LETTER Communicated by Ron Meir

Efficient Neural Codes That Minimize L_p Reconstruction Error

Zhuo Wang
wangzhuo@nyu.edu
Department of Mathematics, University of Pennsylvania, Philadelphia, PA 19104, U.S.A.

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

Daniel D. Lee
ddlee@seas.upenn.edu
Departments of Electrical and Systems Engineering, Computer and Information Science, and Bioengineering, University of Pennsylvania, Philadelphia, PA 19104, U.S.A.

The efficient coding hypothesis assumes that biological sensory systems use neural codes that are optimized to best possibly represent the stimuli that occur in their environment. Most common models use information–theoretic measures, whereas alternative formulations propose incorporating downstream decoding performance. Here we provide a systematic evaluation of different optimality criteria using a parametric formulation of the efficient coding problem based on the L_p reconstruction error of the maximum likelihood decoder. This parametric family includes both the information maximization criterion and squared decoding error as special cases. We analytically derived the optimal tuning curve of a single neuron encoding a one-dimensional stimulus with an arbitrary input distribution. We show how the result can be generalized to a class of neural populations by introducing the concept of a meta–tuning curve. The predictions of our framework are tested against previously measured characteristics of some early visual systems found in biology. We find solutions that correspond to low values of p, suggesting that across different animal models, neural representations in the early visual pathways optimize similar criteria about natural stimuli that are relatively close to the information maximization criterion.

Neural Computation **28, 2656–2686** (2016)
doi:10.1162/NECO_a_00900

1 Introduction

The efficient coding hypothesis states that biological sensory systems have limited coding resources and therefore seek to employ coding strategies that are optimally adapted to the statistical structure of their sensory environment (Attneave, 1954; Barlow, 1961; Maddess & Laughlin, 1985; Theunissen & Miller, 1991; Fitzpatrick, Batra, Stanford, & Kuwada, 1997; Harper & McAlpine, 2004). Several studies have experimentally demonstrated that sensory neural codes seem to indeed follow input distribution statistics in order to reach higher coding efficiency (Brenner, Bialek, & de Ruyter van Steveninck, 2000; Twer & MacLeod, 2001; Dean, Harper, & McAlpine, 2005; Ozuysal & Baccus, 2012). A large fraction of previous work assumed that neural representations are tuned to maximize the mutual information they are able to convey about the stimulus values given some overall constraints on available metabolic costs—total number of spikes, for example (Laughlin, 1981; Linsker, 1989; Atick & Redlich, 1990; van Hateren, 1993; Seung & Sompolinsky, 1993; Nadal & Parga, 1994; Brunel & Nadal, 1998; Zhang & Sejnowski, 1999; Pouget, Deneve, Ducom, & Latham, 1999; Kang, Shapley, & Sompolinsky, 2004; Sharpee et al., 2006; McDonnell, & Stocks, 2008; Nikitin, Stocks, Morse, & McDonnell, 2009; Tkacik, Prentice, Balasubramanian, & Schneidman, 2010; Yarrow, Challis, & Seriès, 2012; Kastner, Baccus, & Sharpee, 2015). This Infomax criterion has been a preferred choice because it does not require making any further assumptions about potential downstream computations and tasks the encoded stimulus may be involved in. A few studies have taken a downstream perspective and have argued for optimality criteria that consider how well the stimulus information can be reconstructed from the neural representations. They often use a metric criterion in terms of the mean squared reconstruction error (Bethge, Rotermund, & Pawelzik, 2002, 2003; Berens, Gerwinn, Ecker, & Bethge, 2009; Yaeli & Meir, 2010; Doi & Lewicki, 2011). This reconstruction metric has been shown to optimize performance in perceptual estimation and classification tasks (Salinas, 2006). Recently there has been increasing interest in comparing the information with the metric approach (Ganguli & Simoncelli, 2010; Wang, Stocker, & Lee, 2013; Gjorgjieva, Sompolinsky, & Meister, 2014; Grabska-Barwinska & Pillow, 2014). However, a unified comparison and evaluation of these different approaches is still lacking.

Here, we provide a unified framework to compare these optimal criteria. We introduce a parametric formulation of the efficient coding problem in terms of minimizing the overall reconstruction error according to the L_p norm as a function of the norm parameter p. We assume reconstruction from a maximum likelihood estimate (MLE) decoder in the asymptotic time limit. More specifically, we consider a one-dimensional stimulus s with distribution $f(s)$ that is encoded in m neurons with tuning curve(s) $\boldsymbol{h}(s)$. While the mapping $\boldsymbol{h}(s)$ is deterministic, we assume the neural response $\mathbf{r}$ to follow a distribution $P(\mathbf{r}|\boldsymbol{h}(s))$ according to neural noise. For both

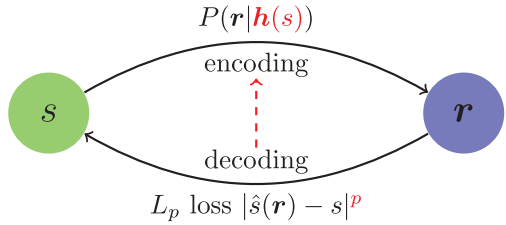


Figure 1: Efficient coding problem in terms of reconstruction error. A one-dimensional stimulus s is encoded in a neural response pattern r. We define the optimal tuning curve(s) $\boldsymbol{h}(s)$ as the one that minimizes the overall L_p reconstruction error according to an MLE decoder. We study how the optimal coding strategy is dependent on the norm parameter p. The Infomax solution is equivalent to the optimal encoder for $p \to 0$.

Poisson and gaussian noise, we analytically derive the optimal tuning curve $\boldsymbol{h}$ to achieve minimal L_p mean reconstruction error for arbitrary stimulus distributions. This framework includes both the Infomax and mean-squared error optimal solutions in the limit of $p \to 0$ and $p = 2$, respectively (see Figure 1). We first focus on solutions for the optimal tuning curve $h(s)$ of a single (sigmoidal) neuron encoding the stimulus. We then show how the single neuron tuning curve solution can be naturally extended to populations of neurons. Under certain assumptions, the optimal single neuron tuning curve $h(s)$ can be related to an optimal meta-tuning curve of the neural population, from which the individual tuning characteristics of the population of neurons can be determined.

In the context of this theoretical framework, we investigate how known tuning characteristics of biological sensory systems can be explained. We compare the measured tuning characteristics of early sensory representations in the fly, the cat, and the monkey for known stimulus statistics with predictions from our framework. For the examples we tested, the biological tuning characteristics are quite well predicted by our framework and are best matched for small values of the norm parameter p. We conclude that early sensory representations in biological systems may be optimized to convey maximal information.

2 Optimal Neural Coding for a Single Neuron

We start with the case where a single neuron is encoding a one-dimensional stimulus variable s. We assume that s follows a distribution density $f(s)$. We also assume that the neuron's average firing rate is determined by a sigmoidal function $h(s)$. The actual observed firing rate r is subject to neural noise, whose variability is described by a stochastic model $P(r|h(s))$.

We do not limit the noise to be defined by a canonical Poisson spiking model. Rather, we only assume that the mean firing rate is equal to the output of the tuning curve $\langle r \rangle = h(s)$ and the spike generating process is independent of the neuron's spiking history. With sufficient encoding time or with independent observations of identical neurons, the accumulated noise is asymptotically normal with zero mean and fixed variance according to the central limit theorem. In order to decode the input stimulus s, we take the maximum likelihood estimator (MLE) $\hat{s}(r)$, which is asymptotically unbiased and efficient (Cover & Thomas, 1991).

In order to find the L_p optimal tuning curve for a one–dimensional stimulus s, we need to minimize the mean L_p loss of the maximum likelihood estimator. The only constraint for the sigmoidal tuning curve is the saturation limit of the firing rate. Within the regime of low noise limit, the maximum firing rate does not affect optimality. Therefore, we assume $0 \le h(s) \le 1$ without loss of generality, which leads to the optimization problem

$$\text{minimize} \quad \left\langle |\hat{s}(r) - s|^p \right\rangle_{s,r} \tag{2.1}$$

$$\text{subject to} \quad 0 \le h(s) \le 1. \tag{2.2}$$

2.1 Objective Functions in Terms of Fisher Information. To get insight into the optimization problem, we analyze the Fisher information. The Fisher information $\mathcal{I}(s)$ describes the encoding precision for each specific individual stimulus s. For any s, $\mathcal{I}(s)$ can be calculated according to its definition

$$\mathcal{I}(s) = \left\langle \left(\frac{\partial}{\partial s} \log p(r|s) \right)^2 \middle| s \right\rangle_r, \tag{2.3}$$

where the conditional distribution $p(r|s)$ describes the stochastic neural response for a given stimulus and the average is taken over r but not s. It has been shown that in the asymptotic limit of long encoding time, the total Fisher information characterizes the precision of the ML estimator $\hat{s}$ in reconstructing the stimulus s (see section A.1 in appendix A):

$$(\hat{s}(r) - s) \sim \text{Normal}(0, \mathcal{I}(s)^{-1}), \tag{2.4}$$

$$\left\langle |\hat{s}(r) - s|^p \middle| s \right\rangle_r = const(p) \cdot \mathcal{I}(s)^{-p/2}. \tag{2.5}$$

It is clear from equation 2.5 that larger Fisher information leads to smaller L_p error. One example is the Cramer-Rao lower bound when $p = 2$. The more general equation 2.5 establishes the connection between L_p loss in equation 2.1 and the Fisher information. This results in an equivalent optimization

in terms of Fisher information:

$$\text{minimize} \quad \left\langle \mathcal{I}(s)^{-p/2} \right\rangle_s . \tag{2.6}$$

In addition to the L_p-error minimization problem, we also consider the well-known Infomax optimization, which maximizes the mutual information between the response and the stimulus. It has previously been shown that Fisher information can be related to mutual information (Brunel & Nadal, 1998). In our framework, Infomax is equivalent to optimizing the logarithm of Fisher information:

$$I(\mathbf{r}, s) = \frac{1}{2} \left\langle \log \mathcal{I}(s) \right\rangle_s + const, \tag{2.7}$$

$$\text{minimize} \quad - \left\langle \log \sqrt{\mathcal{I}(s)} \right\rangle_s . \tag{2.8}$$

2.2 Constraints in Terms of Fisher Information. Next, we show how to incorporate constraints in equation 2.2 into the same framework. For a one-dimensional stimulus variable, the Fisher information of a neuron is fully determined by the nonlinear tuning curve $h(s)$ and the noise model. Here we show the results for both Poisson noise (P) and constant gaussian noise (cG), with details provided in the section A.2:

$$\textbf{P}: \quad \mathcal{I}(s) \propto \frac{h'(s)^2}{h(s)}, \tag{2.9}$$

$$\textbf{cG}: \quad \mathcal{I}(s) \propto h'(s)^2. \tag{2.10}$$

These formulations can easily be inverted. For any given Fisher information allocation $\mathcal{I}(s)$, the corresponding nonlinear tuning curve is

$$\textbf{P}: \quad h(s) \propto \left(\int_{-\infty}^{s} \sqrt{\mathcal{I}(\xi)}\, d\xi \right)^2, \tag{2.11}$$

$$\textbf{cG}: \quad h(s) \propto \int_{-\infty}^{s} \sqrt{\mathcal{I}(\xi)}\, d\xi . \tag{2.12}$$

Given bound constraints on the tuning curve in equation 2.2, we have

$$\textbf{P}: \quad \int_{-\infty}^{\infty} \sqrt{\mathcal{I}(s)}\, ds \propto \int_{-\infty}^{\infty} \frac{h'(s)}{\sqrt{h(s)}}\, ds = 2\sqrt{h(s)}\Big|_{-\infty}^{\infty} \leq const., \tag{2.13}$$

$$\textbf{cG}: \quad \int_{-\infty}^{\infty} \sqrt{\mathcal{I}(s)}\, ds \propto \int_{-\infty}^{\infty} h'(s)\, ds = h(s)\big|_{-\infty}^{\infty} \leq const. \tag{2.14}$$

Ignoring irrelevant constant scalar terms that do not affect the optimal form, these constraints can be unified:

$$\textbf{P or cG:} \quad \text{subject to} \quad \int_{-\infty}^{\infty} \sqrt{\mathcal{I}(s)}\, ds \leq const. \tag{2.15}$$

Since it is always better to have more Fisher information, equality in equation 2.15 must hold for optimality. To summarize, the objective function in equation 2.6, attempts to optimally allocate the Fisher information $\mathcal{I}(s)$ across the space of the stimulus variable s with distribution $f(s)$ under the integral constraint in equation 2.15. After determining the optimal allocation $I^*(s)$, the optimal nonlinearity $h^*(s)$ can then be determined using equation 2.11 or 2.12 depending on the neural noise model.

2.3 Single Neuron Results. According to the above analysis, solving the L_p reconstruction error minimization problem is equivalent to solving the Fisher information allocation problem. For each p value in the L_p-minimum decoding loss criterion, the optimization problem is

$$\text{minimize} \quad \left\langle (\mathcal{I}(s))^{-p/2} \right\rangle_s = \int f(s)\, (\mathcal{I}(s))^{-p/2}\, ds, \tag{2.16}$$

$$\text{subject to} \quad \int \sqrt{\mathcal{I}(s)}\, ds \leq const. \tag{2.17}$$

This variational problem can easily be solved; the optimal solution is

$$I^*(s) \propto f(s)^{2/(1+p)}, \tag{2.18}$$

$$\textbf{P:} \quad h^*(s) = \left(\frac{\int_{-\infty}^{s} f(\xi)^{1/(1+p)}\, d\xi}{\int_{-\infty}^{\infty} f(\xi)^{1/(1+p)}\, d\xi} \right)^2, \tag{2.19}$$

$$\textbf{cG:} \quad h^*(s) = \frac{\int_{-\infty}^{s} f(\xi)^{1/(1+p)}\, d\xi}{\int_{-\infty}^{\infty} f(\xi)^{1/(1+p)}\, d\xi}. \tag{2.20}$$

A simple comparison between the two noise models reveals that the optimal tuning curve for a neuron with Poisson noise is exactly the square of the optimal tuning curve for a neuron with constant gaussian noise. This relationship was first reported by Bethge et al. (2002) and Johnson and Ray (2004). The squaring transformation shows that the optimal coding under Poisson noise tends to use more reliable low firing rates rather than more unreliable higher rates. Below we focus on the constant gaussian noise solution and discuss the link between our general formula and several results that have been previously reported in the literature:

- When $p = 0$, the L_0-minimum solution is given by the cumulative function of the input distribution,

$$h^*(s) \propto \int_{-\infty}^{s} f(\xi)\, d\xi. \tag{2.21}$$

- When $p = 2$, the L_2-minimum solution is given by the cumulative function of the cube root of the input distribution,

$$h^*(s) \propto \int_{-\infty}^{s} f(\xi)^{1/3}\, d\xi. \tag{2.22}$$

- When $p \to \infty$, the optimal tuning curve $h^*(s)$ converges to a linear function because its derivative approaches a constant function of s and the prior $p(s)$ is no longer relevant. However, this usually requires the stimulus to be bounded $s \in [s_{\min}, s_{\max}]$; otherwise, the integral of $f(s)^{1/(1+p)}$ will diverge for sufficiently large p.

Note that optimizing the L_p-min problem, equation 2.16, for $p \to 0$ leads to the same optimal solution as the Infomax problem in equation 2.8. This solution, first proposed in Laughlin (1981) and Nadal and Parga (1994), is known as the output equalization rule because the output $h^*(s)$ is uniformly distributed within its range limit. We informally refer to both "L_0-min" and the Infomax solution in the remainder of this letter. When $p = 2$, the optimal solution in equation 2.22 minimizes the mean square error of the reconstructed stimulus. This solution was first proposed for optimal RGB color perception (Twer & MacLeod, 2001) and discussed in Wang, Stocker, and Lee (2012).

To summarize, the solution in equation 2.20 provides a systematic understanding of the optimal nonlinearities for the various criteria as a function of by p. In Figure 2, we illustrate different L_p optimal tuning curves for a standard gaussian stimulus prior. Intuitively, the efficient coding problem can be understood as optimizing the allocation of neural descriptive power across an inhomogeneous stimulus distribution. Depending on the value of p, the optimal allocation strategy balances between more frequently appearing stimuli with less frequent ones. Strategies corresponding more with Infomax (p near 0) emphasize stimuli with a higher likelihood of appearing. L_p-optimal strategies with large p are more conservative and need to spend more resources to encode more surprising stimuli since the error penalty is larger.

2.4 Examples of Various Stimulus Prior Distributions. We derived optimal tuning curves for a few example stimulus distributions (priors). In particular, we considered prior distributions that can be expressed as generalized gaussian distributions with scale parameter c and shape parameter β. For simplicity, we show solutions only for the constant gaussian noise

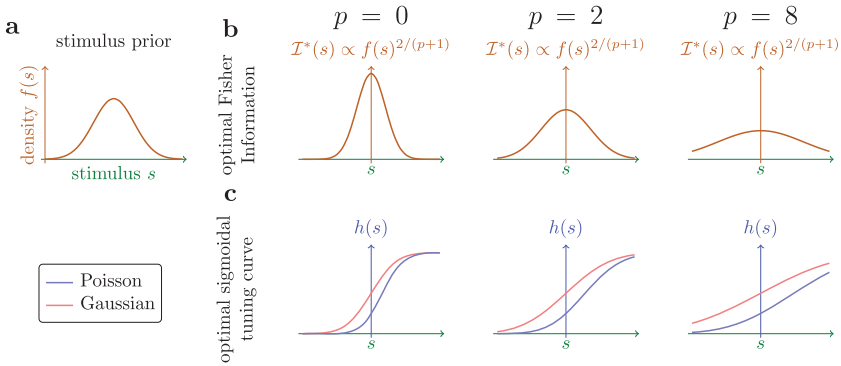


Figure 2: The L_p optimal sigmoidal tuning curves for for $p = 0, 2, 8$ for both Poisson and constant gaussian noise models. (a) The gaussian stimulus distribution (prior). (b) For each p, the optimal Fisher Information $I^*(s)$ is derived based on the prior distribution. (c) The corresponding optimal tuning curves for Poisson noise (blue lines) or constant gaussian noise (red lines).

assumption. From equation 2.20, the L_p-optimal tuning curve is related to the input stimulus distribution:

$$f(s) \propto \exp\left(-c|s|^{\beta}\right), \tag{2.23}$$

$$h'(s) \propto f(s)^{1/(1+p)} \propto \exp\left(-c\left(\frac{|s|}{(1+p)^{1/\beta}}\right)^{\beta}\right). \tag{2.24}$$

Therefore, for a certain value of p, the nonlinearity is simply a rescaled version of the cumulative function of $f(s)$. The scalar $(1+p)^{1/\beta}$ is a decreasing function of β. In Figure 3, we illustrate three cases: in the extreme of uniform distribution case where $\beta = \infty$, the scalar remains a constant and there is no difference across all the L_p-optimal tuning curves; for the gaussian distribution case where $\beta = 2$, the scalar grows sublinearly as $(1+p)^{1/2}$; and for the Laplacian distribution case where $\beta = 1$, the scalar grows linearly as $(1+p)$.

Another important conclusion we highlight is that all the L_p-optimal solutions except L_0 are not invariant under nonlinear stimulus transformations. For example, the L_2-optimal solution for a positive valued stimulus is not identical to the L_2-optimal solution for the same stimulus transformed to a logarithmic scale. The L_0-optimal solution is the only one that is invariant under any one-to-one stimulus transformations (Cover & Thomas, 1991). This fact again demonstrates the intuition that L_p-min strategies are highly task driven: the solution changes if the stimulus variable undergoes some nonlinear transformation before being processed.

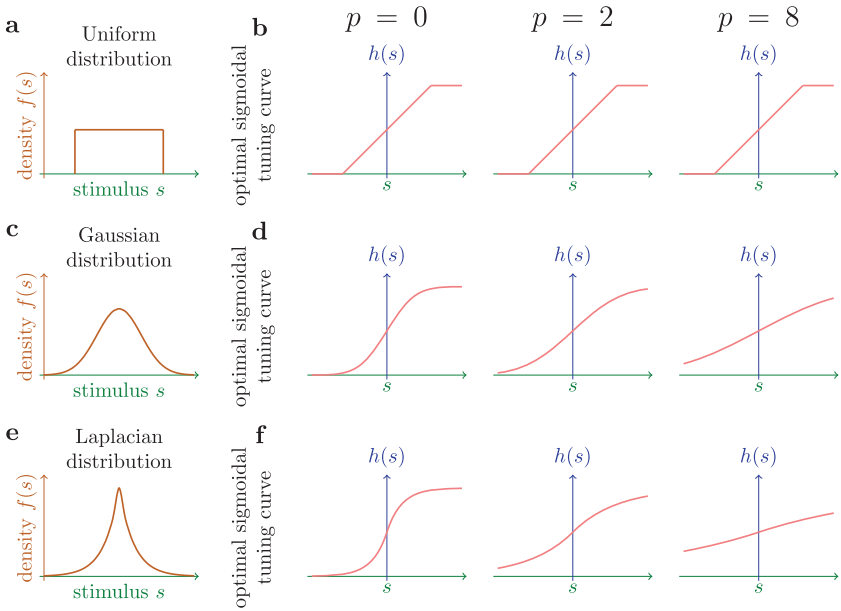


Figure 3: The L_p optimal sigmoidal tuning curves for a single neuron with constant gaussian noise model. Here we compare the results for various forms of prior distributions: uniform distribution (a, b), Gaussian distribution (c, d), and Laplacian (or double exponential) distribution (e, f).

3 Generalization to Neural Populations

The tuning characteristics of optimal neural population codes have been studied (Zhang & Sejnowski, 1999; Pouget et al., 1999; Kang et al., 2004; McDonnell & Stocks, 2008; Nikitin et al., 2009; Ganguli & Simoncelli, 2010; Yaeli & Meir, 2010). The conclusion from these studies is that the solutions largely depend on the individual assumptions made in the corresponding derivations.

3.1 Neural Population Assumptions. Certain assumptions are necessary to derive a well-constrained optimization problem. Rather than allowing all neurons in the population to independently exhibit arbitrary nonlinear tuning curves, we assumed the tuning curve of the kth neuron to have the following form:

$$h_k(s) = h_0(\psi(s) - \psi(s_k)). \tag{3.1}$$

We refer to $\psi(s)$ as the meta-tuning curve that transforms the stimulus s to neural space. For each neuron, s_k is the characteristic stimulus associated

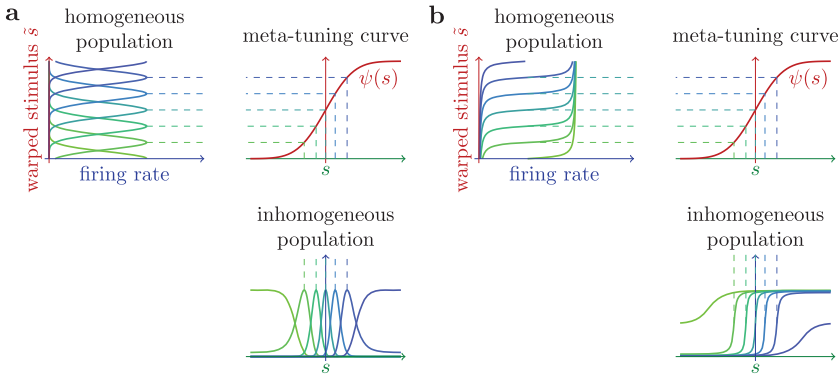


Figure 4: Under our assumptions, the inhomogeneous neural population tuning is derived by warping a homogeneous tuning description through the meta-tuning curve; that is, the stimulus space is nonlinearly transformed according to the meta-tuning curve via the sigmoidal meta-tuning curve $\psi(s)$. Two representative choices of h_0 are (a) unimodal and (b) sigmoidal.

with that neuron. For example, s_k can be the preferred stimulus (at which the neuron elicits maximum neural response) for neurons with unimodal tuning curves or the semisaturation stimulus (at which the neuron elicits half of the maximum neural response) for neurons with sigmoidal tuning curves.

Below, we denote $\tilde{s} = \psi(s)$ and $\tilde{s}_k = \psi(s_k)$ resulting from the output of the meta-tuning curve. Further assumptions are:

1. All neurons in the population share the same given nonlinearity $h_0(\tilde{s} - \tilde{s}_k)$.
2. The characteristic stimuli $\tilde{s}_k$ are uniformly distributed; in other words, the spacing $\Delta\tilde{s} = \tilde{s}_k - \tilde{s}_{k-1}$ between adjacent neurons is a constant.
3. h_0 and h_0' are slowly varying when measured at the scale of $\Delta\tilde{s}$, that is, $h_0(\tilde{s}_k) \approx h_0(\tilde{s}_k + \Delta\tilde{s})$ and $h_0'(\tilde{s}_k) \approx h_0'(\tilde{s}_k + \Delta\tilde{s})$. When $\Delta\tilde{s}$ is small, this constraint is equivalent to h_0 and h_0' being continuous.
4. The neurons have independent output noise, so the total Fisher information of the population is the linear sum of the Fisher information of each individual neuron, $I_{\text{total}}(s) = \sum_k I_k(s)$ (see section A.3 for the proof).

These assumptions are sometimes referred to as the uniform tiling properties of a neural population (Ganguli & Simoncelli, 2010; Grabska-Barwinska & Pillow, 2014). It is important to note that assumptions 1 and 2 limit the solutions to a subspace of all possible population codes for which the mapped stimulus $\tilde{s}$ is encoded by a homogeneous population (see Figure 4). In our model, the total Fisher information of the population

with either the Poisson noise or constant gaussian noise (see equation 2.9 or 2.10) becomes

$$I_0 \approx I_{\text{total}}(\tilde{s}) = \sum_k I_k(\tilde{s}) = \sum_k \frac{h_0'(\tilde{s} - \tilde{s}_k)^2}{h_0(\tilde{s} - \tilde{s}_k)} \quad \text{or} \quad \sum_k h_0'(\tilde{s} - \tilde{s}_k)^2. \tag{3.2}$$

The form of $h_0(\cdot)$ is fixed and often assumed but not limited to be either unimodal or sigmoidal. In Figure 4, we illustrate how to determine the individual tuning curves of the inhomogeneous neural population.

3.2 Optimal Meta-Tuning Curve. For any meta-tuning curve $\tilde{s} = \psi(s)$, we can calculate the Fisher information of the kth neuron and the total Fisher information for the population with respect to the original stimulus s as

$$\textbf{P}: \quad I_k(s) \propto \frac{h_0'(\psi(s) - \tilde{s}_k)^2}{h_0(\psi(s) - \tilde{s}_k)} \cdot \psi'(s)^2, \tag{3.3}$$

$$\textbf{cG}: \quad I_k(s) \propto h_0'(\psi(s) - \tilde{s}_k)^2 \cdot \psi'(s)^2, \tag{3.4}$$

$$\textbf{P or cG:} \quad I_{\text{total}}(s) = \sum_k I_k(s) \approx I_0 \cdot \psi'(s)^2. \tag{3.5}$$

In the population coding case, the mean L_p reconstruction error of s is related to the total Fisher information, and we need to minimize the following term,

$$\left\langle \left(I_{\text{total}}(s)\right)^{-p/2} \right\rangle_s = \int f(s) \left(I_{\text{total}}(s)\right)^{-p/2} ds, \tag{3.6}$$

where $f(s)$ is the prior distribution of the stimulus s. We can limit the output of a nondecreasing meta-tuning curve to the range $0 \le \psi(s) \le const$. Then, minimizing the L_p reconstruction error is equivalent to the following optimization in terms of the meta-tuning curve $\psi(s)$:

$$\text{minimize} \quad \left\langle \left(I_{\text{total}}(s)\right)^{-p/2} \right\rangle_s \approx I_0^{-p/2} \cdot \int f(s) \psi'(s)^{-p} ds \tag{3.7}$$

$$\text{subject to} \quad \int \psi'(s)\, ds \le const. \tag{3.8}$$

This optimization problem is the same as the constant gaussian noise case we previously discussed in section 2.3. This leads to a solution for the optimal meta-tuning curve $\psi^*(s)$ with corresponding total Fisher information:

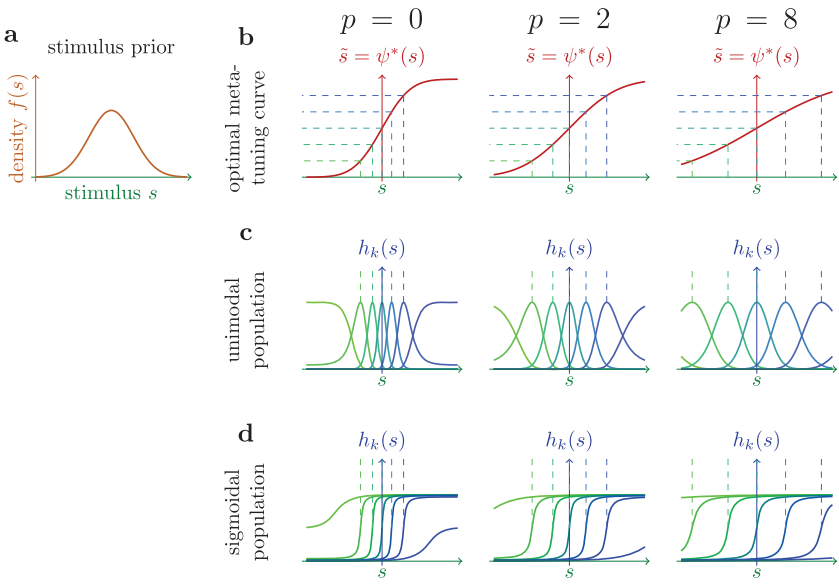


Figure 5: The L_p optimal neural populations for $p = 0, 2, 8$ and a gaussian stimulus distribution. (a, b). Replicated from Figure 2. The optimal meta-tuning curve for the population is identical to the optimal tuning curve of a single neuron with constant gaussian noise. Here we show two different kinds of optimal neural population, where each neuron has (c) unimodal tuning curves or (d) sigmoidal tuning curves.

$$\psi^{*\prime}(s) \propto f(s)^{1/(1+p)}, \quad I^*_{\text{total}}(s) \propto f(s)^{2/(1+p)}, \tag{3.9}$$

$$\psi^*(s) = \frac{\int_{-\infty}^{s} f(\xi)^{1/(1+p)}\, d\xi}{\int_{-\infty}^{\infty} f(\xi)^{1/(1+p)}\, d\xi}. \tag{3.10}$$

This result illustrates that under our model, the Fisher information allocation for the population is entirely determined by the meta-tuning curve $\psi(s)$, in the same way as the Fisher information allocation is determined by the sigmoidal tuning curve $h(s)$ of a single neuron with constant gaussian noise. In Figure 5, we show the L_0, L_2, and L_8 optimal neural populations for encoding a stimulus variable with a gaussian distribution. Compared to previous work by Ganguli and Simoncelli (2010), our framework considers a more constrained class of neural populations because it assumes a fixed gain across neurons. Our formulation, however, allows us to specify an entire family of L_p-optimal solutions that smoothly incorporate the special cases of the Infomax and the MSE solutions.

4 Relaxing the Asymptotic Assumptions

For both the single neuron case and the neural population case, our results so far have relied on several key assumptions. The most restrictive one is the assumption that neurons are operating in the asymptotic long time limit. In this limit, the optimal decoder naturally converges to the MLE. In contrast, in a more realistic scenario where encoding time is short, it is generally the case that a Bayesian (and usually biased) decoder will perform better. Unfortunately, it is difficult to derive analytic solutions in this case, yet numerical efforts have been made (Bethge et al., 2003; Nikitin et al., 2009). Furthermore, the derivation of the optimal Bayesian decoder can be intractable for arbitrary prior distributions.

In order to provide a sense of how well our derived analytic solutions hold for shorter encoding times, we compared their predicted performance to the actual measured performance obtained by numerical simulations. The decoding performance of our L_p optimized coding solutions can be easily simulated for arbitrary encoding times. For simplicity, we considered a standard gaussian stimulus distribution $p(s)$ in our simulations. The encoding process is straightforward: stimuli are sampled and encoded by the L_p optimal code with additional Poisson spiking noise. For the decoding process, we examined both the assumed unbiased, MLE and the maximum a posteriori estimator (MAPE). In both cases, the iterative gradient descent method (Newton's method) was used to find the stimulus with maximal likelihood (for MLE) or maximal posterior likelihood (for MAPE). The mean L_p decoding error was then calculated over a large set of generated stimuli and compared to the theoretical prediction.

For a neuron with maximum firing rate $r_{\max}$ and a fixed length of the time window T, the key variable is the maximum allowed spike count $N_{\max} = r_{\max}T$. For each value of $N_{\max}$, we ran 100 independent trials, and in each trial, 100,000 stimuli were randomly generated. This experiment was done for both a single neuron with sigmoidal tuning curve and a population of neurons with unimodal tuning curves. Results are shown in Figure 6. As expected, the theoretical predictions were more accurate when $N_{\max}$ was large, with the critical value for $N_{\max}$ increasing as a function of p. For a shorter encoding time, our result shows that the MAPE is a better estimator despite the similar performance for larger $N_{\max}$. The performance of the MLE seems to be lower–bounded by our theoretical prediction (see the solid line), but the MAPE benefits from the prior information and is upper–bounded by a constant related to that prior.

In the single neuron case, the critical spike count $N_{\max}$ ranges from approximately 10^2 (for $p = 0.01$) to approximately 10^4 spikes (for $p = 2$). For some sensory neurons, such as the H1 neuron of a blowfly (see section 5.1), the maximal firing rate $r_{\max}$ can be as high as 100 Hz, which means that the critical time for the long encoding assumption to be valid is around $T \geq 1$

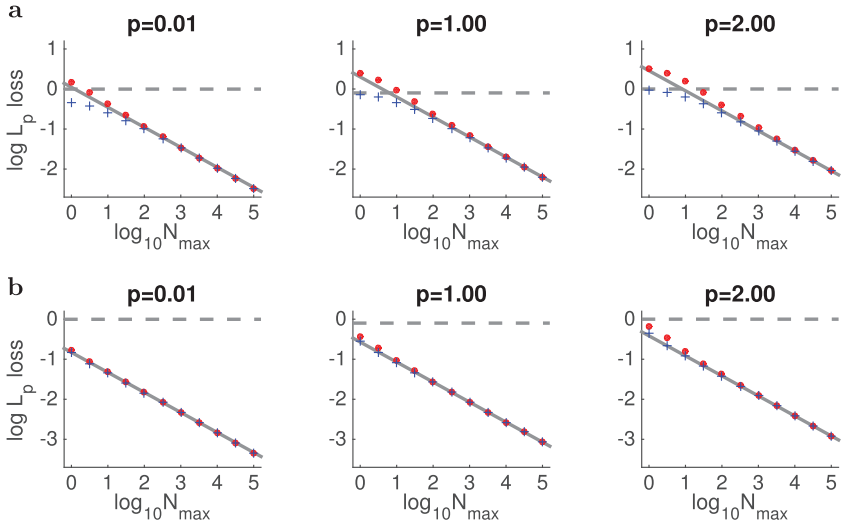


Figure 6: The simulated L_p encoding error (MLE: red dot, MAPE: blue cross) versus theoretical prediction assuming unbiased estimator (solid lines) or using only prior information (dashed lines). The markers indicates the median over 100 trials. (a) The performance of a single neuron with sigmoidal tuning curve (see, e.g., Figure 3d). (b) The performance of a population with $K = 11$ neurons with unimodal tuning curves (see, e.g., Figure 5c). The vertical axis is the mean L_p loss $\langle|\hat{s} - s|^p\rangle^{1/p}$, and the horizontal axis is $N_{\max}$, both in logarithm space with base 10.

sec (for $p = 0.01$) to $T \geq 100$ sec (for $p = 2$). In the neural population case, we run simulations with $K = 11$ neurons with unimodal tuning curves. As expected, the performance in terms of the L_p error is one order of magnitude better than for the single neuron case. Correspondingly, the critical spike count $N_{\max}$ is much smaller: from approximately $10^{0.5}$ (for $p = 0.01$) to approximately $10^{1.5}$ spikes (for $p = 2$). For small p values, the performance matches the theoretical prediction for populations containing as few as 11 neurons with $N_{\max} \geq 3$ spikes per neuron. For a larger p value such as $p = 2$, this number may increase to $N_{\max} \geq 30$ spikes per neuron.

In sum, we found that depending on the value of p, the long time-limit assumptions can be reasonably relaxed for short encoding times. In particular, we find that the critical spike count can be as low as $N_{\max} = 3$ to 30 spikes per neuron, which justifies the biological relevance of our result. Generally the predictions of our framework are much less constrained for smaller p values. We have also found that the performance of a Bayesian decoder (the MAPE) tends to be better than the MLE decoder, which shows that the optimality of our solution (MLE) strongly relies on the unbiased

assumption. Fortunately, this limitation is subordinated to the short encoding time limitation. The MAPE itself is asymptotically unbiased and has similar performance as the MLE decoder once the critical $N_{\max}$ is reached.

5 Efficient Codes in Visual Perception

Our theoretical analysis raises the question of which efficiency criterion the brain actually uses to encode information. In this section, we consider several different sensory modalities in early vision: motion encoding, orientation encoding, and contrast encoding. In each case, we attempted to estimate the prior distribution of the input stimulus and compared the tuning characteristics of the predicted efficient coding model with published physiological data.

5.1 Speed Encoding by a Single Blowfly H1 Neuron. We first analyzed data from the H1 neuron of blowfly, which encodes the speed s of a horizontally moving bar. The analyzed data set (de Ruyter van Steveninck, Lewen, & Bialek, 1997) was collected from a fly H1 neuron responding to a stochastically generated visual motion stimulus. The data were taken for 20 minutes at a sampling rate of 500 Hz. We binned the neural data into 1200 bins with duration $\Delta t = 1$ second and calculated the average stimulus s_i and the number of spikes N_i for $i = 1, \ldots, 1200$ in each bin. This stimulus-response relation is plotted in Figure 7a.

The natural speed prior for the blowfly is unknown. However, based on the investigation of natural movie clips, previous research has proposed that the prior distribution for visual speed should follow a power law function of the form $f(s) \propto (1 + |s|/v_0)^{-2}$, where $v_0 > 0$ is a scale parameter (van Hateren, 1993; Dong & Atick, 1995; Stocker & Simoncelli, 2006). For this particular form of the prior, the optimal L_p tuning curve $h_p^*(s)$ for a neuron with Poisson noise can be analytically computed:

$$h_p^{*\prime}(s) \propto f(s)^{\frac{1}{1+p}} \quad \Rightarrow \quad h_p^*(s) \propto \left(1 + \text{sign}(s)\left(1 - \frac{1}{(1 + |s|/v_0)^{\frac{1-p}{1+p}}}\right)\right)^2. \tag{5.1}$$

It can be seen that for this parametric form of the prior distribution, the L_p optimal solution exists only for $0 \le p \le 1$. In order to infer the prior distribution and the optimal norm parameter, we optimized for the parameters v_0 and p that maximized the data likelihood. The result, shown in Figure 7b, represents the predicted speed prior distribution to which the H1 neuron is optimally adapted to. In Figures 7c and 7d, we can see that parameter values $v_0 = 21.3$ deg/sec and $p = 0$ lead to the highest data likelihood. However, other pairs of (p, v_0) for $p < 0.8$ also yield good likelihood scores.

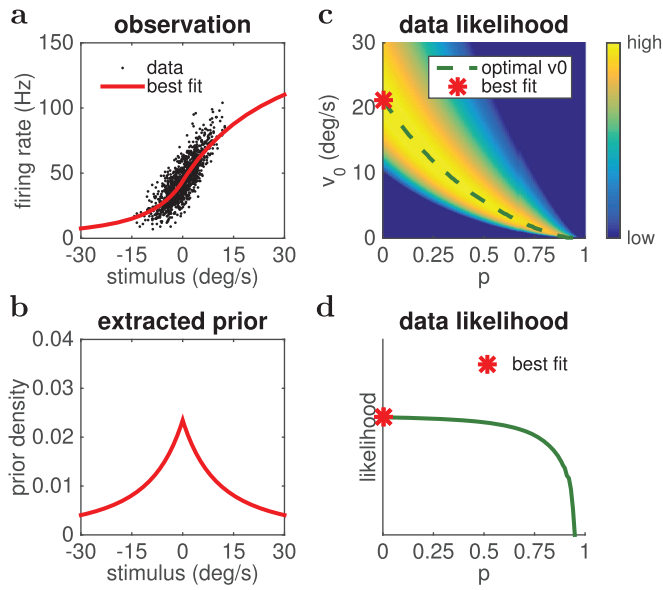
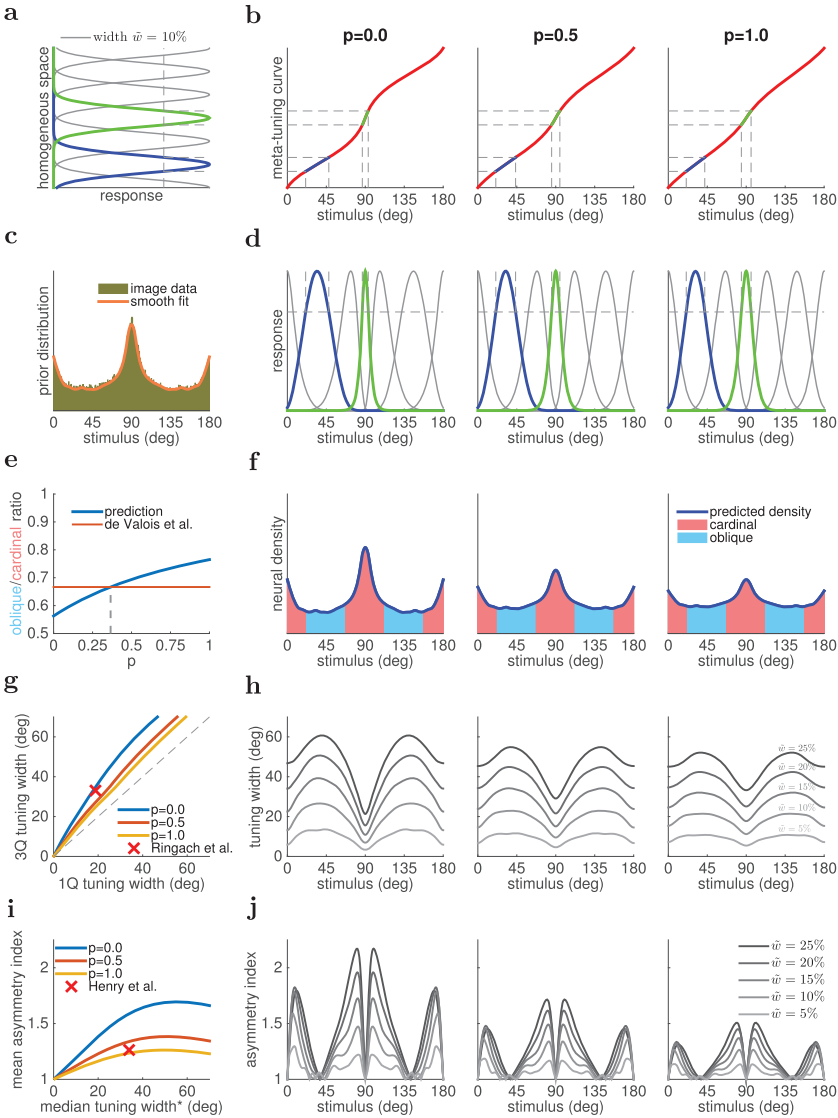


Figure 7: (a) The stimulus-response data collected from a fly H1 neuron (van Steveninck et al., 1997). We plot the best tuning curve using the parametric model in equation 5.1. (b) The predicted prior distribution to which the fly H1 neuron is most likely adapted. (c) The optimal parameter v_0 and p is chosen to maximize the data likelihood. Dashed line shows the optimal parameter $v_0(p)$ as a function p. (d) The maximum data likelihood for each pair $(p, v_0(p))$ as a function of p.

5.2 Orientation Encoding with Neural Populations. We also applied our proposed framework to analyze biological neural populations that encode local visual orientation. We first estimated the prior distribution $f(\theta)$ of local visual orientation θ from a natural image data set (van Hateren & van der Schaaf, 1998) using a filter analysis at a single spatial scale (detailed description in appendix B). The resulting prior distribution, shown in Figure 8c, is very similar to previously estimated distributions (see Girshick, Landy, & Simoncelli, 2011). Based on the estimated prior density, we derived the optimal meta-tuning curves $\psi(\theta)$ for various values of the norm parameter p (see Figure 8b). The unimodal tuning curves of the population (see Figure 8d) were then determined as described in section 3.2 assuming a homogeneous population of certain tuning width $\tilde{w}$ (see Figure 8a). Below, we compare predictions of the model population with measured biophysical characteristics of orientation tuned neurons.

The first prediction is with regard to neural density. De Valois, Yund, and Hepler (1982) reported that the ratio between neurons tuned for oblique



versus cardinal orientations is about 0.66 in area V1 of the macaque. In our framework, the neural density as a function of θ is directly related to the derivative of the meta-tuning curves (see Figure 8f). In order to compute the ratio between the number of neurons tuned for the oblique versus the cardinal orientations, we binned the neural population into two sub-populations shown as blue or red regions in Figure 8f. The predicted

ratio is a function of the norm parameter p (see Figure 8e); for $p \approx 0.37$, the ratio of the model population matches the ratio found for neurons in V1.

We can also predict how the tuning width depends on the preferred stimulus of the neurons. Following the definition of Ringach, Shapley, and Hawken (2002), we defined the tuning width w as the length of the orientation interval over which a neuron's mean response is at least $1/\sqrt{2}$ of its peak firing rate. Figure 8h shows the predicted tuning width $w(\theta)$ as a function of the preferred orientation θ of a neuron in the model population. Each curve shows the tuning width $w(\theta)$ for a different assumed constant tuning width $\tilde{w}$ in the homogeneous population (see Figure 8a). From these continuous functions, we calculated the first and third quartiles w_{1Q}, w_{3Q} of the tuning widths across the inhomogeneous population. For each p value, the possible values of $w_{1Q}(\tilde{w})$ and $w_{3Q}(\tilde{w})$ form a curve with parameter $\tilde{w}$ as shown in Figure 8g. A comparison of the quartile predictions with physiological data from neurons in area V1 of the macaque (Ringach et al., 2002) suggests that the model best matches the data for a norm parameter of value $p = 0.08$.

Finally, we can make predictions about tuning curve asymmetries. Specifically, we compared the predicted asymmetry index (Henry, Dreher, & Bishop, 1974) of our model population with the values found for biological neurons. Similar to the tuning width, the predicted asymmetry index is also a function of the assumed tuning width $\tilde{w}$ of the neurons in the homogeneous population (see Figure 8j). We computed the predicted relationship between the mean asymmetry index and the median tuning width for different p values and compared it with measurements from simple cells in striate cortex of the cat (Henry et al., 1974). The reported median tuning width (measured at 1/2 peak amplitude; we have rectified our predictions accordingly) of 34° and asymmetry index 1.26 matches our predictions for $p \approx 0.85$ (see Figure 8i).

Figure 8: Comparison between theoretically predicted and physiologically measured tuning characteristics of orientation tuned neural populations. (a–d) Cartoon examples of L_p-optimal neural population derived based on a homogeneous neural population and the optimal meta-tuning curve, which is determined by the prior distribution extracted from natural images. The p values are 0, 0.5 and 1. (e, f) The oblique versus cardinal ratio prediction is compared with previous results (De Valois et al., 1982) on macaque V1 foveal neurons, which suggests $p \approx 0.37$. (g, h) The first and third quartile tuning width prediction is compared with previous results (Ringach et al., 2002) on macaque V1, which suggests $p \approx 0.08$. (i, j) The asymmetry index and median tuning width(*) prediction is compared with previous results (Henry et al., 1974) on cat's striate cortex, which suggests $p \approx 0.85$. (* The tuning width here is measured at half-amplitude to be consistent with the previous study.)

In summary, we found that the measured orientation tuning characteristics of neurons in primary visual cortex of the macaque and the cat match those model predictions that correspond to fairly low values of p.

5.3 Contrast Encoding with Neural Populations. We also applied our framework to make predictions for the contrast gain characteristics of neurons in early visual cortex. The contrast of natural images has been defined in multiple ways in the literature. Two standard definitions of local contrast are the root-weighted-mean-square contrast (Najemnik & Geisler, 2005; Mante, Frazor, Bonin, Geisler, & Carandini, 2005) and the equivalent-Michelson contrast (Brady & Field, 2000; Tadmor & Tolhurst, 2000; Clatworthy, Chirimuuta, Lauritzen, & Tolhurst, 2003). We use the equivalent-Michelson contrast in order to match our predictions with recorded physiological data (Clatworthy et al., 2003). We gathered 200,000 patches of size 32×32, randomly sampled from natural images from the data set (van Hateren & van der Schaaf, 1998). The histogram of their equivalent-Michelson contrast is regarded as the prior distribution of the environment (see Figure 9c). The detailed description of this process is discussed in appendix C.

In early visual perception systems, contrast information is encoded by a population of neurons with contrast selectivity in a soft-thresholding manner. One traditional model characterizes the neuron's response as a function of the contrast c via the Naka-Rushton equation (Naka & Rushton, 1966),

$$h(c) = h_{\max} \cdot \frac{c^q}{c_{50}^q + c^q}, \tag{5.2}$$

where $h_{\max}$ is the maximum possible firing rate, c_{50} is the semisaturation contrast so that $h(c_{50}) = 0.5 \cdot h_{\max}$, and q is an exponent parameter characterizing the steepness of the curve near c_{50}. Using our framework, we can predict the distribution of the semisaturation constant c_{50} within a population and compare that to physiological data (Clatworthy et al., 2003) (see Figure 9e). Our prediction suggests that monkey V1 neurons are roughly performing Infomax encoding ($p \approx 0.15$), while cat striate cortex neurons are optimized for a larger value of p ($p \approx 0.75$). As we can see from Figure 9e, the fit for the c_{50} distribution of cat striate neurons is worse than for monkey VI neurons. The neural population in cat V1 seems to be adapted to smaller contrast values. This may be due to the mismatch between the natural image data set and the true visual environment of the animal.

6 Discussion

We have proposed a family of efficiency criteria for neural coding. Each efficiency criterion uniquely determines an optimal way of encoding a scalar

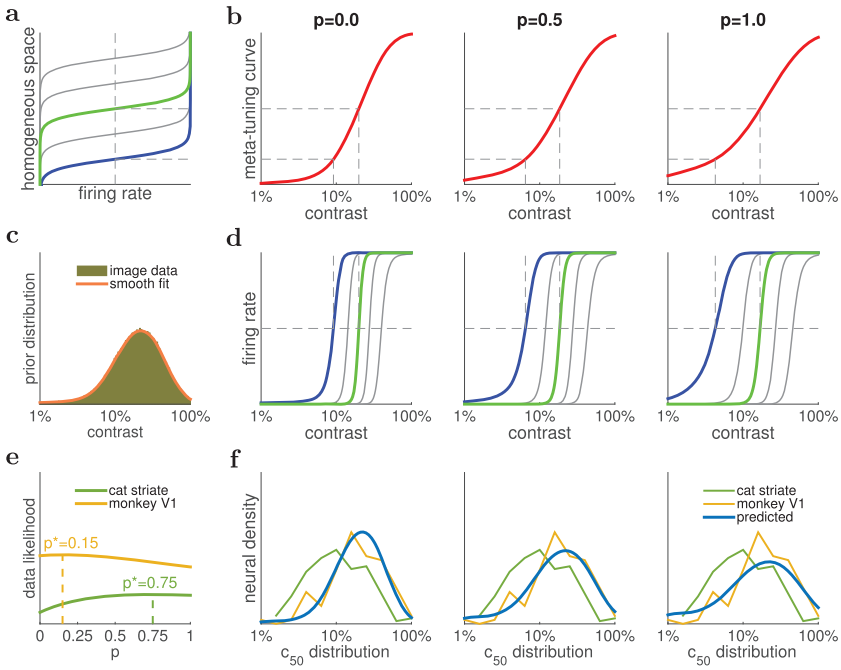


Figure 9: Analysis of L_p optimal population encoding of contrast in natural images. (a–d) Examples of L_p-optimal neural populations derived based on a homogeneous neural population and the optimal meta-tuning curve, which is determined by the prior distribution of equivalent-Michelson contrast extracted from natural images. The p values are 0, 0.5 and 1. (e, f) The predicted distribution of the c_{50} value for the entire population is compared with physiological data from Clatworthy et al. (2003) on cat's striate cortex and monkey's V1, best fits are $p \approx 0.15$ for the monkey and $p \approx 0.75$ for the cat.

stimulus with an arbitrary prior distribution. The efficiency criteria are parameterized by a parameter $p \geq 0$ associated with the underlying goal of minimizing the L_p reconstruction error when using a maximum likelihood decoder. These efficiency criteria naturally generalize several special cases that have received much attention in the literature—for example, the Infomax case ($p \to 0$) or the minimal mean squared error (MMSE) case ($p = 2$).

For each optimality criterion and a stimulus with a known prior, we analytically derived the optimal tuning curve for a single neuron. To extend this result to neural populations, we introduced the concept of the meta-tuning curve and showed that the optimal meta-tuning curve is identical to the optimal tuning curve for a single neuron with gaussian noise. These predictions based on different optimality criteria are tested against previously measured characteristics of several early visual systems for different

animals. Predictions corresponding to low values of p provide the best match, which suggests that the optimality criterion is near Infomax for the neural representations considered.

In our model and analysis, we have made the key assumption that the decoder is asymptotically unbiased. This implies that the results are strictly valid only in the low-noise regime, for example, when there is sufficient encoding time or a sufficient number of neurons. However, based on numerical simulations, we found that it is reasonably safe to relax the long encoding time assumption in particular if the neural population size is large or the optimal criterion parameter p is small.

Many behavioral studies also suggest that human and other animals make decisions that are often biased due to the effects of prior beliefs (Knill & Richards, 1996; Wei & Stocker, 2015). With numerical simulations, we showed that at short encoding times, the Bayesian MAPE decoder is indeed performing better than the unbiased MLE decoder and slightly better than our analytic predictions. In fact, the performance of the MLE is lower bounded by our theoretical predictions (solid lines in Figure 6), while the performance of the MAPE benefits from the prior information. Thus, our results are strictly valid only when assuming an MLE decoder.

In section 2.2, we analyzed the Poisson noise model and the constant gaussian noise model. Similar analysis can be applied to other noise models where the output variance depends on the output mean. For neural populations, we assumed that the output noise of an individual neuron is independent from the others, thus simplifying the computation of the total Fisher information of the population. If the output noise has a correlated structure, then the total Fisher information is no longer the sum of the Fisher information of individual neurons. Analysis of neural populations described by a meta-tuning curve with correlated noise is a subject for further investigation.

In conclusion, we believe that our results demonstrate the utility of exploring different reconstruction error criteria for analyzing neural responses in perceptual systems. The parameter p describes whether the neural system is adapted to more or less robust error statistics, and we have obtained some estimates of this parameter from data on early visual processing neurons in a number of different animals. It will be interesting to explore how the parameter p changes as information propagates through various stages of the perceptual system. We are also investigating how this analysis can be extended to higher-dimensional stimuli and to more complex noise models.

Appendix A: Fisher Information

The concept of Fisher information provides a statistical characterization of how well a random variable $\mathbf{r}$ can be used to estimate an underlying parameter s under a stochastic model $p(\mathbf{r}|s)$.

If a family of distributions $p(\mathbf{r}|s)$ is characterized by a one-dimensional parameter s, then the Fisher information is defined as (see Cover & Thomas, 1991),

$$\mathcal{I}(s) = \left\langle \left(\frac{d}{ds} \log p(\mathbf{r}|s) \right)^2 \middle| s \right\rangle_{p(\mathbf{r}|s)}. \tag{A.1}$$

A.1 Links to Popular Loss Functions

A.1.1 Mutual Information Limit. One possible measurement of neural coding quality is the mutual information. Measuring mutual information does not require an explicit estimator $\hat{s}(\mathbf{r})$. Instead, it directly measures the level of dependency between the neural response $\mathbf{r}$ and the input stimulus s. The link between mutual information $I_{\text{mutual}}(\mathbf{r}, s)$ and the Fisher information matrix was established in Brunel and Nadal (1998):

$$I_{\text{mutual}}(\mathbf{r}, s) = \frac{1}{2} \left\langle \log \mathcal{I}(s) \right\rangle_s + const. \tag{A.2}$$

Here we will not repeat the careful and delicate derivation, but the main idea is based on the fact that an efficient and unbiased estimator $\hat{s}$ is approximately gaussian with mean s and variance $\mathcal{I}(s)^{-1}$. The conditional entropy of such gaussian random variable is locally $1/2 \cdot \log(\mathcal{I}(s)^{-1}) + const$, and by averaging the local conditional entropy, we can get the mutual information. In terms of Fisher information matrix, we want to maximize the right side of equation A.2.

A.1.2 Cramer-Rao Lower Bound. Another possible way to measure coding quality is to use the L_2 norm to measure the error vector $\hat{s} - s$. Such an L_2 norm is related to the Fisher information matrix via the Cramer-Rao lower bound (Cover & Thomas, 1991). For any unbiased estimator $\hat{s}(\mathbf{r})$, for example, the maximum likelihood estimator (MLE),

$$\mathbf{Var}[\hat{s}(\mathbf{r}) - s \mid s] \geq \mathcal{I}(s)^{-1}. \tag{A.3}$$

As a lower bound, the Cramer-Rao bound can be attained by the MLE $\hat{s}(\mathbf{r})$ due to its asymptotic efficiency (Cover & Thomas, 1991).

In order to calculate the mean L_2 error, one can find the attainable lower bound locally at a given point s or globally averaged over all s:

$$\langle |\hat{s} - s|^2 \mid s \rangle_{\mathbf{r}} = \mathbf{Var}[\hat{s}(\mathbf{r}) - s \mid s] \geq \mathcal{I}(s)^{-1}, \tag{A.4}$$

$$\left\langle |\hat{s} - s|^2 \right\rangle_{\mathbf{r},s} \geq \left\langle \mathcal{I}(s)^{-1} \right\rangle_s. \tag{A.5}$$

Comparing this with equation A.2, we now derive another way of evaluating the Fisher information matrix. In order to minimize the mean L_2 error, one should minimize the right side of equation A.5. For a more complete work regarding the relationship between Fisher information and the Cramer-Rao lower bound, see Pilarski and Pokora (2015).

A.1.3 Asymptotic L_p Limit. A natural generalization of the L_2 metric to evaluate the difference $\hat{s} - s$ is the L_p metric for other values of p. In order to obtain the optimal L_p population code, one can instead solve the optimization problem to minimize the mean L_p norm of the difference $\hat{s} - s$ by evaluating the pth absolute moment of a gaussian random variable with zero mean and variance $\mathcal{I}(s)^{-1}$. Such a family of optimization problems parameterized by p can provide a natural connection between two traditional optimal criteria: the Infomax and MMSE (L_2-min):

$$\left\langle |\hat{s}(\mathbf{r}) - s|^p \right\rangle_{\mathbf{r},s} \approx const(p) \cdot \left\langle \mathcal{I}(s)^{-p/2} \right\rangle_s . \tag{A.6}$$

When $p = 2$, it is clear that the right side of equation A.6 is the same as the Cramer-Rao lower bound in equation A.5 up to some constant. In the limit of $p \to 0$, we can use the replica trick to show that minimizing the right side of equation A.6 is equivalent to maximizing the mutual information term in equation A.2:

$$\lim_{p \to 0} \frac{\mathcal{I}(s)^{-p/2} - 1}{p} = -\frac{1}{2} \log \mathcal{I}(s). \tag{A.7}$$

These characterizations of loss functions in equations A.2, A.5, and A.6 by using Fisher information simplifies the process of finding the optimal neural codes.

A.2 Fisher Information Examples. In order to apply the concept of Fisher information to analyze the performance of neural codes, here we calculate the Fisher information for a single neuron with a Poisson noise model or a constant gaussian noise model.

A.2.1 Poisson Spiking Model. The first model is the Poisson spiking model. If the neuron elicits a random number of spikes r during a given time window, ΔT is a Poisson random variable with rate $\Delta T \cdot h(s)$:

$$P(r = N|s) = \frac{1}{N!} \left(\Delta T \cdot h(s)\right)^N \exp(-\Delta T \cdot h(s)), \tag{A.8}$$

$$\log P(r = N|s) = -\log(N!) + N \log \left(\Delta T \cdot h(\mathbf{s})\right) - \Delta T \cdot h(\mathbf{s}), \tag{A.9}$$

$$\frac{d}{dx}\log P(r=N|s)=h'(\mathbf{s})\left(\frac{N}{h(s)}-\Delta T\right). \tag{A.10}$$

For Poisson random variable N with rate $\Delta T\cdot h(s)$, we know that

$$\langle N\rangle=\Delta T\cdot h(s),\quad \left\langle N^2\right\rangle=\Delta T\cdot h(s)+(\Delta T\cdot h(s))^2. \tag{A.11}$$

Using this result, we know

$$\begin{aligned}\mathcal{I}(s)=\left\langle\left(\frac{d}{ds}\log P(r=N|s)\right)^2\middle|s\right\rangle&=h'(s)^2\left\langle\left(\frac{N}{h(s)}-\Delta T\right)^2\right\rangle\\&=\Delta T\cdot\frac{h'(s)^2}{h(s)}.\end{aligned} \tag{A.12}$$

If the optimal Fisher information $\mathcal{I}(s)$ is known, the optimal nonlinearity $h(s)$ can be derived by solving the above ordinary differential equation:

$$h(s)\propto\left(\int_{-\infty}^{s}\sqrt{\mathcal{I}(\xi)}\,d\xi\right)^2. \tag{A.13}$$

A.2.2 Constant Gaussian Noise Model. In the second model, we assume the random number of spikes can be any real number. The additive noise in each unit time window is σ_0^2; therefore, the total number of spikes r_k that has been observed over a time window of length ΔT is a gaussian random variable with mean $\Delta T\cdot h(s)$ and variance $\sigma_0^2\Delta T$:

$$p(r|s)=\frac{1}{\sqrt{2\pi\sigma_0^2\Delta T}}\exp\left(-\frac{1}{2\sigma_0^2\Delta T}(r-\Delta T\cdot h(s))^2\right), \tag{A.14}$$

$$\log p(r|s)=-\frac{1}{2}\log(2\pi\sigma_0^2\Delta T)-\frac{1}{2\sigma_0^2\Delta T}(r-\Delta T\cdot h(s))^2, \tag{A.15}$$

$$\frac{d}{ds}\log p(r|s)=\frac{h'(s)}{\sigma_0^2}(r-\Delta T\cdot h(s)). \tag{A.16}$$

Using this result, we know that the Fisher information for a neuron with constant gaussian noise is

$$\mathcal{I}(s)=\left\langle\left(\frac{d}{ds}\log p(r|s)\right)^2\middle|s\right\rangle=\frac{h'(s)^2}{\sigma_0^4}\left\langle\left(r-\Delta T\cdot h(s)\right)^2\right\rangle=\frac{\Delta T}{\sigma_0^2}\cdot h'(s)^2. \tag{A.17}$$

If the optimal Fisher information $\mathcal{I}(s)$ is known, the optimal nonlinearity $h(s)$ can be derived by solving the above ordinary differential equation:

$$h(s) \propto \int_{-\infty}^{s} \sqrt{\mathcal{I}(\xi)}\, d\xi. \tag{A.18}$$

In equations A.12 and A.17, we have derived the Fisher information of a single neuron with a Poisson or constant gaussian noise model. In order to generalize from a single neuron to a population of neurons, we need the following result to measure the overall goodness of a population code.

A.3 Fisher Information for Neurons with Independent Noise. When each neuron in the population has independent noise, we prove that the total Fisher information of the population is the linear sum of the Fisher information contributed by each individual neuron:

$$p(\mathbf{r}|s) = \prod_{k=1}^{m} p(r_k|s) \quad \Rightarrow \quad I_{\text{total}}(s) = \sum_{k=1}^{m} I_k(s). \tag{A.19}$$

Using the definition of Fisher information, we know

$$I_{\text{total}}(s) = \left\langle \left(\frac{d}{ds} \log p(\mathbf{r}|s) \right)^2 \middle| s \right\rangle, \tag{A.20}$$

$$= \left\langle \left(\sum_{k=1}^{m} \frac{d}{ds} \log p(r_k|s) \right)^2 \middle| s \right\rangle. \tag{A.21}$$

When $k \neq l$, we know the neural response r_k, r_l are independent conditioned on s. Therefore:

$$\left\langle \frac{d}{ds} \log p(r_k|s) \cdot \frac{d}{ds} \log p(r_l|s) \middle| s \right\rangle = \left\langle \frac{d}{ds} \log p(r_k|s) \middle| s \right\rangle \cdot \left\langle \frac{d}{ds} \log p(r_l|s) \middle| s \right\rangle = 0, \tag{A.22}$$

which is because

$$\left\langle \frac{d}{ds} \log p(r_k|s) \middle| s \right\rangle = \int \frac{\frac{d}{ds} p(r_k|s)}{p(r_k|s)} \cdot p(r_k|s) dr_k = \frac{d}{ds} \left(\int p(r_k|s)\, dr_k \right) = 0. \tag{A.23}$$

As a conclusion, the total Fisher information for a population of neurons with independent Poisson/constant gaussian noise is equal to the linear sum of the Fisher information of each neuron.

Appendix B: Estimating the Distribution over Local Orientation

We extracted orientation statistics for natural images from a standard image database (van Hateren & van der Schaaf, 1998). First, we randomly sampled 200,000 square patches (16 pixels-by-16 pixels) across the entire database. We then created a set of sine-wave grating filters with a fixed spatial frequency that was close to the human peak sensitivity (approximately 4 cycles per visual degree or 8 pixels per cycle) but various phase and 360 different orientations (0° to 179.5° with 0.5° spacing). The dominant orientation of each patch was determined by the maximum response across all these filters. To mitigate the effect of pixel-wise noise or quantization effects, we used only patches with high filter response levels (top 50%). The resulting prior distribution (shown in Figure 8c) is very similar to previously measured distributions (e.g., Girshick et al., 2011) and is shown in Figure 8c. We used a spline function to fit the cumulative of the empirical histogram in order to obtain a smooth version of the density $f(\theta)$.

Appendix C: Equivalent-Michelson Contrast

Originally, the Michelson contrast was defined for sinusoid gratings based on its max/min luminance:

$$c = \frac{L_{\max} - L_{\min}}{L_{\max} + L_{\min}}. \tag{C.1}$$

It is clear the the Michelson contrast has a value between 0 and 1. For any patches of nonsinusoid gratings, we determine its equivalent Michelson contrast in the following way.

For each image patch, we use a set of 64 odd-Gabor filters $g_{\text{gabor}}(x, y)$ of different orientation θ and wavelength λ to convolute with natural image patches to obtain local responses. Specifically, the Gabor filters are

$$g_{\text{gabor}}(x, y) = g_{\text{normal}}(x, y) \cdot g_{\text{sinusoid}}(x, y), \tag{C.2}$$

$$g_{\text{normal}}(x, y) = \exp\left(-\frac{x'^2 + y'^2}{2\sigma^2}\right), \quad g_{\text{sinusoid}}(x, y) = \sin\left(2\pi \frac{x'}{\lambda}\right), \tag{C.3}$$

$$x' = x\cos\theta + y\sin\theta, \quad y' = -x\sin\theta + y\cos\theta,$$

$$\sigma = \frac{1}{\pi}\sqrt{\frac{\ln 2}{2}}\frac{2^b + 1}{2^b - 1}\lambda, \tag{C.4}$$

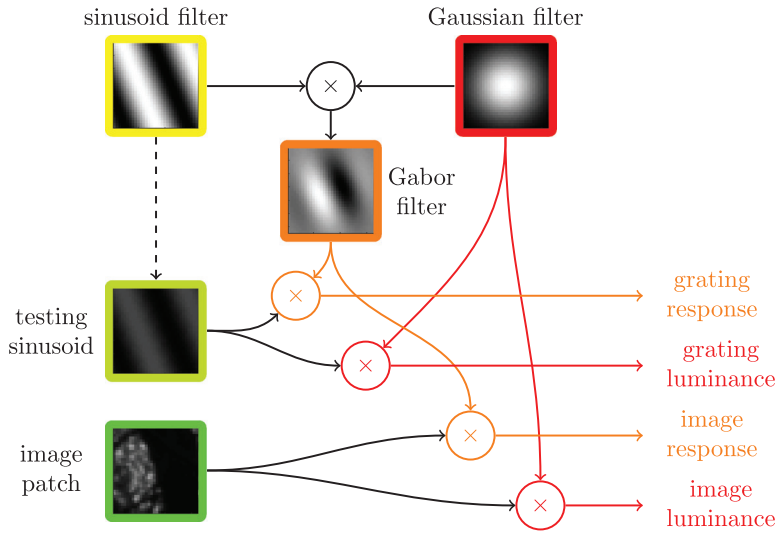


Figure 10: The process to determine equivalent-Michelson contrast for an image patch with respect to certain Gabor filter.

where the orientation θ takes eight values uniformly sampled from the range $[0, \pi]$ and the wavelength λ takes eight values uniformly sampled in the logarithm space from 4 to 85.3 pixels per cycle. The size of gaussian filter σ is automatically determined by the wavelength λ and a fixed octave value $b = 1.5$ in order to best match the properties of simple cells in the primary visual cortex.

With such a filter bank of 64 Gabor filters, we calculate the equivalent Michelson contrast for each image patch. For each Gabor filter, we use the corresponding gaussian filters $g_{\text{normal}}(x, y)$ to compute the local mean luminance to model luminance adaptation. We also use the corresponding sinusoid filter $g_{\text{sinusoid}}(x, y)$ to construct a testing sinusoid grating $L_{\text{ave}} + L_{\text{amp}} \cdot g_{\text{sinusoid}}(x, y)$. By properly choosing the parameters L_{ave} and L_{amp}, we can match both the Gabor filter response and the Gaussian filter response. The equivalent Michelson contrast is then determined by the Michelson contrast of this testing grating:

$$L_{\max} = L_{\text{ave}} + |L_{\text{amp}}|, \quad L_{\min} = L_{\text{ave}} - |L_{\text{amp}}| \quad \Rightarrow \quad c = \frac{|L_{\text{amp}}|}{L_{\text{ave}}}. \quad \text{(C.5)}$$

The process is summarized in Figure 10. The local contrast value of each image patch is then determined by taking the maximum among the 64 equivalent Michelson contrast values calculated using the Gabor filter bank. This

max operation is taken in order to match the normalization computation taking place in the visual perception pathway (Carandini & Heeger, 2012). Neurons that are responding to a low contrast value often appear to be silent (normalized out) when there is a neighbor neuron responding to a significantly larger contrast.

Acknowledgments

This work has been supported by grants from the Office of Naval Research and Air Force Office of Scientific Research. We also thank Xue-Xin Wei for fruitful discussion and feedback on the manuscript.

References

Atick, J. J., & Redlich, A. N. (1990). Towards a theory of early visual processing. *Neural Computation*, *2*(3), 308–320.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychol. Rev.*, *61*, 183–193.

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

Berens, P., Gerwinn, S., Ecker, A., & Bethge, M. (2009). Neurometric function analysis of population codes. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems, 22* (pp. 90–98). Red Hook, NY: Curran.

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

Bethge, M., Rotermund, D., & Pawelzik, K. (2003). Optimal neural rate coding leads to bimodal firing rate distributions. *Netw. Comput. Neural Syst.*, *14*, 303–319.

Brady, N., & Field, D. J. (2000). Local contrast in natural images: Normalisation and coding efficiency. *Perception*, *29*, 1041–1055.

Brenner, N., Bialek, W., & de Ruyter van Steveninck, R. (2000). Adaptive rescaling maximizes information transmission. *Neuron*, *26*, 695–702.

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

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Review Neuroscience*, *13*, 51–62.

Clatworthy, P. L., Chirimuuta, M., Lauritzen, J. S., & Tolhurst, D. J. (2003). Coding of the contrasts in natural images by populations of neurons in primary visual cortex (V1). *Vision Research*, *43*, 1983–2001.

Cover, T. M., & Thomas, J. (1991). *Elements of information theory*. Hoboken, NJ: Wiley.

de Ruyter van Steveninck, R., Lewen, G. D., & Bialek, W. (1997). Reproducibility and variability in neural spike trains. *Science*, *275*, 1805.

De Valois, R. L., Yund, E. W., & Hepler, N. (1982). The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research*, *22*, 531–544.

Dean, I., Harper, N. S., & McAlpine, D. (2005). Neural population coding of sound level adapts to stimulus statistics. *Nature Neuroscience*, *8*, 1684–1689.

Doi, E., & Lewicki, M. S. (2011). Characterization of minimum error linear coding with sensory and neural noise. *Neural Computation*, *23*, 2498–2510.

Dong, D. W., & Atick, J. J. (1995). Statistics of natural time-varying images. *Network: Computation in Neural System*, *6*, 345–358.

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

Ganguli, D., & Simoncelli, E. P. (2010). Implicit encoding of prior probabilities in optimal neural populations. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Carlotta (Eds.), *Advances in neural information processing systems, 23* (pp. 658–666). Red Hook, NY: Curran.

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*, 926–932.

Gjorgjieva, J., Sompolinsky, H., & Meister, M. (2014). Benefits of pathway splitting in sensory coding. *Journal of Neuroscience*, *34*, 12127–12144.

Grabska-Barwinska, A., & Pillow, J. W. (2014). Optimal prior-dependent neural population codes under shared input noise. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, 27* (pp. 1880–1888). Red Hook, NY: Curran.

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

Henry, G. H., Dreher, B., & Bishop, P. O. (1974). Orientation of cells in cat striate. *Journal of Neurophysiology*, *37*, 1394–1409.

Johnson, D., & Ray, W. (2004). Optimal stimulus coding by neural populations using rate code. *J. Comput. Neurosci*, *16*, 129–138.

Kang, K., Shapley, R. M., & Sompolinsky, H. (2004). Information tuning of populations of neurons in primary visual cortex. *Journal of Neuroscience*, *24*, 3726–3735.

Kastner, D. B., Baccus, S. A., & Sharpee, T. O. (2015). Critical and maximally informative encoding between neural populations in the retina. *Proc. National Acad., Sci. U.S.A.*, *112*, 2533–2538.

Knill, D. C., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge: Cambridge University Press.

Laughlin, S. B. (1981). A simple coding procedure enhances a neuron's information capacity. *Z. Naturforschung, 36c*(3), 910–912.

Linsker, R. (1989). An application of the principle of maximum information preservation to linear systems. In D. S. Touretzky (Ed.), *Advances in neural information processing systems, 1* (pp. 186–194). Cambridge, MA: MIT Press.

Maddess, T. M., & Laughlin, S. B. (1985). Adaptation of the motion-sensitive neuron H1 is generated locally and governed by contrast frequency. *Proc. R. Soc. Lond. B Biol. Sci.*, *225*, 251–275.

Mante, V., Frazor, R. A., Bonin, V., Geisler, W. S., & Carandini, M. (2005). Independence of luminance and contrast in natural scenes and in the early visual system. *Nature Neuroscience*, *8*, 1690–1697.

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

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

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387–391.

Naka, K. I., & Rushton, W. A. H. (1966). S-potentials from luminosity units in the retina of fish (Cyprinidae). *Journal of Physiology*, *185*, 587–599.

Nikitin, A. P., Stocks, N. G., Morse, R. P., & McDonnell, M. D. (2009). Neural population coding is optimized by discrete tuning curves. *Phys. Rev. Lett.*, *103*, 138101.

Ozuysal, Y., & Baccus, S. A. (2012). Linking the computational structure of variance adaptation to biophysical mechanisms. *Neuron*, *73*, 1002–1015.

Pilarski, S., & Pokora, O. (2015). On the Cramer-Rao bound applicability and the role of Fisher information in computational neuroscience. *BioSystems, 136*, 11–22.

Pouget, A., Deneve, S., Ducom, J.-C., & Latham, P. E. (1999). Narrow versus wide tuning curves: What's best for a population code? *Neural Computation, 11*, 85–90.

Ringach, D. L., Shapley, R. M., & Hawken, M. J. (2002). Orientation selectivity in macaque V1: Diversity and laminar dependence. *Journal of Neuroscience*, *22*, 5639–5651.

Salinas, E. (2006). How behavioral constraints may determine optimal sensory representations. *PLoS Biology*, *4*, 2383–2392.

Seung, H. S., & Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *Proc. National Acad. Sci. U.S.A.*, *90*, 10749–10753.

Sharpee, T. O., Sugihara, H., Kurgansky, A. V., Rebrik, S. P., Stryker, M. P., & Miller, K. D. (2006). Adaptive filtering enhances information transmission in visual cortex. *Nature*, *439*, 936–942.

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, *9:4*, 578–585.

Tadmor, Y., & Tolhurst, D. J. (2000). Calculating the contrasts that retinal ganglion cells and LGN neurones encounter in natural scenes. *Vision Research*, *40*, 3145–3157.

Theunissen, F. E., & Miller, J. P. (1991). Representation of sensory information in the cricket cercal sensory system. II. Information theoretic calculation of system accuracy and optimal tuning-curve widths of four primary interneurons. *J. Neurophysiol.*, *66*, 1690–1703.

Tkacik, G., Prentice, J. S., Balasubramanian, V., & Schneidman, E. (2010). Optimal population coding by noisy spiking neurons. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 14419–14424.

Twer, T. V. D., & MacLeod, D. I. A. (2001). Optimal nonlinear codes for the perception of natural colours. *Network: Computation in Neural Systems*, *12*, 395–407.

van Hateren, J. H. (1993). Spatiotemporal contrast sensitivity of early vision. *Vision Research*, *33*, 257–267.

van Hateren, J. H., & van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings: Biological Sciences*, *265*, 359–366.

Wang, Z., Stocker, 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). Red Hook, NY: Curran.

Wang, Z., Stocker, A., & Lee, D. D. (2013). Optimal neural population codes for high-dimensional stimulus variables. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, 26* (pp. 297–305). Red Hook, NY: Curran.

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

Yaeli, S., & Meir, R. (2010). Error-based analysis of optimal tuning functions explains phenomena observed in sensory neurons. *Front Comput Neurosci*, *4*, 130.

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

Zhang, K., & Sejnowski, T. J. (1999). Neuronal tuning: To sharpen or broaden? *Neural Computation, 11*, 75–84.

Received February 10, 2016; accepted July 25, 2016.