomes larger and larger the more tuning functions are taken into account by the MS estimator (an example will be given). Furthermore, we will show that tuning functions that lead to large Fisher information are particularly likely to underestimate the MMSE by far. This becomes a severe problem when Fisher information is used as an objective function in order to determine optimal encodings.

Another fundamental problem is that Fisher information can be used for only those encodings for which x is identifiable, which here means that the mapping of the tuning functions is one-to-one. If x is not identifiable, Fisher information may either underestimate or overestimate the true error by far. For example, a single symmetric tuning function centered at $1/2$ (the middle of the interval $(0, 1)$) cannot improve the mean squared error at all for any x, while Fisher information can be arbitrarily large everywhere. Conversely, the Fisher information of a single tuning curve that is constant somewhere within an arbitrary small but finite interval predicts a diverging error within this interval, while the c.m.s.e. in fact depends on the length of this interval and χ^2_{MS} can never be larger than the variance of the a priori distribution (see equation 2.12). Therefore, Fisher optimality in this article is defined to require both a minimal MASE and a one-to-one mapping of the tuning function array.

3 Optimal Gaussian Tuning Depends on Available Decoding Time

Consider the example of equidistant gaussian tuning curves on the unit interval

$$f_j^{gauss}(x) = f_{\max} \exp\left\{-\frac{1}{2}\left(\frac{x - j/N}{\sigma}\right)^2\right\}, j = 1, \ldots, N. \tag{3.1}$$

In order to determine the optimal scale with respect to the MASE, the corresponding Fisher information is calculated by inserting equation 3.1 into equation 2.7:

$$J[\{f_j^{gauss}(x)\}_{j=1}^N] = \frac{Tf_{\max}}{\sigma^2} \sum_{j=1}^{N} \left(\frac{x - j/N}{\sigma}\right)^2 \exp\left\{-\frac{1}{2}\left(\frac{x - j/N}{\sigma}\right)^2\right\}. \tag{3.2}$$

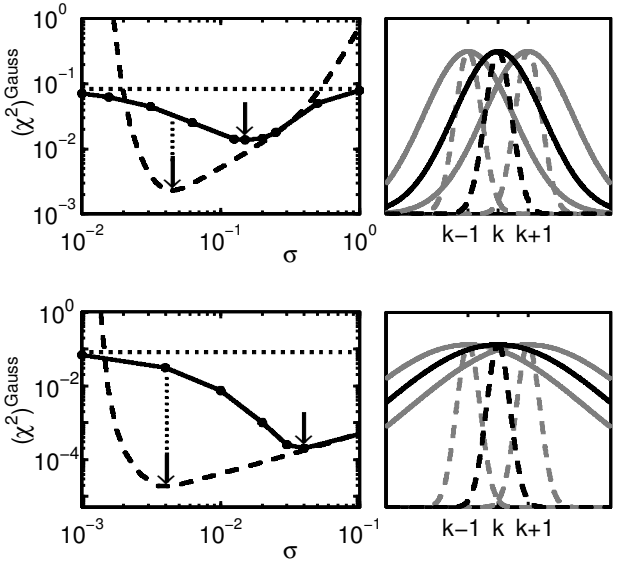

Figure 2: (Left) Log-log plot of the minimum mean squared error $(\chi^2_{MS})^{gauss}$ as a function of the scale σ for $f_{\max}T = 1$ (solid) compared with the mean asymptotic error $f_{\max}T(\chi^2_{AS})^{gauss} = (\chi^2_{AS})^{gauss} = \mathrm{E}[1/J]$ (dashed) in the case of $N = 10$ (upper) and $N = 100$ neurons (lower). The variance of the a priori distribution $\mathrm{Var}[x] = 1/12$ (dotted) provides an upper bound for χ^2_{MS}. The arrows indicate the different minima. (Right) Comparison of the optimal tuning curves with respect to the mean asymptotic error $(\chi^2_{AS})^{gauss}$ (dashed) and the minimum mean squared error $(\chi^2_{AS})^{gauss}$ (solid) in the case of $N = 10$ (upper) and $N = 100$ neurons (lower). The gray lines indicate the adjacent tuning curves.

Numerical integration over $1/J[\{f_j^{gauss}(x)\}_{j=1}^N]$ then yields the MASE $(\chi^2_{AS})^{gauss}$. If it is multiplied by $f_{\max}T$, the resulting expression becomes independent of time, which implies that the optimal scale with respect to Fisher information is independent of time too. In Figure 2, $f_{\max}T(\chi^2_{AS}(\sigma))^{gauss}$ is plotted as a function of the scale in the case of $N = 10$ and $N = 100$ exhibiting a unique minimum, for which the corresponding values are as follows.

N	σ_{AS}	$f_{\max}T(\chi^2_{AS})^{gauss}$
10	0.045	$2 \cdot 10^{-3}$
100	0.004	$2 \cdot 10^{-5}$

The use of the MASE as objective function is justified only in the case $T \to \infty$ of asymptotic normality. For finite T, however, it is necessary to check whether the MASE agrees with the MMSE. Hence, we computed

$(\chi^2_{MS})^{gauss}$ directly for the case of $f_{\max}T = 1$ using Monte Carlo methods (see the appendix).

The MMSE as a function of the scale for $f_{\max}T = 1$ is also plotted in Figure 2 (left, solid line), and the values of the minima are as follows:

N	σ_{MS}	$(\chi^2_{MS})^{gauss}$
10	0.11	$1.2 \cdot 10^{-2}$
100	0.04	$2 \cdot 10^{-4}$

By comparison, we find that the optimal scales with respect to $(\chi^2_{MS})^{gauss}$ are about one order of magnitude larger than one would conclude from the MASE. In particular, this difference between the short-term optimum and the long-term optimum scale becomes even larger when increasing the number of neurons from 10 to 100. Figure 2 (right) shows the corresponding tuning curves illustrating this relative increase. While the MASE is close to $(\chi^2_{MS})^{gauss}$ for scales that are larger than the optimal scale, the difference between both increases rapidly the more the scale is reduced from the optimal scale and reaches a maximum at the minimum MASE.

4 Fisher-Optimal Codes Without Tuning Curve Shape Constraints

The analysis of optimal gaussian tuning in the previous section indicates that Fisher-optimal codes are particularly likely to underestimate the MMSE if the time window is small. This mismatch between the MASE and the MMSE becomes even more dramatic in the case of Fisher-optimal codes if one does not stick to the restriction of gaussian-shaped tuning functions. This can be demonstrated by considering Fisher-optimal population codes where the tuning curves are not subjected to a priori constraints apart from a limitation of their dynamic range by a minimum firing rate $f_{\min}$ and a maximum firing rate $f_{\max}$. We will first determine the optimal tuning function in the case of a single neuron and then for multiple neurons.

4.1 Single Neuron. A way to find the tuning function that minimizes the MASE is to start with a calculus of variations for the MASE functional:

$$\frac{1}{T}\int_0^1 \frac{f(x)}{(f'(x))^2}\,dx. \tag{4.1}$$

A necessary condition for a minimum of the MASE functional is given by the corresponding Euler-Lagrange differential equation,

$$\frac{f(x)}{(f'(x))^2} + 2f'(x)\frac{f(x)}{(f'(x))^3} = C, \tag{4.2}$$

which is equivalent to the requirement of a constant Fisher information, because the left-hand side is proportional to $1/J[f]$. The unique solution

satisfying the boundary conditions $f(0) = f_{\min}$ and $f(1) = f_{\max}$ reads

$$f^{opt}(x) = \left[\left(\sqrt{f_{\max}} - \sqrt{f_{\min}}\right) x + \sqrt{f_{\min}}\right]^2. \tag{4.3}$$

While the calculus of variation does not account for solutions with kinks, one can prove with some additional effort that f^{opt} in fact constitutes the Fisher-optimal tuning function in the case of the Poisson noise model. Its Fisher information is $J[f^{opt}(x)] = 4T(\sqrt{f_{\max}} - \sqrt{f_{\min}})^2$. For constant additive gaussian noise, an analog analysis leads to a linear tuning function $f(x) = (f_{\max} - f_{\min})x + f_{\min}$, and in general it can be shown that a constant Fisher information is a necessary condition for Fisher-optimal codes.

4.2 Many Neurons. If x is encoded by more than one neuron, the requirement of identifiability of x does not necessarily imply any more that the tuning functions are monotonic. In particular, if at least one neuron has a strictly monotone tuning curve, all other neurons may have arbitrarily shaped tuning functions. This makes it easy to construct Fisher-optimal codes, for which the MASE vanishes. In particular, we will show that this is already possible for two neurons if they have, for example, the following tuning functions (see Figure 3),

$$f_1^{wave}(x) = f_{\max} x^2, \tag{4.4}$$

$$f_{2,\omega}^{wave}(x) = f_{\max}\,[(\omega x) \bmod 1]^2, \tag{4.5}$$

where we have set $f_{\min} = 0$ for convenience. In this example, the total Fisher information is also a constant:

$$J[\{f_1^{wave}(x), f_{2,\omega}^{wave}(x)\}] = 4 f_{\max} T(\omega^2 + 1). \tag{4.6}$$

Hence, the mean asymptotic error of this wave function encoding scheme equals $(\chi^2_{AS})^{wave} = [4 f_{\max} T(\omega^2 + 1)]^{-1}$, which becomes arbitrarily small with increasing ω. If Fisher information would be a general measure for the precision of population codes, this would imply that all coding problems could be solved with two neurons only. However, if we compare Fisher information with the precision of the MS estimator in the case of $f_{\max} T = 1$, we find that $(\chi^2_{MS})^{wave} > 0.06$ for all $\omega \geq 1$.

In summary, our analysis of Fisher-optimal codes bears two important conclusions. First, with respect to Fisher information, gaussian tuning curves are particularly bad codes, however large or small their tuning widths are. Second, Fisher-optimal codes are not necessarily advantageous in the case of finite time windows.

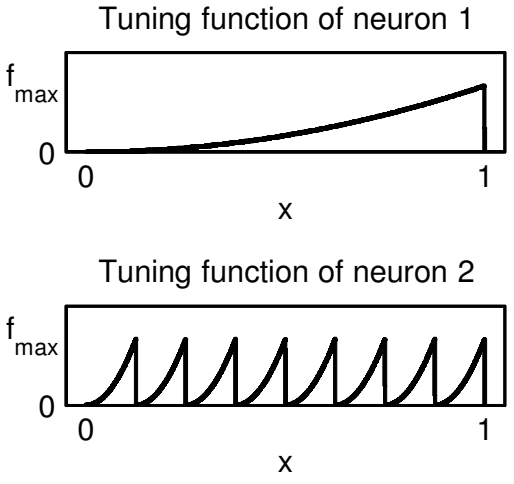

Figure 3: Fisher-optimal wave coding scheme consisting of two neurons. One tuning function ensures the identifiability (top). The other tuning function is a wave function that leads to arbitrary large Fisher information when the wavelength is decreased (bottom).

5 When and Why the Mean Asymptotic Error Fails

In the space of possible arrays of tuning functions, it is difficult to name clear-cut decision boundaries that tell precisely where χ^2_{AS} is a fairly good approximation of χ^2_{MS} and where it is not. Therefore, the goal of this section is to gain intuition about which features of an encoding strategy are most relevant for the correspondence between the MASE and the MMSE.

We begin with a simple example that demonstrates why the MASE and the MMSE need not converge in the large N limit. To this purpose, we construct a sequence of encodings $(f_j^{spec})_{j=1}^N$ by simply extending the wave coding scheme (equations 4.4 and 4.5) to

$$f_j^{spec}(x) := f^{wave}_{2,\omega(j)}(x), \tag{5.1}$$

with $\omega(1) = 1$. For any given T, $\omega(2)$ can be chosen sufficiently large that $(\chi^2_{AS})^{spec}_{N=2} \ll (\chi^2_{MS})^{spec}_{N=2}$. The subsequent $\omega(j)$, $j \geq 3$ are defined recursively by $\omega(j+1) := \omega(j) + j^\alpha$, where we introduced the exponent $\alpha \geq 0$ in order to indicate that Fisher information can increase with N arbitrarily fast. Even in the case of $\alpha = 0$, however, it holds $(\chi^2_{AS})^{spec} \ll (\chi^2_{MS})^{spec}$ for arbitrary N, because $(\chi^2_{AS})^{spec}$ decreases faster than $1/N$ for all N, which is in contradiction to the scaling of the mean squared error risk in the case of asymptotic efficiency.

In general, Fisher information, and hence the MASE, is a separable function in N and T:

$$\chi^2_{AS} = \frac{1}{T} s(N). \tag{5.2}$$

If asymptotic efficiency holds for large N, the MMSE has to decrease as N^{-1}. Therefore, it is a necessary condition for asymptotic efficiency with respect to the limit of large N that $s(N)$ decreases proportional to N^{-1}, too. This condition is not necessarily fulfilled, but as in the example, the population Fisher information can grow much faster with N than linear. Moreover, it is possible to construct encodings for which the MMSE decreases substantially faster than N^{-1} (an example is given below), so that the case of asymptotic efficiency holds only for particularly suboptimal codes, which exhibit a high degree of redundancy.

In the example above, there are tuning functions that map very distant values of x to the same firing rate and the mismatch between the MASE and the MMSE increases with an increasing number of maxima and minima in the tuning functions. In the following example, we will show that a restriction of the number of maxima is not sufficient to ensure $\chi^2_{AS} \approx \chi^2_{MS}$, but the matching of these two quantities crucially depends on nonlinearities in the tuning functions.

Consider the following class of Fisher-optimal codes built with monotonic tuning functions[2]

$$f^{mono}_{j,\nu}(x) = \begin{cases} f_{\max}\left(Nx - \frac{(l-1)N+j-l}{\nu}\right)^2, & \frac{(l-1)N+j-1}{\nu N} < x < \frac{(l-1)N+j}{\nu N} \\ f_{\max}\left(\frac{l}{\nu}\right)^2 \quad , & \frac{(l-1)N+j}{\nu N} < x < \frac{lN+j-1}{\nu N} \end{cases}, \tag{5.3}$$

where j denotes the neuron index and $\nu = 1, 2, 3, \ldots$ specifies the shape of the tuning function array (see Figure 4). Each tuning curve is completely determined, if one let l run through all integer values $l = 1, \ldots, \nu$. The Fisher information of these encodings is independent of ν:

$$J[\{f^{mono}_{j,\nu}(x)\}_{j=1}^N] = 4 f_{\max} T N^2. \tag{5.4}$$

In the limit $\nu \to \infty$, this coding scheme cannot be distinguished from N identical tuning functions $f^{mono}_{j,\infty}(x) = f_{\max} x^2$ (see Figure 4). However, the population Fisher information $J[\{f^{mono}_{j,\nu}(x)\}_{j=1}^N]$ is N times larger than the population Fisher information of the asymptotic tuning functions $\{f^{mono}_{j,\infty}(x)\}_{j=1}^N$

[2] The proof of Fisher optimality is based on the same reasoning as the proof of Fisher optimality for unimodal tuning functions given in section 6. However, the set of encodings $\{\{f^{mono}_{j,\nu}(x)\}_{j=1}^N : \nu = 1, 2, 3, \ldots\}$ does not contain all Fisher-optimal tuning function arrays.

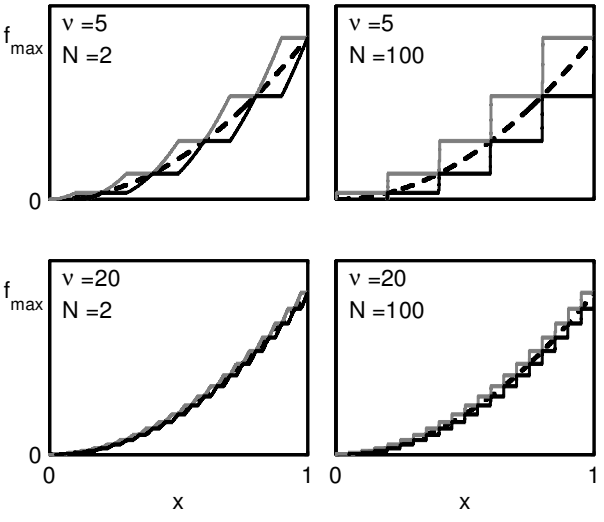

Figure 4: Four examples of Fisher-optimal encodings built with monotonic tuning functions as described by equation 5.3. The left column shows the case of $N = 2$, and the right column shows the case of $N = 100$, where we plotted only the first tuning function ($j = 1$, gray) and the last one ($j = N$, black). The intermediate tuning functions ($j = 2, \ldots, N-1$, not shown) lie between the first and the last one. Independent of N, all tuning functions converge to $f_\infty^{mono}(x) = f_{\max} x^2$, which is illustrated by the comparison of the case $\nu = 5$ (upper row) with the case $\nu = 20$ (lower row).

(this is possible because limiting values are not invariant under a change in the order of limiting processes).

This example nicely demonstrates that Fisher information behaves as if all structures in the tuning functions are of the same relevance independent of their length scale. In fact, however, nonlinearities in the tuning functions become relevant only if they are observable at a scale that is naturally set by the scattering of the noise distribution. Correspondingly, the critical decoding time T_c that is necessary to approach the asymptotic normal case, which is described correctly by Fisher information, increases with increasing ν. For large ν, T_c has to be roughly proportional to ν^2 because the squared deviations of the tuning functions $f_{j,\nu}^{mono}(x)$ from the smoothed tuning functions $f_{j,\infty}^{mono}(x)$ are of the order of $(1/\nu)^2$ and become relevant only if the mean squared error, which scales like $1/T$, is of the same order (or smaller). Therefore, the critical decoding time diverges for a diverging ν, which explains the difference in the population Fisher information between $\{f_{j,\nu}^{mono}(x)\}_{j=1}^N$ and $\{f_{j,\infty}^{mono}(x)\}_{j=1}^N$. Finally, it is worthwhile to note that the ramp coding scheme obtained for $\nu = 1$ can be considered the best among the class of

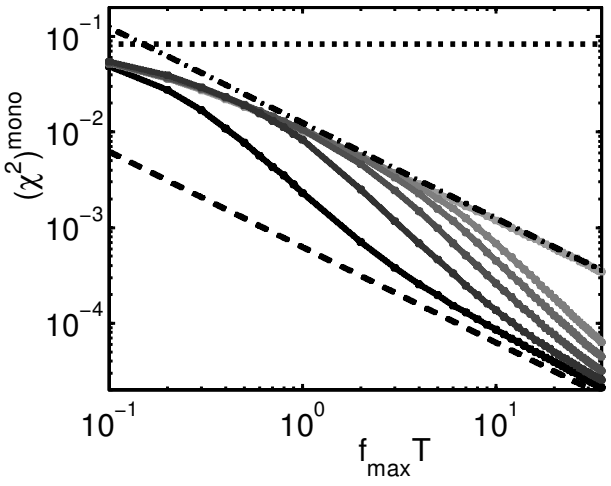

Figure 5: The MMSE of $\{f_{j,\nu}^{mono}(x)\}_{j=1}^{20}$ is displayed as a function of the decoding time T for $\nu = 1, 2, 3, 4, 5$ and $\nu = 20$ (solid, from dark to pale). Although all encodings have the same Fisher information (dashed), the critical decoding time increases with increasing ν. If T is smaller than the critical decoding time and larger than $3/(f_{\max}N)$, the MMSE curves are well described by the Fisher information of the asymptotic tuning functions $J[\{f_{j,\infty}^{mono}(x)\}_{j=1}^{20}]$ (dot-dashed). For $T < 3/(f_{\max}N)$, the bound given by the a priori variance $1/12$ (dotted) is most relevant.

Fisher-optimal coding schemes built with nondecreasing tuning functions, because it has the smallest critical decoding time. This is demonstrated in Figure 5, where we show how the different dependency of $(\chi_{MS}^2)^{mono}$ on the decoding time is affected by the parameter ν.

Taken together, Fisher information is a measure of the long-term coding precision of population codes in the first place, while in the case of finite T, one has to check carefully whether Fisher information provides correct results for the minimum mean squared error. As a rule of thumb, one can say that the smoother the tuning functions, the higher the probability is that the MASE matches the MMSE.

Hitherto, we considered examples only where $\chi_{AS}^2 \leq \chi_{MS}^2$, and one might suspect that this holds true in general according to the Cramér-Rao bound. Apart from the trivial fact that χ_{AS}^2 diverges in the limit $T \to 0$ in contrast to $\chi_{MS}^2 \leq \mathrm{Var}[x]$, we will now show that for arbitrary large T, arrays of tuning functions exist, for which $\chi_{AS}^2 \gg \chi_{MS}^2$. In order to show this, we consider a generalized ramp coding scheme,

$$f_{j,\alpha}^{ramp}(x) = f_{\max}\left([Nx - j + 1]_+ - [Nx - j]_+\right)^{\alpha}, \tag{5.5}$$

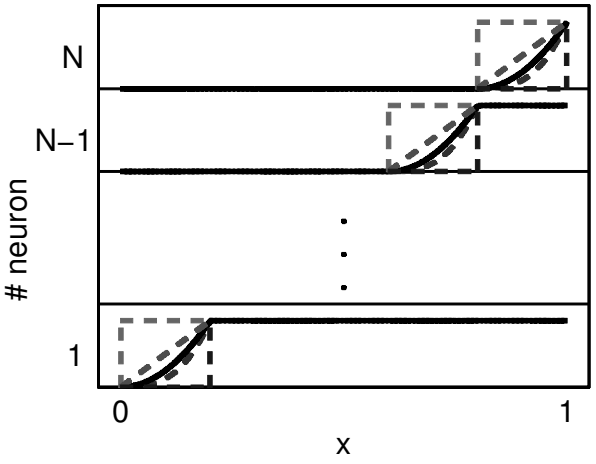

Figure 6: Illustration of the ramp coding schemes for different values of α. The tuning functions differ only in the shape of the ramp, which is determined by α. Apart from the Fisher-optimal encoding, which is given for $\alpha = 2$ (solid), we also plotted some other shapes of the ramp (dashed) that correspond to $\alpha = 0, 1, 3, \infty$ (from pale to dark).

where $[y]_+ = y\Theta(y)$ is the rectifier function. The parameter $\alpha \in [0, \infty)$ can be used to change the tuning curves smoothly from linear ramp functions ($\alpha = 1$) to step functions ($\alpha \to 0$ or $\alpha \to \infty$), which is illustrated in Figure 6. Furthermore, it is important to note that $f_{j,2}^{ramp}(x)$ is identical to $f_{j,1}^{mono}(x)$, which is the Fisher-optimal encoding for nondecreasing tuning functions with the smallest critical decoding time. According to equation 5.5, the MASE becomes

$$(\chi_{AS}^2(\alpha))^{ramp} = \frac{1}{f_{\max} T N^2} \cdot \begin{cases} \frac{1}{\alpha^2(3-\alpha)}, & \alpha \in (0,3) \\ \infty, & \text{otherwise} \end{cases}, \tag{5.6}$$

which implies that for all T, there is an $\alpha < 3$ so that $(\chi_{AS}^2(\alpha))^{ramp} \gg \mathrm{Var}[x] \geq (\chi_{MS}^2(\alpha))^{ramp}$. In the case of $\alpha \geq 3$, the MASE diverges however large T may be. The strong dependence on α in case of $(\chi_{AS}^2(\alpha))^{ramp}$ is not likely to hold for the MMSE too. In particular, it is quite surprising that $(\chi_{AS}^2(3))^{ramp}$ diverges, although the corresponding tuning function array looks very similar to that in the Fisher-optimal case ($\alpha = 2$). The reason for this huge discrepancy is that Fisher information in general cannot account for the precision of encodings, for which $J(x)$ has first-order zeros (or higher order). One could say that the latter is a weaker form of nonidentifiability. Although the encoding is one-to-one, Fisher information cannot account for the precision if the slope of all tuning functions becomes too small somewhere.

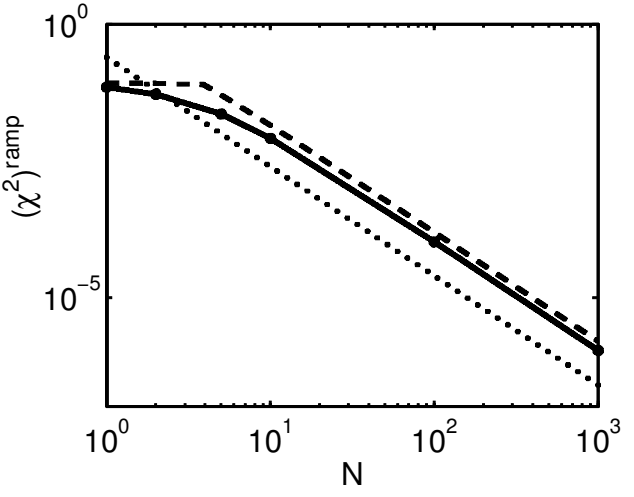

Figure 7: The MMSE (solid) of the Fisher-optimal ramp encoding ($\alpha = 2$) is shown in the case of $f_{\max}T = 1$. It is very close to the upper bound (dashed) that is the minimum of the two upper bounds given by equation 5.7 and the a priori variance. The dotted line indicates the MASE of the Fisher-optimal ramp encoding, which may be considered as a lower bound on the MMSE for all α provided N is sufficiently large. Therefore, all ramp encodings perform similarly well in case of $f_{\max}T = 1$.

In contrast to the strong dependence of the MASE on α, the MMSE in the case of $f_{\max}T = 1$ is very similar for all α. It can be shown analytically (see the appendix) that the following inequality holds for all α,

$$(\chi^2_{MS})^{ramp} \leq \frac{1}{N^2 p} + \frac{1}{N^3}\left(1 - \frac{1-q^N}{p^2}\right), \tag{5.7}$$

where $p = 1 - e^{-f_{\max}T}$ increases and $q = 1 - p$ decreases with the length T of the time window. As one can see in Figure 7, this bound is quite close to $(\chi^2_{AS}(2))^{ramp}$ of the Fisher-optimal code.

To summarize, all examples discussed in this section demonstrate that the matching of the MASE with the MMSE critically depends on the effect of nonlinearities of the tuning functions on the stimulus reconstruction. This is also suggested by the fact that Fisher information is calculated only on the basis of the local shape of the likelihood function $p(\mathbf{k} \mid x)$, which corresponds to a linear extrapolation around the true value x_{true}.

In the case of the Poisson noise model considered in this article, it is possible to give this statement about the locality of Fisher information a precise meaning: If $g_j = f_j^{-1}$ denotes the local inverse function of a tuning

curve f_j, then the conditional variance $\mathrm{Var}[g_j(k_j/T) \mid x]$ at point x can be expressed by

$$\mathrm{Var}[g_j(k_j/T) \mid x] = \mathrm{Var}\left[g_j(f_j(x)) + g_j'(f_j(x))\frac{k_j - Tf_j(x)}{T} + \mathcal{O}(k_j^2) \mid x\right] \quad (5.8)$$

$$= \frac{1}{J[f_j(x)]} + \mathrm{Var}[\mathcal{O}(k_j^2) \mid x]. \quad (5.9)$$

In a similar way, Fisher information shows up if one determines the error of the MS estimator in the limit of vanishing noise. For any given x, the MS estimator can be approximated by a linear function of $\mathbf{k}$ in this limit. In particular, this linear function can be set to the form of a superposition of the inverse tuning functions $g_j = f_j^{-1}$, because the MS estimator is asymptotically unbiased. Therefore, it holds that

$$\hat{x}_{MS}(\mathbf{k}) \approx x + \sum_j W_j g_j'(f(x))\,(k_j/T - f_j(x)) = x + \sum_j W_j \frac{k_j/T - f_j(x)}{f_j'(x)}, \quad (5.10)$$

where the $\{W_j\}_{j=1}^N$ stand for an arbitrary weighting with $\sum_{j=1}^N W_j = 1$. Accordingly, the conditional error variance of the MS estimator is given by

$$\mathrm{E}[(\hat{x}_{MS}(\mathbf{k}) - x)^2 \mid x] = \sum_j W_j^2 \frac{\mathrm{Var}[k_j \mid x]}{(Tf_j'(x))^2}. \quad (5.11)$$

Minimizing equation 5.11 under the constraint of $\{W_j\}_{j=1}^N$ yields

$$W_j = \frac{(Tf_j'(x))^2}{\mathrm{Var}[k_j \mid x] \sum_j \frac{(Tf_j'(x))^2}{\mathrm{Var}[k_j|x]}}, \quad (5.12)$$

and correspondingly, the conditional error variance becomes

$$\mathrm{E}[(\hat{x}_{MS}(\mathbf{k}) - x)^2 \mid x] = \frac{1}{\sum_j \frac{(Tf_j'(x))^2}{\mathrm{Var}[k_j|x]}} = \frac{1}{J[\{f_j(x)\}_{j=1}^N]}. \quad (5.13)$$

This somewhat heuristic calculation suggests that the inverse Fisher information is a good approximation of the risk of the MS estimator, when the scattering of the MS estimator around the true value of x is restricted to a region within which the tuning functions may be considered linear.

6 Are Fisher-Optimal Codes Unique?

In order to test the idea of efficient coding as a first principle for population codes in the brain, it is important to deduce characteristic predictions for the shape of neuronal tuning functions from it that can be directly compared with experimental data. Therefore, it is important to know whether tuning functions of Fisher-optimal codes exhibit unique features that can be observed experimentally. Several theoretical studies have focused on whether a large or a small tuning width is advantageous with respect to its coding efficiency. While Fisher information can diverge in the case of two neurons only, if no particular constraints are imposed on the shape of the tuning function, the MASE of Fisher-optimal codes remains finite if the number of maxima of each tuning function is set to be limited. This is clearly the case for unimodal tuning functions, which have one maximum only. For such encodings, Fisher information cannot increase faster than proportional to N^2, provided identifiability of x.

Here we derive Fisher-optimal unimodal tuning curves. In order to avoid asymmetries due to the boundaries of the interval, we switch to the case of a circular random variable (which could represent the angle of an oriented bar). The ring topology of the circular random variable requires modifying the Euclidean distance slightly,

$$D(x_1, x_2) = \min\{|x_1 - x_2|, 1 - |x_1 - x_2|\}, \tag{6.1}$$

by always choosing the path of smaller distance (Lehmann & Casella, 1999). Accordingly, the MMSE is then given by

$$(\chi^2_{MS})^{uni} = \mathrm{E}[D(\hat{x}_{MS} - x)^2]. \tag{6.2}$$

While this modification reduces the a priori error compared to the case without periodic boundary conditions, it has no effect asymptotically.

For convenience, we assume that x is encoded by unimodal symmetric tuning curves of identical shape with equidistantly distributed centers $c_j = j/N$,

$$f_j^{uni}(x) = \begin{cases} f_{\max} & , \quad D(x, c_j) \le a \\ g\left(\frac{D(x,c_j)-a}{b-a}\right) & , \quad a < D(x, c_j) < b \\ f_{\min} & , \quad b \le D(x, c_j) \le \frac{1}{2} \end{cases}, \tag{6.3}$$

where $g\colon (0, 1) \to [f_{\min}, f_{\max}]$ is a monotone decreasing and otherwise arbitrary function.

Since the Fisher information $J[f_j^{uni}(x)]$ can be positive only for $a < D(x, c_j) < b$, we refer to the corresponding regions as Fisher information regions (F regions) of the tuning functions. Independent from the function $g(z)$,

$J[f_j^{uni}(x)]$ is proportional to the inverse squared F region width $(b-a)^{-2}$. Therefore, it is a necessary condition for a minimum of the MASE that the F regions of different neurons must not overlap, because the contributions of different neurons add at most linearly to the total Fisher information.

Then the evaluation of the MASE can be decomposed,

$$\mathrm{E}\left[\frac{1}{J^{uni}(x)}\right] = \int\limits_{D(x,c_j)<a} \frac{dx}{J^{uni}(x)} + \int\limits_{a<D(x,c_j)<b} \frac{dx}{J[f_j^{uni}(x)]} + \int\limits_{b<D(x,c_j)<0.5} \frac{dx}{J^{uni}(x)}, \tag{6.4}$$

and we can conclude from section 4.1. that $g(z) = ((\sqrt{f_{\max}} - \sqrt{f_{\min}})(1-z) + \sqrt{f_{\min}})^2$ is the minimizer of the second term. It remains to determine the optimal choice of a and b. As we will show, the MASE becomes a minimum if the F region width is set to $b - a = 1/(2N)$, and $b = k/(2N)$ for any $k \in \{1, 2, \ldots, N\}$ (see Figure 8). In this case, the total Fisher information J^{uni} does not depend on x so that it holds

$$\mathrm{E}\left[\frac{1}{J^{uni}}\right] = \frac{1}{\mathrm{E}_x[J^{uni}]} = \frac{1}{16(\sqrt{f_{\max}} - \sqrt{f_{\min}})^2 TN^2}. \tag{6.5}$$

Because $\mathrm{E}[J^{uni}]$ decreases if $b-a$ is increased, it follows with the Jensen inequality,

$$\mathrm{E}\left[\frac{1}{J}\right] \geq \frac{1}{\mathrm{E}_x[J]}, \tag{6.6}$$

that the optimal F region width $b-a$ cannot be larger than $1/(2N)$ (see Figure 9). However, $b-a$ cannot be smaller than $1/(2N)$ due to the requirement of identifiability as well as due to the fact that the MASE diverges for all encoding strategies with $b - a < 1/(2N)$.

Since the identical minimal MASE is achieved for sharp tuning as well as for broad tuning, the length of the tuning width $w := a + b$ is not appropriate to characterize the Fisher-optimal code in the class of unimodal tuning functions. Obviously, this conclusion holds true also for suboptimal coding schemes, such as gaussian or cosine tuning functions. In fact, they have to be considered a special choice out of many equivalent[3] codes that all have the same shape of decay $g(z)$ but are different with respect to their tuning width. Instead of the tuning width, we find that in general, the length of the

[3] Equivalence with respect to the population Fisher information.

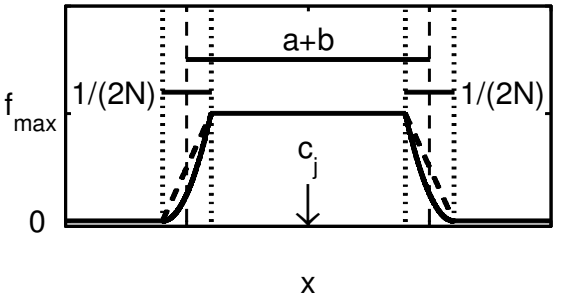

Figure 8: The Fisher-optimal unimodal tuning curve with the smallest critical decoding time is flat with small edges of length $1/(2N)$. Fisher information and the special type of noise model are relevant for the shape of optimal tuning curves only within these edge regions. The solid line refers to a Poisson noise model and the dashed line to additive gaussian noise of arbitrary variance.

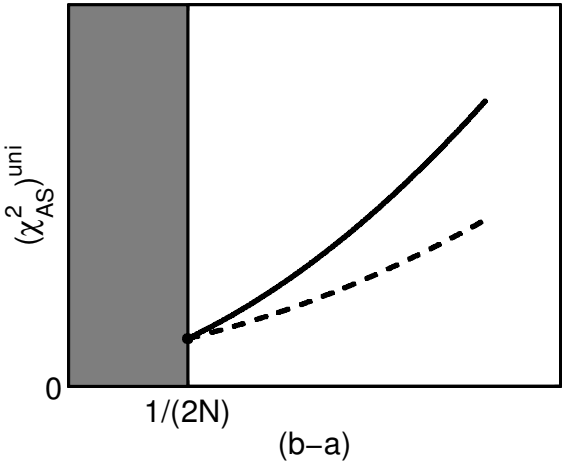

Figure 9: Sketch of the MASE (solid) as a function of the F-region width $(b-a)$. If $(b-a) < 1/(2N)$, there is an interval with zero Fisher information and hence the MASE diverges (gray region). For $(b-a) = 1/(2N)$, we have derived the encoding with minimal MASE, for which we found that it equals $1/\mathrm{E}[J]$ of that encoding (dot). Since the MASE is larger than or equal to $1/\mathrm{E}[J]$ (dashed) and this lower bound is an increasing function of $(b-a)$, the MASE is an increasing function of $(b-a)$ too, independent of the particular shape of the encoding. It follows that the encoding that minimizes the MASE in case of $(b-a) = 1/(2N)$ is also Fisher optimal compared to coding schemes with different F-region widths.

F region width is crucial for Fisher optimality because it has to be as small as possible. Accordingly, steep changes and flat plateaus are the main signatures of Fisher-optimal tuning curves, which is the more true the larger the populations are, because the minimal average F region width scales like $1/N$.

While the tuning width failed to constitute a characteristic feature of Fisher-optimal codes, other global aspects may be suitable for this purpose. The derivation above suggests that Fisher-optimal unimodal tuning functions are approximately box shaped if N is sufficiently large. However, the set of Fisher-optimal codes with unimodal tuning functions considered above is not complete, because we imposed various additional constraints on the tuning functions there. In fact, there are many more Fisher-optimal unimodal tuning functions with the same MASE as given by equation 6.5 if we drop these assumptions. For example, if we require monotony instead of strict monotony for $g(z)$ only, g may have arbitrarily many constant parts. Then similar to the idea underlying the Fisher-optimal class of monotonic tuning function encodings given by equation 5.3, this allows the construction of various Fisher-optimal codes, which have no features in common that could be tested experimentally.

7 Optimal Tuning Width—a Question of Energy?

While it was not possible to determine an optimal tuning width with respect to Fisher information under the constraints considered above, we suspect that in reality, energy consumption plays a crucial role for the tuning widths favoring sparse codes (Levy & Baxter, 1996). Before we derive an optimal width under this additional constraint, we want to know how much the MMSE of the Fisher-optimal tuning curves $\{f_j^{uni}\}_{j=1}^N$ depends on the width. Therefore, we computed the minimum of equation 6.2 numerically for $f_{\min} = 0$ and $f_{\max}T = 1$ (this corresponds to $f_{\max} = 200$ Hz and $T = 5$ ms) in the case of $N = 10$ and $N = 20$ neurons. It turns out that $w \approx 1/2$ is optimal, while the objective function is flat in a wide region around this optimum (see Figure 10). Furthermore, there is a slight asymmetry: $(\chi_{MS}(w))^{uni}$ increases not as fast in the direction of sparse codes as in the other direction. This asymmetry is due to the Poisson noise model, for which the noise variance increases proportionally to the average firing rate.

While the broadness of the minimum of the MMSE as a function of the tuning width does not indicate a substantial advantage of a certain receptive field size, this is likely to change if energy consumption is taken into account. Hitherto, the coding efficiency was limited only by a constraint on the power ($f(x) \leq f_{max}$), which can be motivated, for example, by the refractory period of a neuron after the generation of an action potential. On the other hand, it is likely that energy constraints are relevant because the average interspike intervals of cortical neurons are much larger than their refractory period.

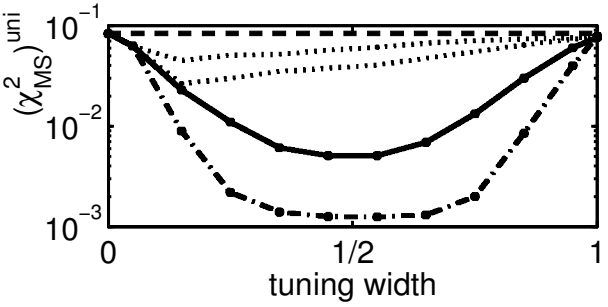

Figure 10: MMSE as a function of the tuning width in the case of $N = 10$ neurons (solid) and $N = 20$ neurons (dot-dashed). The dashed line indicates the a priori variance $\mathrm{Var}[x] = 1/12$ that is an upper bound for χ^2_{MS}. The dotted lines denote the results for $N = 10$ (upper) and $N = 20$ (lower) neurons if the energy constraint (see equation 7.1) is taken into account.

This fact can be taken into account if one assumes an additional upper bound for the mean firing rates $\mathrm{E}[f_j(x)] \leq \langle f \rangle_{\max}$.

In order to demonstrate the effect of this energy constraint, we have also calculated the minimum mean squared error in the case where the mean firing rate is limited by $\langle f \rangle_{\max} = f_{\max}/20$ (this corresponds to a maximum firing rate of 200 Hz and a mean firing rate of 10 Hz). Therefore both constraints together can be expressed by a width-dependent maximum firing rate:

$$\tilde{f}_{\max}(w) = \min\left(f_{\max}, \frac{f_{\max}}{20w}\right). \tag{7.1}$$

The resulting $(\chi^2_{MS})^{uni}_{energy}$ is shown in the case of $f_{\max}T = 1$ in Figure 10 by the dotted line, and it has a distinct minimum located at the maximal w, for which $\tilde{f}_{\max}(w) = f_{\max}$. Furthermore, the MASE exhibits a clear asymmetry toward smaller tuning widths, being minimal for all $w \leq \langle f \rangle_{\max}/f_{\max} \leq 1/20$.

This preference for smaller tuning widths gives a precise meaning to the statement that sparse coding can be explained by constrained energy consumption. In contrast to previous conclusions in van Vreeswijk (2001), this result does not rely critically on the Poisson noise model, but also holds true in the case of a gaussian noise model. In fact, the Fisher-optimal code for the gaussian noise model differs from the Poisson case only by a small change of the function $g(z)$, which becomes $g(z) = (f_{\max} - f_{\min})(1-z) + f_{\min}$ (see Figure 8, dashed line). This leads to a MASE $v/[2N(f_{\max} - f_{\min})]^2$, where $v = \mathrm{Var}[k_j \mid x]$ denotes the constant noise variance. Correspondingly, the MASE is again minimal for all $w \leq \langle f \rangle_{\max}/f_{\max} \leq 1/20$ in the same way as in the Poisson case. Moreover, the Poisson model itself is not sufficient to explain small receptive fields, because $(\chi^2_{AS})^{uni}$ is identical for all tuning

widths $w = 1/N, 2/N, \ldots, 1$ and the asymmetry of $(\chi^2_{MS})^{uni}$ in the absence of energy constraints is very weak.

8 Discussion and Conclusion

This article addresses several questions that arise if efficient coding is used as a first principle (Attneave, 1954; Barlow, 1959) in order to explain response properties of neuronal populations. The questions range from the problem of defining an ideal observer to the issue of how relevant the need for efficient signal transmission is as a determinant of the shape of neural codes compared to other constraints that are additionally imposed.

The ideal observer paradigm corresponds to the idea of evaluating all available information about the stimulus that in principle can be read out from the neuronal response. In almost all relevant cases, however, it is not possible to determine an estimator that is optimal for all possible stimuli. Therefore, it is not a trivial question which objective function is best suited to the problem of optimal coding. Although many studies use Fisher information as a measure for the precision of an ideal observer, this approach is justified only with respect to the limit of asymptotic normality. In this limit, the risk functions of many relevant estimators become equal to the inverse of Fisher information. Among others, this is the case for the risk of the maximum likelihood estimator and the MS estimator. Out of the asymptotic efficient estimators, we chose the MS estimator as an ideal observer because it accounts for all available prior knowledge on the stimulus statistics $p(x)$ and thus reflects the ultimate limit of optimal stimulus reconstruction.

The use of prior knowledge and its proper combination with the transmitted data is indispensable for efficient coding at short timescales. This is due not only to the bias-variance trade-off, but also in case of large N, prior knowledge, like the limitations on the dynamic range, cannot be neglected because it may have a substantial effect on the shape of optimal codes. Therefore, the MS estimator appears to be a reasonable choice because it allows for a very flexible incorporation of all sorts of prior information in a rather straightforward way.

As with any Bayesian type of estimator, one has to take great care with the choice of the a priori distribution. There are, however, well-developed methods to determine noninformative a priori distributions that do not introduce more specifications than are actually known (Lehmann & Casella, 1999). While the uniform prior distribution chosen in this article plays the role of a dummy distribution, it can be seen as a simple choice of a noninformative prior given the dynamic range of the signal x, because it has no effect on the posterior distribution apart from cutting it to the range of x.

For the need of signal processing, it is natural to require a constant upper bound for the risk function in order to achieve a certain precision for all x. This in turn suggests choosing the minimax estimator that minimizes

the worst-case loss functional $\mathcal{F}[r(x)] := \sup_{x\in[0,1]} r(x)$ instead of the mean. Due to the uniform prior, however, the risk function of the MS estimator depends only slightly on x, and its maximum rapidly becomes close to that of the minimax estimator with increasing data. In comparison to the minimax error, the MMSE has the advantage of being less sensitive to special assumptions on the coding model. In particular, the average risk does not depend on singularities of the risk function, and its computation is by far not so demanding. Furthermore, the MS estimator is always admissible, and it is asymptotically efficient, which allows for relating the MMSE to the average inverse Fisher information.

Taken together, we believe that the MMSE provides a reasonable objective function for optimal population codes that holds beyond the scope of asymptotic efficiency. Moreover, it can be estimated from neurophysiological experiments, without the need to specify an estimator a priori to the data (Roddey et al., 2000).

We also computed the mutual information for all examples presented in this article. Like Fisher information, mutual information becomes equivalent to χ^2_{MS} in the case of gaussian distributions so that all three measures can be used equivalently as objective functions in the case of asymptotic normality (see also Clarke & Barron, 1990; Brunel & Nadal, 1998). As an interesting fact, however, we note that mutual information becomes similar to the MMSE more rapidly than the MASE for all examples considered in this article (not shown).

While asymptotic efficiency is obtained for many population codes, we demonstrated that the use of Fisher information to characterize the precision of an encoding for a given decoding time T strongly depends on the particular coding scheme. As an example, we have shown that the optimal width of a population of gaussian tuning curves depends on the available decoding time, while Fisher-optimal codes are always independent of T.

The choice of the counting time window length can be related to the timescale at which neurons integrate over their synaptic inputs. Since the high degree of irregularity of neuronal discharge in cortex (Softky & Koch, 1993) implies that the effective integration time constant is of the order of a few milliseconds, we were most interested in the situation where the spike count of a single neuron is of the order of one.

We demonstrated that the critical decoding time T_c that is necessary for a sufficient matching of the MASE and the MMSE is typically increased by nonlinearities of the tuning functions. In particular, T_c grows with the frequency with which the tuning functions rise and decay between their minimum and maximum firing rate.

If Fisher information is used as an objective function in order to determine optimal coding schemes, one is led to tuning functions with a slope that is as large as possible. In fact, the Fisher information of a (bounded) tuning function behaves like a penalty term for regularization. Because Fisher information is intended to become as large as possible, however, Fisher op-

timality has quite the opposite effect of regularization, and, hence, it cannot be expected to rule out a large number of coding schemes on the basis of Fisher information only.

As we showed with an example in section 4.2, where only $f_{\max}T$ and N are given, the MASE can be reduced to zero with only two neurons. This means that any additional neuron is completely superfluous with respect to asymptotic optimality, and hence no substantial constraint is imposed by Fisher information. In contrast, a clear optimum within the class of bounded functions can be found on the basis of the MMSE, which we will show in a forthcoming paper (Bethge, Rotermund, & Pawelzik, 2002).

If the total number of maxima of the tuning functions is finite, the MASE remains finite too. However, in these cases, there is no unique Fisher-optimal code, but very many encodings achieve the same minimal MASE. This can be conceived from the case of nondecreasing tuning functions, for which we presented an infinitely large set of Fisher-optimal encodings (that was still not complete). The method with which Fisher optimal codes can be derived was presented in the case of unimodal tuning functions. It is not restricted to monotonic or unimodal tuning functions, but can also be applied to tuning functions with more maxima. The crucial point is that the F regions of Fisher-optimal tuning functions do not overlap, and the total length of the F regions of a single tuning function f_j equals $d_j / \sum_{i=1}^{N} d_i$, where d_j denotes the number of how many times $f_j(x)$ is allowed to traverse the dynamic range.[4] In the simplest case, the tuning function f_j then has d_j regions within which the tuning function increases (or decreases) quadratically, as given by equation 4.3. We therefore have the general formula for the Fisher information of Fisher-optimal codes:

$$J = 4T\left(\sqrt{f_{\max}} - \sqrt{f_{\min}}\right)^2 \left(\sum_{j=1}^{N} d_j\right)^2 . \tag{8.1}$$

It is also possible that the quadratic increase itself is interrupted by flat regions as it is the case for $f_{j,\nu}^{mono}(x)$ and $\nu > 1$, so that the F regions can be scattered over the entire range of x. Due to this freedom, the number of Fisher-optimal codes is very large, and there are no common features that could be tested experimentally.

Fisher-optimal codes do not perform equally well with respect to the MMSE. Therefore, it is natural to choose that Fisher-optimal encoding that has the smallest critical decoding time. Accordingly, one generally obtains tuning functions with flat maxima and quadratically decaying edges that are as steep as possible.

[4] For example, $d_j = 1$ in case of monotonic tuning functions and $d_j = 2$ in case of unimodal tuning functions.

Furthermore, we found that the question of whether small or broad tuning widths are advantageous cannot be decided if only the number of neurons N, the available decoding time T, and the maximum firing rate $f_{\max}$ are given. Instead, we demonstrated that a limitation of the average firing rate, which can be motivated by energy consumption, naturally breaks the symmetry toward sparse codes with small tuning widths.

Throughout this article, we considered the estimation of a single parameter only. In the case of multiparameter estimation, the choice of the squared error distance is not sufficient to enable a well-posed comparison of different coding schemes, but additional specifications become necessary. While the optimization can be very complicated if the loss functional $\mathcal{L}[\hat{\mathbf{x}}, \mathbf{x}]$ depends on different dimensions in a nonlinear fashion, it becomes rather simple if it is multilinear, that is,

$$\mathcal{L}[\hat{\mathbf{x}}, \mathbf{x}] = \mathcal{L}(\chi_1^2, \ldots, \chi_D^2) = \sum_{d=1}^{D} c_d \chi_d^2, \tag{8.2}$$

which makes the individual loss χ_d^2 of different parameters commensurable by an appropriate choice of weightings $\{c_d\}_{d=1}^D$. If one further assumes that no statistical dependencies among the different stimulus components exist, (that is, $p(\mathbf{x}) = \prod_{m=1}^D p(x_m)$, it is easy to extend our results to the case where many parameters, say D, have to be inferred simultaneously from the neuronal population activity.[5] Without assuming special constraints, the optimization problem reduces to the single-parameter case, where optimal encoding is simply given by D subpopulations that encode each parameter independently and the number of neurons in each subpopulation has to be chosen such that the contributions $c_d \chi_d$ to the total loss become equal. The more general case, where statistical dependencies among the variables exist, can often be traced back to the case without correlations, provided the weightings c_d are all identical. For example, if $p(\mathrm{x})$ is given by an arbitrary multivariate normal distribution, for which all correlations are determined by the covariance matrix, one can always find another coordinate system $\tilde{\mathrm{x}}$ by the Karhunen-Loeve transformation (Jollife, 1986), for which all variables become independent ($p(\tilde{\mathbf{x}}) = \prod_{m=1}^D p(\tilde{x}_m)$).

From these considerations, we conclude that the shape of optimal tuning functions is not necessarily related to the number of dimensions. Under specific assumptions, however, dependencies on the number of dimensions can emerge. For example, the result in Zhang and Sejnowski (1999) that decreasing the scale maximizes Fisher information in the case of $D < 2$ and the opposite holds in the case of $D > 2$ is a direct consequence of the

[5] This corresponds to the case considered in Zhang and Sejnowski (1999) and Eurich and Wilke (2000).

restriction to radial symmetric tuning curves and the density approximation of the encoding, by which continuous translation invariance of the coding problem over the whole real axis is enforced artificially.

Although the goal of this work was to understand the general principles of optimal population coding constraining the possibilities of interneuronal signal processing in cortex, one might also relate these results to the shape of measured tuning functions as it has been done in literature. For unimodal tuning functions, which are ubiquitous in the sensory pathways, the Fisher-optimal shape with the smallest critical decoding time is flat and has steep quadratic edges (see Figure 8). For large N, these tuning curves become very similar to boxes standing in remarkable contrast to most measured tuning functions of sensory neurons in mammals. This apparent contradiction suggests essentially two alternative explanations. On the one hand, it is by far not evident whether the boxlike tuning curve is indeed favorable. It has been derived with the use of Fisher information, and we have shown how the optimality of a neural encoding relies on additional constraints. Other constraints from those considered in this article may be necessary to explain the shape of experimentally observed tuning functions. For example, it is likely that one needs to consider wiring constraints and natural limitations of the subsequent neuronal readout that prevent minimum mean square estimation. If in the end, however, the neuronal encoding strategy is determined by the choice of additional constraints rather than by optimal coding, the honest conclusion would be that the principle of efficient coding is of rather little relevance for explaining neuronal response properties.

On the other hand, it is also likely that the quantity x actually encoded by the neuron is somehow correlated with the stimulus feature s considered in the experimental tuning curve. If $x = s + c$, where c stands for an arbitrary contextual quantity that is not under the control of the experimentalist, its variability during the measurement would lead to a smoothed version $f(s)$ of the actual tuning function $f(x)$.

In summary this means that both the uncertainty in the assumptions of constraints, as well as the limited control in experiments, may often counteract the goal to explain measured tuning functions by the method of optimal coding. If we nevertheless speculate about theoretical implications of efficient coding that might be observed experimentally, we would conclude from our investigations that place coding is better than intensity coding (using the terms *place coding* and *intensity coding* as defined in Snippe, 1996) in cases where the constraint of decoding time is stronger than the cost of the required number of neurons. This means that it is advantageous if neurons are activated in an all-or-nothing manner rather than in a smooth graded way. Experimentally, this idea can be supported by the fact that bursting cells, which are activated in a bistable way, are ubiquitous in the brain.

Appendix

A.1 Monte-Carlo Integration. While the evaluation of the integrals determining the MS estimator,

$$\hat{x}(\mathbf{k}) = \frac{\int_0^1 x\, p(\mathbf{k} \mid x)\, dx}{\int_0^1 p(\mathbf{k} \mid x)\, dx}, \tag{A.1}$$

can be computed by classical integration routines, the evaluation of the mean,

$$\mathrm{E}[(\hat{x}(\mathbf{k}) - x)^2] = \int_0^1 p(x) \sum_{k_1=0}^{\infty}, \ldots, \sum_{k_N=0}^{\infty} (\hat{x}(\mathbf{k}) - x)^2\, p(\mathbf{k} \mid x)\, dx, \tag{A.2}$$

requires a summation over an N-dimensional space. According to the Monte Carlo technique (Bishop, 1995), we estimate the value of equation A.2 by an average over n trials, for which we draw a particular $(x, \mathbf{k})_i$ randomly from the joint distribution $p(\mathbf{k} \mid x)p(x)$. Then it holds that

$$\mathrm{E}[(\hat{x}(\mathbf{k}) - x)^2] \approx \frac{1}{n} \sum_{i=1}^{n} (\hat{x}(\mathbf{k}_i) - x_i)^2. \tag{A.3}$$

The error of this approximation drops with the number of trials. We evaluated the right-hand side of equation A.3 up to the second relevant digit. As a termination criterion, we stopped the averaging process when there was no change in the value of the second relevant digit during the last 10,000 trials.

A.2 Fisher Information and the Exponential Family. The mean squared error of any estimator can always be decomposed into its bias $b_{\hat{x}}(x) = (g(x) - x)^2$ and its variance $v_{\hat{x}}(x) = \mathrm{E}[(\hat{x} - g(x))^2 \mid x]$, where we introduced $g(x) = \mathrm{E}[\hat{x} \mid x]$ for the sake of clarity. Accordingly, the Cramér-Rao bound (see equation 2.8) can also be given in the form

$$v_{\hat{x}}(x) \geq \frac{g'(x)^2}{J(x)}. \tag{A.4}$$

For this lower bound, it is known that equality holds if and only if $p(\mathbf{k} \mid x)$ constitutes an exponential family. While the Poisson distribution constitutes an exponential family with respect to the mean spike count $\mu_j = f_j(x)T$ for each neuron j, it depends on the shape of the tuning functions whether

an estimator of x exists, for which equality holds in equation A.4. Such an estimator has to satisfy the following equation (Lehmann & Casella, 1999):

$$\hat{x}(\mathbf{k}) = g(x) + \frac{g'(x)}{J(x)} \partial_x \log p(\mathbf{k} \mid x). \tag{A.5}$$

Therefore, the mean squared error of an estimator is completely determined by Fisher information if it is unbiased (i.e., $g(x) = x$), and the right-hand side of equation A.5 is independent from x, that is,

$$x + \frac{\partial_x \log p(\mathbf{k} \mid x)}{J(x)} = const. \tag{A.6}$$

Inserting equation 2.1 and taking the derivative with respect to x yields

$$\frac{(k - \mu)\mu''}{\mu'^2} = 0, \tag{A.7}$$

in the case of $N = 1$. From this, it follows that μ'' equals zero and hence, the tuning function $f(x) = \frac{1}{T}\mu(x)$ is required to be linear. While we did not solve equation A.6 for $N > 1$, this short calculation may hint at the strong restrictions that it imposes on the shape of the tuning functions.

A.3 Derivation of Equation 5.7. The upper bound results from a calculation of the error of an suboptimal estimator:

$$\hat{x}(\mathbf{k}) := \max\left\{ \frac{1}{N}, \frac{j}{N}\Theta\left(k_j - \frac{1}{2}\right) \middle| j = 1, \ldots, N \right\}. \tag{A.8}$$

We decompose the mean squared error,

$$\begin{aligned}(\chi^2(\alpha))^{ramp} &= \int_0^1 \sum_{\mathbf{k}} (x - \hat{x}(\mathbf{k}))^2 p(\mathbf{k} \mid x)\, dx \\ &= \frac{1}{N} \sum_{a=1}^{N} \underbrace{N \int_{\frac{a-1}{N}}^{\frac{a}{N}} \sum_{\mathbf{k}} (x - \hat{x}(\mathbf{k}))^2 p(\mathbf{k} \mid x)\, dx}_{\chi_a^2},\end{aligned} \tag{A.9}$$

and consider the parts χ_a^2, $a = 1, \ldots, N$ separately:

$$\chi_1^2 = N \int_0^{\frac{1}{N}} \left(x - \frac{1}{N}\right)^2 dx \leq \max_{x \in [0, 1/N]} \left(x - \frac{1}{N}\right)^2 = \frac{1}{N^2} \tag{A.10}$$

$$\chi_2^2 = p(k_2 > 0 \mid x \in [1/N, 2/N]) \int_{\frac{1}{N}}^{\frac{2}{N}} \left(x - \frac{2}{N}\right)^2 dx \tag{A.11}$$

$$+ p(k_2 = 0 \mid x \in [1/N, 2/N]) \int_{\frac{1}{N}}^{\frac{2}{N}} \left(x - \frac{1}{N}\right)^2 dx$$

$$\leq p(k_2 > 0 \mid x \in [1/N, 2/N]) \max_{x \in [1/N, 2/N]} \left(x - \frac{2}{N}\right)^2$$

$$+ p(k_2 = 0 \mid x \in [1/N, 2/N]) \max_{x \in [1/N, 2/N]} \left(x - \frac{1}{N}\right)^2$$

$$\leq p(k_2 > 0 \mid x \in [1/N, 2/N]) \frac{1}{N^2} + p(k_2 = 0 \mid x \in [1/N, 2/N]) \frac{1}{N^2}$$

$$= \frac{1}{N^2}$$

$$\chi_3^2 \leq p(k_2 > 0 \mid x \in [2/N, 3/N]) \int_{\frac{2}{N}}^{\frac{3}{N}} \left(x - \frac{2}{N}\right)^2 dx$$

$$+ p(k_2 = 0 \mid x \in [2/N, 3/N]) \left(\chi_2^2 + \frac{1}{N^2}\right)$$

$$\leq \frac{1}{N^2}(1 + p(k_2 = 0 \mid x \in [2/N, 3/N])). \tag{A.12}$$

In general, it holds for all $a \geq 2$:

$$\chi_{a+1}^2 \leq \frac{1}{N^2} + p(k_a = 0 \mid x \in [a/N, (a+1)/N]) \chi_a^2 = \frac{1}{N^2} + q\, \chi_a^2, \tag{A.13}$$

where we introduced the abbreviation q for the probability $p(k_a = 0 \mid x \in [a/N, (a+1)/N]) = e^{-f_{\max} T}$, which does not depend on the neuron index. Together with $p := 1 - q$, it then follows by induction,

$$\chi_a^2 \leq \frac{1}{N^2} \frac{1 - q^{a-1}}{1 - q} = \frac{1}{N^2} \frac{1 - q^{a-1}}{p}, \tag{A.14}$$

because it holds

$$\chi_{a+1}^2 \overset{Eq.\ A.13}{\leq} \frac{1}{N^2} + q \frac{1}{N^2} \frac{1 - q^{a-1}}{p} = \frac{1}{N^2} \left(\frac{p + q - q^a}{p}\right) = \frac{1}{N^2} \frac{1 - q^a}{p}. \tag{A.15}$$

According to equation A.9, we finally obtain

$$
\begin{aligned}
(\chi^2(\alpha))^{ramp} = \frac{1}{N}\sum_{a=1}^{N}\chi_a^2 &\le \frac{1}{N}\left(\frac{1}{N^2}+\sum_{a=2}^{N}\frac{1}{N^2}\,\frac{1-q^{a-1}}{p}\right)\\
&= \frac{1}{N^3}\left(1+\frac{1}{p}\sum_{a=2}^{N}(1-q^{a-1})\right)\\
&= \frac{1}{N^3}\left(1+\frac{N-1}{p}-\frac{1}{p}\sum_{a=0}^{N-2}q^{a+1}\right)\\
&= \frac{1}{N^3}\left(1+\frac{N-1}{p}-\frac{q}{p}\,\frac{1-q^{N-1}}{1-q}\right)\\
&= \frac{1}{N^2 p}+\frac{1}{N^3}\left(1-\frac{p+q-q^N}{p^2}\right)\\
&= \frac{1}{N^2 p}+\frac{1}{N^3}\left(1-\frac{1-q^N}{p^2}\right). \qquad \text{(A.16)}
\end{aligned}
$$

Acknowledgments

We thank S. Wilke and C. Eurich for fruitful discussions. This work was supported by the DFG (SFB 517, Neurocognition).

References

Aitken, A. C., & Silverstone, H. (1942). On the estimation of statistical parameters. *Proc. Roy. Soc. Edin.*, A, *62*, 369.

Attneave, F. (1954). Informational aspects of visual perception. *Psychological Review, 61*, 183–193.

Baldi, P., & Heiligenberg, W. (1988). How sensory maps could enhance resolution through ordered arrangements of broadly tuned receivers. *Biological Cybernetics, 59*, 313–318.

Barlow, H. B. (1959). Sensory mechanisms, the reduction of redundancy, and intelligence. In *The mechanisation of thought processes* (pp. 535–539). London: Her Majesty's Stationery Office.

Bethge, M., Rotermund, D., & Pawelzik, K. (2002). *Optimal neural rate coding leads to bimodal firing rate distributions*. Manuscript submitted for publication.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. New York: Oxford University Press.

Braitenberg, V., & Schüz A. (1991). *The anatomy of the cortex: Statistics and geometry*. Berlin: Springer-Verlag.

Brunel, N., & Nadal, J.-P. (1998). Mutual information, Fisher information, and population coding. *Neural Computation, 10*, 1731–1757.

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

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

Cramér, H. (1946). A contribution to the theory of statistical estimation. *Aktuariestidskrift, 29,* 458–463.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience.* Cambridge, MA: MIT Press.

Eurich, C. W., & Wilke, S. D. (2000). Multi-dimensional encoding strategy of spiking neurons. *Neural Computation, 12,* 1519–1529.

Frechet, M. (1943). Sur l'extension de certaines evaluations statistiques de petits echantilons. *Rev. Int. Statist., 11,* 182–205.

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science, 233,* 1416–1419.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing* (Vol. 1, pp. 77–109). Cambridge, MA: MIT Press.

Johnson, D. H. (1996). Point process models of single-neuron discharges. *J. Comput. Neurosci., 3,* 275–299.

Jollife, I. T. (1986). *Principal component analysis.* New York: Springer-Verlag.

Lehmann, E. L., & Casella, G. (1999). *Theory of point estimation.* New York: Springer-Verlag.

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

Panzeri, S., Biella, G., Rolls, E. T., Skaggs, W. E., & Treves, A. (1996). Speed, noise, information and the graded nature of neuronal responses. *Network: Comp. in Neural Syst., 7,* 365–370.

Panzeri, S., Treves, A., Schultz, S., & Rolls, E. T. (1999). On decoding the responses of a population of neurons from short time windows. *Neural Computation, 11,* 1553–1577.

Paradiso, M. A. (1988). A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. *Biological Cybernetics, 58,* 35–49.

Rao, C. R. (1946). Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc., 37,* 81–91.

Roddey, J. C., Girish, B., & Miller, J. P. (2000). Assessing the performance of neural encoding models in the presence of noise. *J. Comput. Neurosci., 8,* 95–112.

Rolls, E. T., & Cowey, A. (1970). Topography of the retina and striate cortex and its relationship to visual acuity in rhesus monkeys and squirrel monkeys. *Exp. Brain Research, 10,* 298–310.

Rolls, E. T., & Tovee, M. J. (1994). Processing speed in the cerebral cortex and the neurophysiology of visual masking. *Proc. R. Soc. Lond. B, 257,* 9–15.

Salinas, E., & Abbott, L. F. (1994). Vector reconstruction from firing rates. *J. Comput. Neurosci., 1,* 89–107.

Seung, H. S., & Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *PNAS, 90,* 10749–10753.

Snippe, H. (1996). Parameter extraction from population codes: A critical assessment. *Neural Computation, 8*, 511–529.
Snippe, H., & Koenderink, J. J. (1992). Discrimination thresholds for channel-coded systems. *Biol. Cybern., 66*, 543–551.
Softky, W. R., & Koch, C. (1993). The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *J. Neurosci., 13*(1), 334–350.
Thorpe, S., Fitz, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature (London), 381*, 520–522.
van Vreeswijk, C. (2001). Whence sparseness? In T. Leen, T. Dietterich, & V. Tresp (Eds.), *Neural information processing systems, 13*. Cambridge, MA: MIT Press.
Zhang, K., & Sejnowski, T. J. (1999). Neuronal tuning: To sharpen or broaden? *Neural Computation, 11*, 75–84.

Received December 20, 2000; accepted April 12, 2002.