LETTER **Communicated by Guillaume Dehaene**

Mutual Information, Fisher Information, and Efficient Coding

Xue-Xin Wei
weixxpku@gmail.com
Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104, U.S.A.

Alan A. Stocker
astocker@sas.upenn.edu
Departments of Psychology and Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, U.S.A.

Fisher information is generally believed to represent a lower bound on mutual information (Brunel & Nadal, 1998), a result that is frequently used in the assessment of neural coding efficiency. However, we demonstrate that the relation between these two quantities is more nuanced than previously thought. For example, we find that in the small noise regime, Fisher information actually provides an upper bound on mutual information. Generally our results show that it is more appropriate to consider Fisher information as an approximation rather than a bound on mutual information. We analytically derive the correspondence between the two quantities and the conditions under which the approximation is good. Our results have implications for neural coding theories and the link between neural population coding and psychophysically measurable behavior. Specifically, they allow us to formulate the efficient coding problem of maximizing mutual information between a stimulus variable and the response of a neural population in terms of Fisher information. We derive a signature of efficient coding expressed as the correspondence between the population Fisher information and the distribution of the stimulus variable. The signature is more general than previously proposed solutions that rely on specific assumptions about the neural tuning characteristics. We demonstrate that it can explain measured tuning characteristics of cortical neural populations that do not agree with previous models of efficient coding.

1 Introduction

The efficient coding hypothesis is an important proposal of how neural systems may represent (sensory) information (Barlow, 1961; Attneave, 1954;

Neural Computation **28, 305–326** (2016)
doi:10.1162/NECO_a_00804

Linsker, 1988). Common formulations of efficient coding are based on the assumption that a neural system is adapted to the statistical structure of the environment in a way that the mutual information (Shannon & Weaver, 1949) between the stimulus variable and its neural representation (e.g., as reflected in the firing activity of a neural population) is maximized subject to certain resource constraints. However, the test of this prominent hypothesis is impeded by the fact that mutual information is analytically tractable only for simple coding problems (Laughlin, 1981; Atick, 1992). One way to work around this difficulty is to relate mutual information to Fisher information (Fisher, 1922). For many neural population coding models, Fisher information is relatively easy to compute and interpret with regard to neurophysiological parameters (e.g., neural response gain and dynamic range), as well as psychophysical behavior (e.g., discrimination threshold; Seung & Sompolinsky, 1993; Seriès, Stocker, & Simoncelli, 2009).

In a seminal paper, Brunel and Nadal (1998) argued that Fisher information provides a lower bound on mutual information. This result has been widely applied in various studies aimed at testing the efficient coding hypothesis (Harper & McAlpine, 2004; McDonnell & Stocks, 2008; Wang, Stocker, & Lee, 2012; Ganguli & Simoncelli, 2010, 2014). These studies have derived efficient coding solutions by maximizing the proposed lower bound (in terms of Fisher information) rather than directly maximizing mutual information. This approach, however, can be problematic because recent theoretical and numerical analyses suggest that Fisher information can be an imprecise measure of coding accuracy (Bethge, Rotermund, & Pawelzik, 2002) and may actually represent an upper rather than a lower bound on mutual information (Yarrow, Challis, & Seriès, 2012). What is currently missing is a clear understanding of the conditions under which Fisher information serves as a (lower or upper) bound on mutual information and when this bound is reasonably tight (i.e., Fisher information provides a good proxy for mutual information).

In this letter, we revisit the formal link between Fisher and mutual information. We first reexamine the conditions for which the lower bound proposed by Brunel & Nadal (1998) holds. We show that the derivation of the bound is based on assumptions that make it automatically tight, thus defying the meaning of a bound. We then formally derive the relation between Fisher and mutual information in a standard input-output model under more general conditions. We discuss the possible interpretations of our derivation in terms of both upper and lower bounds on mutual information. We further derive the conditions under which Fisher information provides a good approximation of mutual information. Finally, we discuss the implications of our results in the context of efficient coding. Our results provide an important step toward a more detailed and rigorous understanding of Fisher information as a characteristic measure of neural codes and their efficiency.

2 Examining the Derivation of a Lower Bound on Mutual Information

Brunel and Nadal (1998) derived a lower bound on the mutual information contained in a neural code using Fisher information. Their formulation viewed neural coding as a channel coding problem where an input (stimulus variable) θ is encoded in the output (measurement) m of a noisy channel. Rather than directly computing the mutual information $I[\theta, m]$ between the stimulus variable and the sensory measurement, Brunel and Nadal considered a substitute problem by computing the mutual information between θ and $\hat{\theta}$, where $\hat{\theta}$ is the output of an unbiased efficient estimator with mean θ and variance $1/J(\theta)$, and

$$J(\theta) = \int \left(\frac{\partial \ln p(m|\theta)}{\partial \theta} \right)^2 p(m|\theta) dm \tag{2.1}$$

is the Fisher information of the estimate with regard to the input variable θ. The mutual information between θ and $\hat{\theta}$ then can be written as

$$I[\theta, \hat{\theta}] = H[\hat{\theta}] - \int d\theta\, p(\theta) H[\hat{\theta}|\theta]. \tag{2.2}$$

The moment-entropy inequality (Cover & Thomas, 2012) states that for a continuous random variable with given variance, the Shannon entropy (Shannon & Weaver, 1949) is maximal if and only if the variable is gaussian distributed. Thus, we can consider

$$H[\hat{\theta}|\theta] \leq H[Z] = \frac{1}{2} \ln \left(\frac{2\pi e}{J(\theta)} \right), \tag{2.3}$$

where Z is a gaussian random variable with mean θ and variance $1/J(\theta)$, and rewrite the mutual information as the inequality

$$I[\theta, \hat{\theta}] \geq H[\hat{\theta}] - \int d\theta\, p(\theta) \frac{1}{2} \ln \left(\frac{2\pi e}{J(\theta)} \right). \tag{2.4}$$

Due to the data processing inequality, $I[\theta, m] \geq I[\theta, \hat{\theta}]$. Assuming the asymptotic limit $H[\hat{\theta}] \to H[\theta]$, we finally arrive at

$$I[\theta, m] \geq \underbrace{H[\theta] - \int d\theta\, p(\theta) \frac{1}{2} \ln \left(\frac{2\pi e}{J(\theta)} \right)}_{I_{\text{Fisher}}}, \tag{2.5}$$

which states that Fisher information (i.e., I_{Fisher}) provides a lower bound on mutual information.

2.1 Limitations of the Formulation. Although the derivation of the lower bound (equation 2.5) is technically correct, it relies on assumptions that compromise its interpretation as a bound. First, there is the assumption that an unbiased efficient estimator $\hat{\theta}(m)$ exists. This assumption implies that the noise model must be a member of the exponential family and θ has to be the natural parameter of the particular exponential family (Lehmann & Casella, 1998; Amari & Nagaoka, 2007). Second, the lower bound is derived for the asymptotic limit where any noise model from the exponential family is largely equivalent to a gaussian model (Lehmann & Casella, 1998). These two assumptions essentially require the noise to be small and gaussian, in which case the bound, equation 2.5 is tight and becomes an identity (Clarke & Barron, 1990; Rissanen, 1996). Thus, while the identity is with no doubt an interesting and useful result, the notion that equation 2.5 represents a lower bound on mutual information, however, seems not particularly meaningful. Also, the derivation in Brunel and Nadal (1998) implies that for any nongaussian noise, mutual information would exceed I_{Fisher}. As we will show later, this implication is incorrect. Rather, deviations from a gaussian noise model have exactly the opposite effect.

In sum, Brunel and Nadal (1998) addressed the connection between mutual information and I_{Fisher} only for the limited case of small gaussian noise. It is easy to verify that in this case, the mutual information is identical to I_{Fisher}. What remains unclear is the relation between mutual information and Fisher information for other, more general noise models. This is what we derive in the following and what represents one of the main contributions of the letter.

3 A New Look at the Connection between Mutual Information and Fisher Information

We revisit the formal link between Fisher and mutual information in particular with regard to nongaussian noise models that are often relevant for the assessment of neural coding models. For analytical convenience, we consider a standard one-dimensional input-output model (Laughlin, 1981; Nadal & Parga, 1994; Bell & Sejnowski, 1995) between the sensory variable θ and its neural representation m. More specifically, we assume

$$m = f(\theta) + \delta, \tag{3.1}$$

where θ has a continuous prior distribution $p(\theta)$, $f(\theta)$ is an invertible transfer function that is bounded, and δ represents arbitrary additive noise

with a smooth density $q(.)$. Note that this model is a one-dimensional approximation of the more general neural coding problem (see section 4).

3.1 Stam's Inequality. We first introduce Stam's inequality (Stam, 1959), which is often applied in information theory yet is little known in the neural coding literature. The inequality plays an important role in the derivation of our main result. We begin by reformulating Fisher information with regard to the input-output model, equation 3.1. With $\tilde{\theta} = f(\theta)$, Fisher information with respect to $\tilde{\theta}$ is given as

$$J(\tilde{\theta}) = \int \left(\frac{\partial \ln p(m|\tilde{\theta})}{\partial \tilde{\theta}} \right)^2 p(m|\tilde{\theta}) dm. \tag{3.2}$$

Because we assume additive noise with density $q(.)$, we can write $p(m|\tilde{\theta}) = q(m - \tilde{\theta})$. In this case, $J(\tilde{\theta})$ becomes independent of $\tilde{\theta}$ and thus constant (Stam, 1959) and can be rewritten as

$$J[\delta] := \int \left(\frac{\partial \ln\, q(\delta)}{\partial \delta} \right)^2 q(\delta) d\delta. \tag{3.3}$$

This quantity is referred to as Fisher information of a random variable with respect to a scalar translation parameter (Dembo, Cover, & Thomas, 1991). Note that we use a different notation $J[\delta]$ in order to distinguish it from the standard formulation of Fisher information $J(\tilde{\theta})$. Conceptually, the constant $J[\delta]$ summarizes the total local dispersion of a distribution.

The Shannon entropy $H[m|\tilde{\theta}]$ (Shannon & Weaver, 1949) is also independent of $\tilde{\theta}$ and identical to the noise entropy,

$$H[\delta] = -\int q(\delta) \ln q(\delta) d\delta. \tag{3.4}$$

Stam's inequality specifies the relation between Fisher information $J[\delta]$ and Shannon entropy $H[\delta]$ as the following: For a given amount of Fisher information, the Shannon entropy of a continuous random variable is minimized if and only if the variable is gaussian distributed (Stam, 1959).

Thus with the notation above (see equations 3.3 and 3.4), Stam's inequality implies that

$$H[\delta] \geq \frac{1}{2} \ln \left(\frac{2\pi e}{J[\delta]} \right). \tag{3.5}$$

As a remark, equation 3.5 is equivalent to the isoperimetric inequality for entropies in the information theory literature (Dembo et al., 1991). For

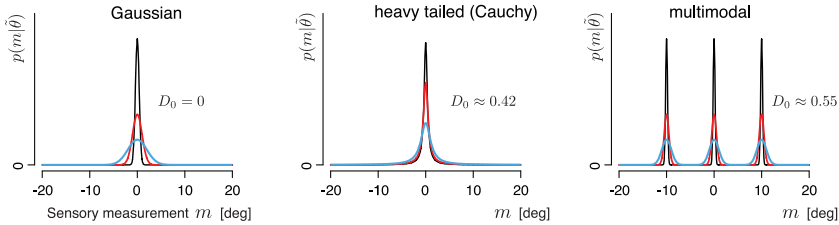

Figure 1: Quantifying the nongaussianity of noise distributions. D_0 (as defined in equation 3.6) provides a measure for the nongaussianity of different noise distributions $p(m|\tilde{\theta})$. For a gaussian distribution, $D_0 = 0$. For a Cauchy distribution that has a heavier tail than a gaussian, $D_0 \approx 0.42$. For a multimodal mixture of N gaussians (same height, peaks well separated), $D_0 \approx \frac{1}{2}\ln(N)$ (here: $D_0 \approx \frac{1}{2}\ln(3) \approx 0.55$). Crucially, D_0 is independent of the width of the distribution (i.e., D_0 has the same value for the black, red, and blue curves).

gaussian distributed δ with variance σ^2, Shannon entropy and Fisher information are $H[\delta] = \frac{1}{2}\ln(2\pi e\sigma^2)$ and $J[\delta] = 1/\sigma^2$, respectively. In this case, equation 3.5 is an equality. By defining

$$D_0 = H[\delta] - \frac{1}{2}\ln\left(\frac{2\pi e}{J[\delta]}\right), \tag{3.6}$$

we obtain a measure of the nongaussianity of the noise distribution (see Figure 1). For example, let us consider the noise to follow a Cauchy distribution with density

$$q(\theta) = \frac{\gamma}{\pi}\frac{1}{\gamma^2 + \theta^2}, \tag{3.7}$$

where γ controls the width of the distribution. For a Cauchy distribution, which exhibits substantially heavier tails than a gaussian,

$$D_0 = H[\delta] - \frac{1}{2}\ln\left(\frac{2\pi e}{J[\delta]}\right) = \frac{1}{2}(\ln(4\pi) - 1) > 0. \tag{3.8}$$

Critically, D_0 is independent of the value of γ and thus provides a generic and, to the best of our knowledge, new measure of nongaussianity largely independent of the magnitude of the noise. This is illustrated in Figure 1 together with another example of a nongaussian noise distribution (multimodal). D_0 provides a measure that is different from those defined on the measured moments of a distribution. Note that $D_0 > 0$ for distributions with lighter tails than a gaussian, as well as for distributions that are asymmetric.

3.2 When Is Fisher Information a Good Approximation of Mutual Information? With the above results, we can now express mutual information in terms of Fisher information for arbitrary noise distributions. Because the transfer function f is invertible, $I[\theta, m] = I[\tilde{\theta}, m]$.[1] Thus, we can write mutual information as

$$I[\tilde{\theta}, m] = H[m] - \int d\tilde{\theta} p(\tilde{\theta}) H[m|\tilde{\theta}]. \tag{3.9}$$

As shown in section 3.1, $H[m|\tilde{\theta}] = H[\delta]$ and thus

$$I[\tilde{\theta}, m] = H[m] - \int d\tilde{\theta} p(\tilde{\theta}) H[\delta]. \tag{3.10}$$

With D_0 representing the entropy difference between the noise δ and a gaussian with the same amount of Fisher information $J[\delta]$, equation 3.6, we can rewrite, generally equation 3.10 as

$$I[\tilde{\theta}, m] = H[m] - \int d\tilde{\theta} p(\tilde{\theta}) \left(\frac{1}{2} \ln \left(\frac{2\pi e}{J[\delta]} \right) + D_0 \right). \tag{3.11}$$

Because $J[\delta] = J(\tilde{\theta})$ (see section 3.1) we replace $J[\delta]$ with $J(\tilde{\theta})$ in equation 3.11 and obtain

$$\begin{aligned} I[\tilde{\theta}, m] &= H[m] - \int d\tilde{\theta} p(\tilde{\theta}) \left(\frac{1}{2} \ln \left(\frac{2\pi e}{J(\tilde{\theta})} \right) + D_0 \right) \\ &= H[m] - \int d\tilde{\theta} p(\tilde{\theta}) \frac{1}{2} \ln \left(\frac{2\pi e}{J(\tilde{\theta})} \right) - \int d\tilde{\theta} p(\tilde{\theta}) D_0 \\ &= (H[m] - H[\tilde{\theta}]) + H[\tilde{\theta}] - \int d\tilde{\theta} p(\tilde{\theta}) \frac{1}{2} \ln \left(\frac{2\pi e}{J(\tilde{\theta})} \right) - D_0. \end{aligned} \tag{3.12}$$

Using the formulas for a change of variables, it is straightforward to verify that

$$H[\tilde{\theta}] - \int d\tilde{\theta} p(\tilde{\theta}) \frac{1}{2} \ln \left(\frac{2\pi e}{J(\tilde{\theta})} \right) = H[\theta] - \int d\theta p(\theta) \frac{1}{2} \ln \left(\frac{2\pi e}{J(\theta)} \right). \tag{3.13}$$

[1] Although $H[\theta] \neq H[f(\theta)]$.

Thus we can rewrite equation 3.12 as

$$\begin{aligned} I[\theta, m] &= I[\tilde{\theta}, m] \\ &= (H[m] - H[f(\theta)]) + H[\theta] - \int d\theta\, p(\theta) \frac{1}{2} \ln\left(\frac{2\pi e}{J(\theta)}\right) - D_0. \end{aligned} \tag{3.14}$$

Finally, with the definition of $C_0 = H[m] - H[f(\theta)]$ we arrive at the following expression for mutual information:

$$I[\theta, m] = \underbrace{H[\theta] - \int d\theta\, p(\theta) \frac{1}{2} \ln\left(\frac{2\pi e}{J(\theta)}\right)}_{I_{\text{Fisher}}} + C_0 - D_0. \tag{3.15}$$

Equation 3.15 is a general formulation of the relation between mutual information and Fisher information and constitutes one of the main results of the letter. It illustrates that the degree to which mutual information is well approximated by Fisher information (I_{Fisher}) depends on the relative magnitudes of C_0 and D_0. Both terms are nonnegative and quantify two very different aspects of the noise: C_0 is monotonic in the magnitude of the noise (i.e., $H[\tilde{\theta} + \delta] - H[\tilde{\theta}]$), while D_0 represents the nongaussianity of the noise. Equation 3.15 also shows that the interpretation of I_{Fisher} as a bound on mutual information critically depends on the magnitudes of the two constants C_0 and D_0, as we will discuss in the following:

- Lower bound on mutual information. On one hand, if the noise is gaussian, $D_0 = 0$. And $C_0 \geq 0$ because adding additive noise cannot decrease the entropy. Therefore,

$$I[\theta, m] \geq H[\theta] - \int d\theta\, p(\theta) \frac{1}{2} \ln\left(\frac{2\pi e}{J(\theta)}\right). \tag{3.16}$$

  That is, if and only if the noise is gaussian, I_{Fisher} is guaranteed to represent a lower bound on mutual information.
- Upper bound on mutual information. On the other hand, because Stam's inequality tells us that $D_0 \geq 0$, the first three terms on the right-hand side of equation 3.15 provide an upper bound on mutual information; thus,

$$I[\theta, m] \leq H[\theta] - \int d\theta\, p(\theta) \frac{1}{2} \ln\left(\frac{2\pi e}{J(\theta)}\right) + C_0. \tag{3.17}$$

  In particular, in the case of vanishing noise, $H[\delta] \to 0$, and thus $C_0 \to 0$, and it follows that

$$I[\theta, m] \leq H[\theta] - \int d\theta\, p(\theta) \frac{1}{2} \ln\left(\frac{2\pi e}{J(\theta)}\right). \tag{3.18}$$

As a result, I_{Fisher} generally represents an upper bound on mutual information in the small noise regime, which is exactly the opposite of what Brunel and Nadal (1998) postulated.

- Exact approximation. Only in the case where the noise entropy goes to zero and the noise converges to a gaussian at the same time, both C_0 and D_0 converge to zero and make the approximation exact; thus,

$$I[\theta, m] = \underbrace{H[\theta] - \int d\theta\, p(\theta) \frac{1}{2} \ln\left(\frac{2\pi e}{J(\theta)}\right)}_{I_{\text{Fisher}}} . \tag{3.19}$$

This small gaussian noise regime is what Brunel and Nadal (1998) assumed in their derivation.

In sum, our analysis paints a more detailed picture of the conditions under which Fisher information serves as a proxy for mutual information. In general, I_{Fisher} is a good approximation of mutual information only if the difference $(C_0 - D_0)$ is small. In particular, with regard to neural coding applications, there are no simple assumptions and rules that would guarantee that this is generally the case other than the obvious assumption that the noise is small ($C_0 \to 0$) and gaussian ($D_0 = 0$). Also, whether I_{Fisher} represents an upper or lower bound on mutual information is determined by the relative magnitudes of C_0 and D_0. It is important to note that even when the noise vanishes, that is, $C_0 \to 0$, D_0 can still be large and positive. This is the case for noise distributions that exhibit either heavier or lighter tails than a gaussian, such as the Cauchy noise we discussed above (see Figure 1). Thus, assuming a small noise regime does not guarantee that I_{Fisher} well approximates mutual information. Finally, it is worth noting that we do not make any assumption about the shape of the additive noise. The noise can be asymmetric (unlike the three examples shown in Figure 1), in which case m represents a biased measurement of the stimulus value. Any noise asymmetry would be reflected in an increase in nongaussianity as quantified by D_0.

4 Implications for Neural Coding Models

In this section, we discuss the implications of our results for models of neural coding. Our derivation above was based on a standard input-output model where we assumed that both the input and the output variable are one-dimensional and the noise is additive. Neural coding models, however, frequently address the case where a one-dimensional stimulus variable θ is represented in the high-dimensional activity vector R of a population of noisy neurons whose individual response noise is not additive (e.g., Poisson distributed). Thus technically, computing the mutual information $I[\theta, R]$ is not precisely the same problem we have already addressed. However, we

can approximate the problem within our framework by formulating mutual information for a quantity that is the projection from R back to a invertible map of the stimulus space (given by the function $f(.)$, equation 3.1). Denoting the image of such projection as m, we can consider the quantity $I[\theta, m]$ a surrogate of $I[\theta, R]$. Although it is likely that the projection results in some loss of information, the model in equation 3.1 could still provide a good and tractable approximation of the more complicated neural population coding model. This is particularly true if the projection is such that m preserves most of the information in R about θ, and the noise of the projection is approximately additive.

The noise in m can be thought of as the "effective noise", which summarizes the noise characteristics of the whole neural population with regard to the stimulus dimension (Rieke, Bodnar & Bialek, 1995). It is important to emphasize that despite the fact that the noise of individual neurons is often nonadditive, the effective noise can be approximately additive. Our derivation is general for any invertible $f(.)$, and often there exists a mapping $f(.)$ for which Fisher information is uniform. Uniform Fisher information implies that the noise is additive to a first-order approximation. Note that previous studies have used similar surrogate formulations of mutual information by considering m a particular estimator of θ, yet without invoking the mapping $f(.)$ (see Bialek, Rieke, de Ruyter van Steveninck, & Warland, 1991; Rieke et al., 1995; Brunel & Nadal, 1998).

4.1 Information Measures of Neural Codes. Our theoretical predictions based on the input-output model are supported by recent simulation results. Yarrow et al. (2012) numerically computed the mutual information as well as the Fisher information in the response of a population of neurons with bell-shaped tuning curves. They varied population size and levels of response variability. Figure 2 depicts one of their simulation results for gaussian neural noise with different Fano factors and different population sizes. The results show that under these conditions, I_{Fisher} consistently overestimates mutual information. This supports our finding that Fisher information generally does not provide a lower bound on mutual information. Only as the population size or the integration time t increases does the effective noise become small but also more gaussian, and Fisher information serves as an accurate proxy for mutual information as we would predict.

In general, our analysis suggests that in order to assume I_{Fisher}, as in equation 2.5, a good approximation for the amount of information conveyed in a neural code, one should first examine the underlying noise characteristics and confirm that they are close to gaussian. Deviations from gaussianity can result in a severe overestimation of mutual information, even in the small noise regime. This is especially important when studying neural codes based on multimodal tuning curves, such as the code used by grid cells (Hafting, Fyhn, Molden, Moser & Moser, 2005). Fisher information has been a popular quantity to analyze the code of the grid cell system.

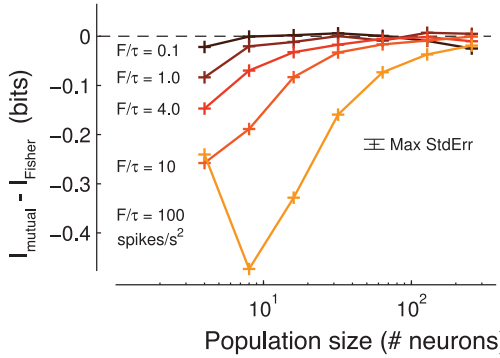

Figure 2: Fisher information generally overestimates mutual information. Shown is the difference between the numerically computed mutual information $I[\theta, R]$ and the Fisher information I_{Fisher} (as in equation 2.5) for the simulated responses of a population of neurons (replotted from Yarrow et al., 2012). The spike counts of individual neurons were assumed to follow a gaussian truncated at zero (to ensure positive firing rates), with variance proportional to the mean (Fano factor F). Individual curves show the difference between mutual and Fisher information as a function of the population size and different ratios between Fano factor and integration time τ. As the population size is small and the "effective noise" (noise on m) is large and nongaussian, Fisher information significantly overestimates mutual information.

For example, it has been argued that the grid cell code has exponentially large capacity because Fisher information can grow exponentially with the number of neurons (Sreenivasan & Fiete, 2011; Mathis, Herz, & Stemmler, 2012). Our results suggest that statements about the coding capacity based on measures of Fisher information may be misleading. Fisher information of a neural code can be arbitrarily large without changing its mutual information if the effective noise in the neural representation is very nongaussian. Importantly, the mismatch between Fisher and mutual information can be substantial even when assuming large populations or vanishing noise. Neurons that exhibit multimodal tuning curves are not uncommon and, besides grid cells, also include disparity-tuned neurons in primary visual cortex (Cumming & Parker, 2000; Fleet, Wagner, & Heeger, 1996) or ITD tuned neurons in owls (*e.g.*, Carr & Konishi, 1990). Equation 3.15 provides a way to directly quantify the amount of overestimation of the mutual information and, thus, determine the conditions under which Fisher information is a good approximation of mutual information.

5 Efficient Coding Interpretation

Efficient coding models are quite often formulated in terms of maximizing mutual information (Barlow, 1961; Linsker, 1988). In the following, we show

how we can rewrite equation 3.15 such that it provides an intuitive interpretation of this optimization problem in terms of the population Fisher information $J(\theta)$. Specifically, we exploit the fact that

$$\begin{aligned} H[\theta] - \int d\theta\, p(\theta) \frac{1}{2} \ln\left(\frac{2\pi e}{J(\theta)}\right) \\ = \int d\theta\, p(\theta) \left(-\ln p(\theta) - \frac{1}{2}\ln\left(\frac{2\pi e}{J(\theta)}\right)\right) \\ = \ln\left(\frac{\int \sqrt{J(\theta)}d\theta}{\sqrt{2\pi e}}\right) - \int d\theta\, p(\theta) \ln\left(\frac{p(\theta)}{\frac{1}{S}\sqrt{J(\theta)}}\right) \\ = \ln\left(\frac{\int \sqrt{J(\theta)}d\theta}{\sqrt{2\pi e}}\right) - KL(p(\theta)||\frac{1}{S}\sqrt{J(\theta)}), \end{aligned} \tag{5.1}$$

where $KL(.)$ is the Kullback-Leibler divergence between two probability distributions (Kullback & Leibler, 1951), and $S = \int \sqrt{J(\theta)}d\theta$ plays the role of a normalization constant that ensures that $\frac{1}{S}\sqrt{J(\theta)}$ is a proper probability density. With equation 5.1, we can express mutual information, equation 3.15 in terms of four meaningful and intuitive components:

$$I[\theta, m] = \frac{1}{2}\ln\left(\frac{S^2}{2\pi e}\right) - KL(p(\theta)||\frac{1}{S}\sqrt{J(\theta)}) - D_0 + C_0. \tag{5.2}$$

The first term can be interpreted as the total available coding capacity (in units of Fisher information). It is independent of the input distribution, implying that it is an intrinsic property of the system. The second term, $KL(p(\theta)||\frac{1}{S}\sqrt{J(\theta)})$, characterizes the information loss due to the mismatch between the input distribution and the way the coding resources are distributed. The third term, D_0, evaluates the information loss due to nongaussianity of the effective noise, and the fourth term, C_0, is quantifying the overall magnitude (entropy) of the effective noise.

5.1 Maximizing Mutual Information. Unlike previous studies (see Harper & McAlpine, 2004; Ganguli & Simoncelli, 2014) we do not attempt to first reexpress mutual information in terms of a specific neural population model and then maximize with regard to the model's neural tuning parameters. Instead, we directly approach the efficient coding problem at the more abstract yet also more general level of the Fisher information $J(\theta)$ of a neural population, which is not subject to the implicit assumptions imposed by any particular neural model (e.g., constraints on neural density, tuning curve shapes, tuning widths). Our goal is to derive general constraints and characteristics of an efficient neural representation expressed in terms of its Fisher information and noise characteristics.

Vanishing noise regime. We start with the above expression for mutual information, equation 5.2. We first consider the condition when the noise is vanishing, that is, the entropy of the noise $H[\delta] \to 0$. Because $C_0 = H[m] - H[f(\theta)] \to 0$,

$$I[\theta, m] = \frac{1}{2} \ln \left(\frac{S^2}{2\pi e} \right) - KL(p(\theta)||\frac{1}{S}\sqrt{J(\theta)}) - D_0. \tag{5.3}$$

It is easy to see that maximizing mutual information must be subject to a constraint on the overall coding capacity available (mutual information must be bound from above). We consider the total Fisher information $S = \int \sqrt{J(\theta)}d\theta$ to manifest the total capacity of the code because it is proportional to the number of discriminable states the code permits.[2] The central question then is how to maximize mutual information subject to the constraint that the coding resources are limited:

$$\int \sqrt{J(\theta)} \leq C. \tag{5.4}$$

To maximize $I[\theta, m]$ with respect to this constraint, it is necessary that both the $KL(.)$ term and D_0 are zero. This provides two necessary conditions for an efficient neural code. First, in order to minimize the KL divergence $p(\theta) = \frac{1}{S}\sqrt{J(\theta)}$, that is, the neural system should distribute its total available coding resources according to the input distribution. This can be viewed as a probabilistic reformulation of histogram equalization (Laughlin, 1981) with the important difference that what is equalized is not firing rates of neurons but rather the square root of Fisher information. Using Fisher information has the advantage that we can formulate efficient coding solutions without being limited to specific neural coding characteristics (tuning curves). Second, in order to minimize D_0, an efficient neural representation should exhibit an effective noise characteristics (noise in m) that is as close as possible to gaussian.

Nonvanishing noise regime. When the noise is large, C_0 is nonzero. However, if C_0 does not depend on any of the other terms on the right-hand side of equation 5.2, then, again, an efficient representation is one whose Fisher information (square root) matches the input distribution and whose noise is gaussian. Generally we find that the form of the transfer function f may slightly change the difference between $H[m]$ and $H[f(\theta)]$, and therefore C_0, because of boundary effects induced by the limited output space. In this case, the relation described above, $p(\theta) = \frac{1}{S}\sqrt{J(\theta)}$, is no longer guaranteed

[2]Note that S is invariant with respect to any reparameterization of θ. This is a desirable property of the constraint because the objective function $I[\theta, m]$ itself is invariant (the amount of information does not dependent on the parameterization of θ).

to be exact. The dependence, however, is typically weak for noise that is not very large.

5.1.1 Signatures of Efficient neural representations. Based on the above derivation, we can identify two characteristic signatures of a neural system that efficiently represents sensory information according to mutual information. The first is that the system's coding resources are allocated such that

$$p(\theta) \propto \sqrt{J(\theta)}. \tag{5.5}$$

This simple relation is reminiscent of the optimal input distribution for a given noise channel in statistics (Clarke & Barron, 1990; Rissanen, 1996). With this signature, we can probe the efficient coding hypothesis by computing the Fisher information of an entire neural population based on electrophysiological measurements and then compare this distribution to the input (stimulus) distribution.

The second signature is with regard to perceptual behavior. More specifically, we can establish a direct link between Fisher information $J(\theta)$, the stimulus (input) distribution $p(\theta)$, and perceptual discrimination threshold $d(\theta)$. It has been shown that the inverse of the square root of Fisher information provides a lower bound on the discrimination threshold for unbiased (Seung & Sompolinsky, 1993) or biased (Seriès et al., 2009) estimators. The bound is tight if the noise is gaussian, which is one of the conditions we identified above to maximize mutual information. Thus, if the encoding is efficient, the discrimination threshold $d(\theta)$ is determined as

$$d(\theta) \propto \frac{1}{p(\theta)}, \tag{5.6}$$

that is, the discrimination threshold should be directly inversely proportional to the stimulus distribution. Again, we expect this relation to hold regardless of the specific neural tuning characteristics. As we show later in Figure 5d, this prediction is well in line with measured discrimination thresholds for perceived heading direction.

5.2 Implications for Neural Models of Efficient Coding. Formulating the efficient coding problem at the level of Fisher information has allowed us to define general signatures of an efficient neural representation that are independent of a specific neural model. This is quite different from previous work that has typically defined the efficient coding problem in terms of specific neural tuning parameters such as neural density or tuning curve shapes (Laughlin, 1981; Harper & McAlpine, 2004; McDonnell & Stocks, 2008; Wang et al., 2012; Ganguli & Simoncelli, 2010, 2014). The advantage

of our approach is that it provides a characterization and understanding of an efficient neural representation that is not limited to the relatively narrow parameter space imposed by considering a specific neural model. Although the neural solutions proposed in these previous studies typically satisfy our signature (equation 5.5), the specific solutions can be quite different.

For example, in a recent study, Ganguli & Simoncelli (2014) proposed that the neural density of an efficient neural population—the distribution of the neurons' preferred tuning values—should match the stimulus distribution $p(\theta)$. This solution is illustrated in Figure 3b, which shows the tuning curves of such a population for the particular stimulus distribution in Figure 3a. However, this characteristic feature is a consequence of the implicit assumptions imposed by the chosen neural model—in this case, the fact that the model allows the tuning curves to be arbitrarily distorted copies along the stimulus dimension of a generic unimodal, symmetric tuning curve. If we impose a different constraint, for example, that all tuning curves are relatively wide and shifted copies of each other, then the solution looks very different. In this case, the neural density is highest at the troughs of the distribution, thus showing exactly the opposite characteristic (see Figure 3c). Yet both populations have the same Fisher information (up to a scale factor), as shown in Figures 3d and 3e, and thus match both of the signatures we identified above.

The power of formulating the efficient coding problem at the more abstract level of Fisher information is that we were able to identify the key features of an efficient code independent of the specific characteristics imposed by a particular neural model. This covers a much larger parameter space than previous approaches that formulated the optimization problem with regard to particular neural constraints (e.g., total firing rate, tuning widths limits). The signature, equation 5.5, can be thought of as providing a manifold in the space of all the possible neural configurations (see Figure 4).

The new formulation has concrete applications in understanding some of the measured tuning properties of neurons in the primate visual pathway. It may explain why some of the known neural density distributions for sensory variables match the encoding accuracy of these variables (e.g., orientation and spatial frequency tuning in primary visual cortex; Ganguli & Simoncelli, 2010) while others do not. One example for a mismatch is the neural representation of heading direction. Neurons in area MST of the macaque monkey are tuned for heading directions. Figure 5a shows the histogram over the measured preferred heading directions for a large pool of MST neurons (replotted from Gu, Fetsch, Adeyemo, DeAngelis, & Angelaki, 2010). More neurons are tuned to lateral directions while the population Fisher information (see Figure 5b) is maximal for forward- and backward-heading directions. This is consistent with our derived signature, equation 5.5, as demonstrated in Figure 3b. Although measurements of the distribution of heading directions for a behaving primate do not exist, we can predict this distribution based on the measured Fisher information of

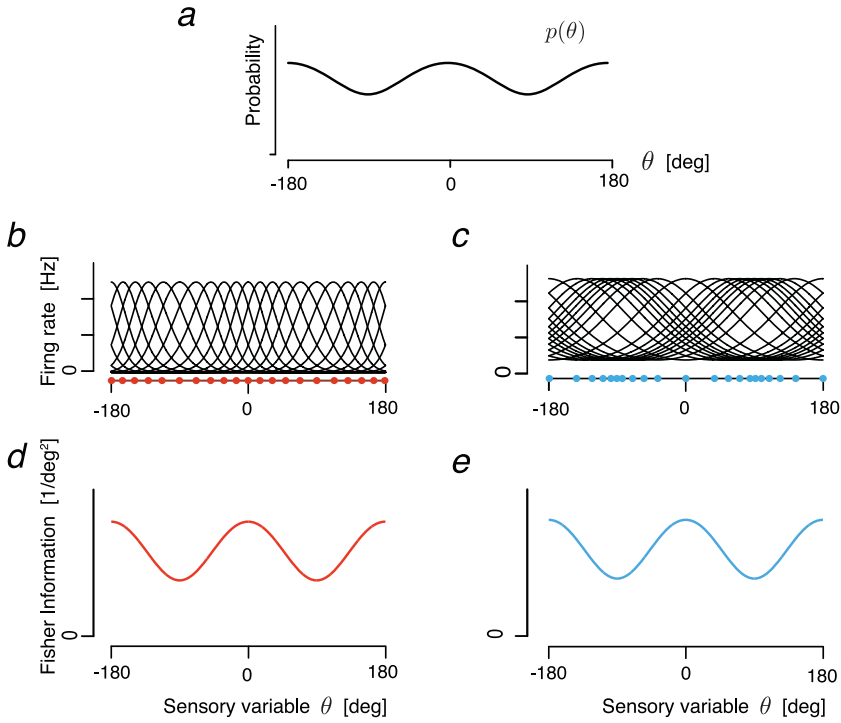

Figure 3: Different population tuning solutions can represent equivalent efficient coding solutions. (a) A stimulus θ with arbitrary distribution $p(\theta)$. (b, c) The tuning curves of two neural populations that maximize mutual information between the stimulus and the population response. Each neuron's preferred tuning is indicated with a dot. The population shown in panel b represents the Infomax solution proposed by Ganguli and Simoncelli (2014). It exhibits the highest neural density at the peaks of the stimulus distribution. The population in panel c is constrained by assuming that all tuning curves are relatively wide and are identical in shape (shifted copies). The neural density of this population has its peaks at $\pm$ 90 degrees, and thus at the troughs of the stimulus distribution. (d, e) Despite this striking difference, the Fisher information is identical for both populations (up to a scale factor) and proportional to the square of $p(\theta)$. Thus, both neural populations can be considered equivalent efficient coding solutions for the same stimulus distribution, each subject to different constraints in terms of the specific tuning parameters.

the MST neural population as shown in Figure 5c. Other examples that demonstrate a mismatch between Fisher information and measured neural population density have also been reported (Fitzpatrick, Batra, Stanford, & Kuwada, 1997; Harper & McAlpine, 2004; Stecker, Harrington, &

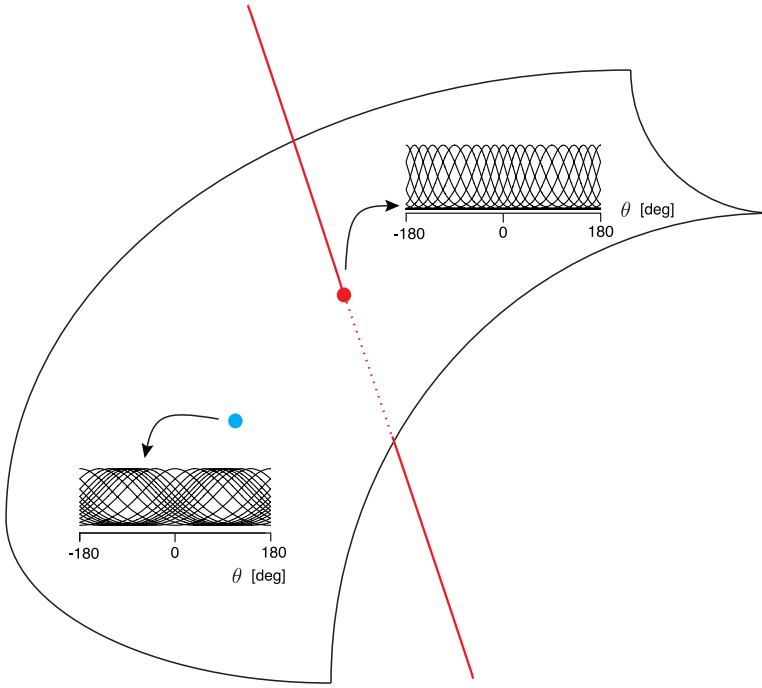

Figure 4: Optimal manifold of efficient coding solutions. The constraint we impose on the overall coding capacity, $S \leq C$, results in the necessary condition $\sqrt{J(\theta)} \propto p(\theta)$. This condition defines an optimal manifold of efficient coding solutions in the space of all possible neural configurations. Previous approaches mainly assumed a parametric description of a neural model first and then solved the optimization problem with regard to these particular tuning parameters. Those solutions start from a much lower-dimensional subspace of all possible neural configurations, as indicated by the red line. The optimal solution (red dot) is part of the manifold. However, because of the specific assumptions of the chosen neural model, this approach cannot account for solutions that are subject to different neural constraints (blue dot). See also Figure 3.

Middlebrooks, 2005). Thus our approach may help to reconcile the different interpretations of these observed tuning characteristics with regard to coding efficiency.

6 Discussion

We have revisited and clarified the relation between Fisher information and mutual information in the context of neural coding. We have derived a new result that describes a more general connection between Fisher and

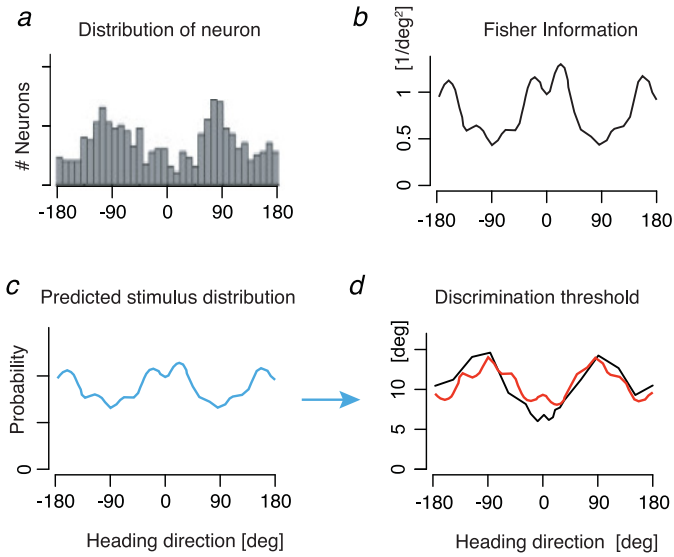

Figure 5: Encoding of heading direction by neurons in area MST of the macaque. (a) Distribution of measured preferred heading directions of MST neurons. More neurons are tuned for lateral rather than back-and-forth directions. (b) Population Fisher information, however, shows peaks at forward- and backward-heading directions, indicating that those directions are most accurately encoded. (c) Based on our signature of efficient coding, equation 5.5, we predict the distribution of heading direction to follow the square root of Fisher information (see panel b). While exact statistical measurements for the distribution of heading direction are missing, the prediction suggests that forward and backward headings are most frequent, which seems to be in agreement with everyday observations. (d) Psychophysically measured discrimination thresholds for heading direction (black curve) nicely reflect the predicted discriminability based on the stimulus distribution (red curve), which represents our second signature. Data in panels a, b, and d are replotted from Gu et al. (2010) (vestibular signals only).

mutual information. In particular, we show that Fisher information (I_{Fisher}, equation 2.5) does not necessarily constitute a lower bound, as has been frequently assumed based on the derivation by Brunel and Nadal (1998). Rather, we have found that the relation between mutual information and Fisher information is more nuanced. It can be a lower bound but only if the (effective) noise is gaussian. However, in the vanishing noise regime (asymptotic small noise limit), Fisher information actually provides an upper bound on mutual information.

If the noise is both gaussian and small, then Fisher information (I_{Fisher}) provides a tight bound for mutual information. It is important to note that by just assuming small noise, it is not guaranteed that mutual information

is actually well approximated by Fisher information. The key coding characteristics that determine by how much Fisher information overestimates mutual information is the degree to which the noise is gaussian. Previous numerical analyses support our result, showing that I_{Fisher} is generally larger than mutual information (Yarrow et al., 2012) and that Fisher information of a single neuron can be made arbitrary high for a fixed positive mean squared error (Bethge et al., 2002).

We emphasize that our analytical results are based on a standard one-dimensional nonlinear input-output model with additive noise. It is important to note that the noise in this model represents the effective noise, that is, the one-dimensional noise component represented in the measurement m, and does not directly refer to the underlying noise characteristics of individual neurons. Since our derivation holds for any invertible $f(.)$ there often, yet not always, exists a mapping for which the effective noise is approximately additive. Despite its limitations in precisely reflecting the neural coding problem, the simple input-output model has proven to be a useful approximation of the more detailed formulation at the neural level (Nadal & Parga, 1994; Bell & Sejnowski, 1995). Nonetheless, an interesting future avenue to explore is to try to characterize the relation between mutual information and Fisher information under more general conditions. This includes nonadditive noise conditions as well as n-dimensional input-output models, which likely require a generalization of Stam's inequality to the multidimensional case (Lutwak, Lv, Yang, & Zhang, 2012).

Our revised derivation of the relation between Fisher information and mutual information has a profound impact on the assessment of neural codes and theories of efficient coding. If indeed I_{Fisher} was assumed to strictly provide a lower bound on mutual information, measures that increased Fisher information of a neural code would automatically signal an increase in coding capacity. This assumption is incorrect and could give rise to incorrect proposals about how a neural system could increase its coding capacity. We argue that the focus should be shifted toward measures that make the bound tight rather than measures that assume that the bound could be raised. This is an important conceptual difference.

Our results also suggest that the total Fisher information $S = \int \sqrt{J(\theta)}d\theta$ provides a more meaningful bound on mutual information than I_{Fisher}.[3] It is an upper bound in the small noise regime, representing the total coding capacity. With this new formulation, we were able to identify two necessary constraints for an efficient code. First, the square root of the Fisher information has to match the input distribution. And second, the effective noise should be gaussian distributed.

In addition to the prediction of discrimination thresholds as described above, we have recently demonstrated that in combination with a Bayesian

[3]To be more precise: $\frac{1}{2}\ln(\frac{S^2}{2\pi e})$. See equation 5.2.

decoder, our formulation of an efficient sensory representation also allows us to make predictions for perceptual biases (Wei & Stocker, 2015). More specifically, we proposed a new Bayesian observer model that is constrained by assuming that the sensory information is efficiently represented as specified by the first signature, equation 5.5. The current work allows us to further refine this observer model by showing that for an efficient code, the effective noise should be gaussian. In this context, our work also has important implications for modeling perceptual behavior.

Finally, these results may provide an explanation for some of the reported differences in the coding strategies of biological neural systems. Formulated with regard to Fisher information, we can specify multiple equivalent efficient coding solutions for neural representations that yet are severely different in terms of their underlying neural tuning characteristics such as their neural density. This allows us to explain correlations between neural density and stimulus distribution that are inconsistent with previously proposed theories of efficient coding that were directly formulated at the level of the neural tuning characteristics. An exciting line of future research will be to understand in detail what additional constraints favor one solution over the others.

Acknowledgments

We thank Zhuo Wang for helpful comments on the mathematical exposition of this letter. Also, we thank the anonymous reviewer(s) for their input and suggestions. The work has been partially supported by ONR grant N000141110744.

References

Amari, S.-I., & Nagaoka, H. (2007). *Methods of information geometry*. Vol. 191, *Translations of mathematical monographs*. Providence, RI: American Mathematical Society.

Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems*, *3*(2), 213–251.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*(3), 183.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129–1159.

Bethge, M., Rotermund, D., & Pawelzik, K. (2002). Optimal short-term population coding: When Fisher information fails. *Neural Computation*, *14*(10), 2317–2351.

Bialek, W., Rieke, F., de Ruyter van Steveninck, R., & Warland, D. (1991). Reading a neural code. *Science*, *252*(5014), 1854–1857.

Brunel, N., & Nadal, J.-P. (1998). Mutual information, Fisher information, and population coding. *Neural Computation*, *10*(7), 1731–1757.

Carr, C., & Konishi, M. (1990). A circuit for detection of interaural time differences in the brain stem of the barn owl. *Journal of Neuroscience*, *10*(10), 3227–3246.

Clarke, B. S., & Barron, A. R. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, *36*(3), 453–471.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. Hoboken, NJ: Wiley Sons.

Cumming, B. G., & Parker, A. J. (2000). Local disparity not perceived depth is signaled by binocular neurons in cortical area V1 of the macaque. *Journal of Neuroscience*, *20*(12), 4758–4767.

Dembo, A., Cover, T. M., & Thomas, J. A. (1991). Information theoretic inequalities. *IEEE Transactions on Information Theory*, *37*(6), 1501–1518.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*, 309–368.

Fitzpatrick, D. C., Batra, R., Stanford, T. R., & Kuwada, S. (1997). A neuronal population code for sound localization. *Nature*, *388*(6645), 871–874.

Fleet, D. J., Wagner, H., & Heeger, D. J. (1996). Neural encoding of binocular disparity: Energy models, position shifts and phase shifts. *Vision Research*, *36*(12), 1839–1857.

Ganguli, D., & Simoncelli, E. P. (2010). Implicit encoding of prior probabilities in optimal neural populations. In J. Lafferty, C. Williams, R. Zemel, J. Shawe-Taylor, & A. Culotta (Eds.), *Advances in neural information processing systems*, *23* (pp. 658–666). Cambridge, MA: MIT Press.

Ganguli, D., & Simoncelli, E. P. (2014). Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. *Neural Computation*, *26*(10), 2103–2134.

Gu, Y., Fetsch, C. R., Adeyemo, B., DeAngelis, G. C., & Angelaki, D. E. (2010). Decoding of MSTd population activity accounts for variations in the precision of heading perception. *Neuron*, *66*(4), 596–609.

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801–806.

Harper, N. S., & McAlpine, D. (2004). Optimal neural population coding of an auditory spatial cue. *Nature*, *430*(7000), 682–686.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*(1), 79–86.

Laughlin, S. B. (1981). A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch*, *36*(910–912), 51.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation*. New York: Springer.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, *21*(3), 105–117.

Lutwak, E., Lv, S., Yang, D., & Zhang, G. (2012). Extensions of Fisher information and Stam's inequality. *IEEE Transactions on Information Theory*, *58*(3), 1319–1327.

Mathis, A., Herz, A. V., & Stemmler, M. (2012). Optimal population codes for space: Grid cells outperform place cells. *Neural Computation*, *24*(9), 2280–2317.

McDonnell, M. D., & Stocks, N. G. (2008). Maximally informative stimuli and tuning curves for sigmoidal rate-coding neurons and populations. *Physical Review Letters*, *101*(5), 058103.

Nadal, J.-P., & Parga, N. (1994). Nonlinear neurons in the low-noise limit: A factorial code maximizes information transfer. *Network: Computation in Neural Systems*, *5*(4), 565–581.

Rieke, F., Bodnar, D., & Bialek, W. (1995). Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *262*(1365), 259–265.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, *42*(1), 40–47.

Seriès, P., Stocker, A. A., & Simoncelli, E. P. (2009). Is the homunculus "aware" of sensory adaptation? *Neural Computation*, *21*(12), 3271–3304.

Seung, H. S., & Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences*, *90*(22), 10749–10753.

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

Sreenivasan, S., & Fiete, I. (2011). Grid cells generate an analog error-correcting code for singularly precise neural computation. *Nature Neuroscience*, *14*(10), 1330–1337.

Stam, A. (1959). Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, *2*(2), 101–112.

Stecker, G. C., Harrington, I. A., & Middlebrooks, J. C. (2005). Location coding by opponent neural populations in the auditory cortex. *PLoS Biology*, *3*(3), e78.

Wang, Z., Stocker, A. A., & Lee, D. D. (2012). Optimal neural tuning curves for arbitrary stimulus distributions: Discrimax, infomax and minimum L_p loss. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, 25* (pp. 2177–2185). Cambridge, MA: MIT Press.

Wei, X.-X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain "anti-Bayesian" percepts. *Nature Neuroscience*, *18*(10), 1509–1517.

Yarrow, S., Challis, E., & Seriès, P. (2012). Fisher and Shannon information in finite neural populations. *Neural Computation*, *24*(7), 1740–1780.

Received February 16, 2015; accepted October 12, 2015.