Bayesian Efficient Coding

Il Memming Park*
Dept. of Neurobiology & Behavior
Stony Brook University
memming.park@stonybrook.edu

Jonathan W. Pillow†
Princeton Neuroscience Institute
Princeton University
pillow@princeton.edu

Abstract

The efficient coding hypothesis, which proposes that neurons are optimized to maximize information about the environment, has provided a guiding theoretical framework for sensory and systems neuroscience. More recently, a theory known as the Bayesian Brain hypothesis has focused on the brain's ability to integrate sensory and prior sources of information in order to perform Bayesian inference. However, there is as yet no comprehensive theory connecting these two theoretical frameworks. Here we bridge this gap by formalizing a Bayesian theory of efficient coding. We define Bayesian efficient codes in terms of four basic ingredients: (1) a stimulus prior distribution; (2) an encoding model; (3) a capacity constraint, specifying a neural resource limit; and (4) a loss function, quantifying the desirability or undesirability of various posterior distributions. Classic efficient codes can be seen as a special case in which the loss function is the posterior entropy, leading to a code that maximizes mutual information, but alternate loss functions give solutions that differ dramatically from information-maximizing codes. In particular, we show that decorrelation of sensory inputs, which is optimal under classic efficient codes in low-noise settings, can be disadvantageous for loss functions that penalize large errors. Bayesian efficient coding therefore enlarges the family of normatively optimal codes and provides a more general framework for understanding the design principles of sensory systems. We examine Bayesian efficient codes for linear receptive fields and nonlinear input-output functions, and show that our theory invites reinterpretation of Laughlin's seminal analysis of efficient coding in the blowfly visual system.

1 Introduction

One of the primary goals of theoretical neuroscience is to understand the functional organization of neurons in the early sensory pathways and the principles governing them. Why do sensory neurons amplify some signals and filter out others? What can explain the particular configurations and types of neurons found in early sensory system? What general principles can explain the solutions evolution has selected for extracting signals from the sensory environment?

Two of the most influential theories for addressing these questions are the "efficient coding" hypothesis and the "Bayesian brain" hypothesis. The efficient coding hypothesis, introduced by Attneave and Barlow more than fifty years ago, uses the ideas from Shannon's information theory to formulate a theory normatively optimal neural coding [1, 2]. The Bayesian brain hypothesis, on the other hand, focuses on the brain's ability to perform Bayesian inference, and can be traced back to ideas from Helmholtz about optimal perceptual inference [3–7].

A substantial literature has sought to alter or expand the original efficient coding hypothesis [5, 8–18], and a large number of papers have considered optimal codes in the context of Bayesian inference [19–26]. However, the two theories have never been formally connected within a single, comprehensive theoretical framework. Here we propose to fill this gap by formulating a general Bayesian theory of efficient coding that unites the two hypotheses. We begin by reviewing the key elements of each theory and then describe a framework for unifying them. Our approach involves combining a prior and model-based likelihood function with a neural resource constraint and a loss functional that quantifies what makes for a "good" posterior distribution. We show that classic efficient codes arise when we use information-theoretic quantities for these ingredients, but that a much larger family of Bayesian efficient codes can be constructed by allowing these ingredients to vary. We explore Bayesian efficient codes for several important cases of interest, namely linear receptive fields and nonlinear response functions. The latter case was examined in an influential paper by Laughlin that examined contrast coding in the blowfly large monopolar cells (LMCs) [27]; we reanalyze data from this paper and argue that LMC responses are in fact better described as minimizing the average square-root error than as maximizing mutual information.

2 Theoretical Background

2.1 Efficient coding hypothesis

The Efficient Coding hypothesis, set forth by Attneave [1] and formalized by Barlow [2], proposes that sensory neurons are optimized to maximize the information they transmit about sensory inputs. This hypothesis represents one of the most influential theories in systems neuroscience, and was the first to apply Shannon's information theory to the problem of neural coding.

*https://orcid.org/0000-0002-4255-7750
†https://orcid.org/0000-0002-3638-8831

Mutual information, as defined by Shannon [28], quantifies (in units of bits) the information that neural responses $\mathbf{y}$ carry about external stimuli $\mathbf{x}$:

$$I(\mathbf{x},\mathbf{y}) \;=\; H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}), \tag{1}$$

where $H(\mathbf{y})$ is the marginal entropy and $H(\mathbf{y}|\mathbf{x})$ is the marginal and conditional (or "noise") entropy of the response:

$$H(\mathbf{y}) = -\int P(\mathbf{y}) \log P(\mathbf{y})\, \mathrm{d}\mathbf{y} \tag{2}$$

$$H(\mathbf{y}|\mathbf{x}) = -\int P(\mathbf{x},\mathbf{y}) \log P(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}. \tag{3}$$

The marginal entropy $H(\mathbf{y})$ quantifies the uncertainty (in bits) of the marginal response distribution $P(\mathbf{y})$, while conditional entropy $H(\mathbf{y}|\mathbf{x})$ quantifies the uncertainty of the conditional response distribution $P(\mathbf{y}|\mathbf{x})$, averaged over the joint distribution $P(\mathbf{x},\mathbf{y})$. The mutual information tells us how much uncertainty about $\mathbf{y}$ is reduced when we know the stimulus $\mathbf{x}$, on average. Remarkably, the mutual information is symmetric, so it can equally be written as $H(\mathbf{x}) - H(\mathbf{x}|\mathbf{y})$, the difference between stimulus entropy $H(\mathbf{x})$ and conditional entropy $H(\mathbf{x}|\mathbf{y})$, corresponding to the average reduction in uncertainty about the stimulus due to an observed neural response [29].

Barlow's proposal, which he termed the *redundancy-reduction hypothesis*, was that neurons maximize $I(\mathbf{x},\mathbf{y})/C$, the ratio between the mutual information between stimulus $\mathbf{x}$ and neural response $\mathbf{y}$, and the channel capacity C, which is an upper bound (considered over all stimulus distributions) on the mutual information. A perfectly efficient code in Barlow's sense is one for which $I(\mathbf{x},\mathbf{y}) = C$, where mutual information equals the channel capacity.

Barlow himself focused on the deterministic case where noise entropy $H(\mathbf{y}|\mathbf{x}) = 0$ and channel capacity is fixed. In this setting, efficiency is achieved by maximizing response entropy $H(\mathbf{y})$. This specific setting yields two predictions most commonly associated with the efficient coding hypothesis: (1) single neurons should nonlinearly transform the stimulus to achieve optimal use of their full dynamic range [27, 30, 31]; and (2) neural populations should decorrelate their inputs, so the marginal response distribution is more independent than their inputs [9, 32–35].

2.2 Bayesian Brain Hypothesis

The Bayesian Brain hypothesis provides a second theoretical perspective on neural coding [4, 6, 7]. The core idea can be traced back to Helmholtz's 19th century work on perception: it proposes that the brain seeks to combine noisy sensory information about the current stimulus with prior information about the environment. The product of these two terms, known as likelihood and prior, can be normalized to obtain the posterior distribution, which captures all information about the state of the environment. In this view, sensory perception is a form of Bayesian inference, and the brain's goal is to compute the posterior distribution over environmental variables given sensory inputs.

The two key ingredients for the Bayesian Brain hypothesis are a prior distribution $P(\mathbf{x})$ over the stimulus and an encoding distribution $P(\mathbf{y}|\mathbf{x})$ that describes the mapping from stimuli $\mathbf{x}$ to neural responses $\mathbf{y}$. These ingredients combine according to Bayes' rule to form the posterior distribution:

$$P(\mathbf{x}|\mathbf{y}) \propto P(\mathbf{x})P(\mathbf{x}|\mathbf{y}), \tag{4}$$

which captures the observer's beliefs about the stimulus $\mathbf{x}$ given the noisy sensory information contained in $\mathbf{y}$. This theory has had major influences on the study of sensory and motor behavior [6, 7, 26, 36–42], as well as on theories of neural population codes that support Bayesian inference [19–21, 43–45]. Despite its emphasis on normatively optimal perception and behavior, the Bayesian brain literature and the efficient coding hypothesis have not not yet been connected in full generality.

3 Bayesian efficient coding

Here we make the connection between the Bayesian brain hypothesis and efficient coding explicit by formulating a more general theory of Bayesian efficient coding, which include classic efficient coding as a special case. We will define a Bayesian efficient code (BEC) as resulting from four basic ingredients:

1. $P(\mathbf{x})$ - A stimulus distribution, or prior.
2. $P(\mathbf{y}|\mathbf{x},\theta)$ - An encoding model, parametrized by θ, describing how stimuli $\mathbf{x}$ are mapped to responses $\mathbf{y}$.
3. $C(\theta)$ - A capacity constraint on the parameters, specifying a neural resource limit.
4. $L(\cdot)$ - A loss functional, quantifying the desirability or undesirability of various posterior distributions.

Given these ingredients, a BEC corresponds to a setting of the model parameters θ that achieves a minimum of the expected loss,

$$\bar{L}(\theta) = \mathbb{E}\Big[L\Big(P(\mathbf{x}|\mathbf{y},\theta)\Big)\Big] = \int P(\mathbf{y}|\theta)\, L\Big(P(\mathbf{x}|\mathbf{y},\theta)\Big)\, \mathrm{d}\mathbf{y}, \tag{5}$$

subject to the capacity constraint $C(\theta) \leq \mathbf{c}$, where $P(\mathbf{x}|\mathbf{y},\theta) \propto P(\mathbf{y}|\mathbf{x},\theta)P(\mathbf{x})$ denotes the posterior over $\mathbf{x}$ given $\mathbf{y}$ and θ, and expectation is taken with respect to the marginal response distribution $P(\mathbf{y}|\theta) = \int P(\mathbf{y}|\mathbf{x},\theta)P(\mathbf{x})\mathrm{d}\mathbf{x}$.

Each of the four ingredients plays a distinct role in determining a Bayesian efficient code (see Fig. 1). The prior is determined by the statistics of the environment, and provides a complete description of the stimuli to be encoded. The model, in turn, determines the form of the probabilistic encoding function to be optimized; it determines the posterior distributions that the organism will utilize when the prior and likelihood are combined after each sensory response $\mathbf{y}^*$.

The capacity constraint C defines a neural resource limitation, such as an energetic constraint on the average spike count or a physiological limit on the maximal firing rate. This makes the problem of determining an optimal code well posed, since without a constraint it is often possible to achieve arbitrarily good codes, e.g., by using arbitrarily large spike counts to encode stimuli with arbitrarily high fidelity.

Finally, the loss functional L is the key component that sets Bayesian efficient codes apart from classic efficient codes: it quantifies how much to penalize different posterior distributions that may arise. For example, the brain might prefer posteriors with small entropy, or small variance, or small standard deviation. (These are not the same, as we shall see shortly.) Note that in our formulation, the loss functional applies to the entire posterior as a distribution function over $\mathbf{x}$, not (for example) a decoded point estimate of the stimulus, as in Bayesian estimation settings. The posterior entropy, for one, cannot be written as a function of the decoded estimate, making it desirable to consider loss functions that apply to the entire posterior distribution. We will discuss motivations for different loss functions in the following sections.

3.1 Bayesian vs. classical efficient codes

Given the above definition, it is natural to ask: when does a Bayesian efficient code correspond to a traditional efficient code as defined by Barlow? The answer is that classical Barlow efficient codes are a special case of Bayesian efficient codes with loss function set to the posterior entropy:

$$L\Big(P(\mathbf{x}|\mathbf{y},\theta)\Big) = -\int P(\mathbf{x}|\mathbf{y},\theta)\log P(\mathbf{x}|\mathbf{y},\theta)\,\mathrm{d}\mathbf{x}. \qquad (6)$$

This results in a code that maximizes mutual information between stimulus and response because the expected loss $\bar{L} = \mathbb{E}_{\mathbf{y}|\theta}[L]$ is the conditional entropy of the responses given the stimuli $H(\mathbf{y}|\mathbf{x};\theta)$; subtracting this from the prior entropy $H(\mathbf{x})$, which is is independent of θ, gives the mutual information $I(\mathbf{x},\mathbf{y})$. Thus, minimizing average posterior entropy is the same as maximizing mutual information.

However, there is nothing privileged about minimizing entropy or maximizing information from a general coding perspective. In the next sections we will show that it is often natural to consider other loss functions, and that the Bayesian efficient codes that result from doing so can differ strongly from classical, information-theoretically optimal codes.

4 Simple Examples

We will motivate a Bayesian theory of efficient coding by showing two simple examples that illustrate the appeal of using loss functions other than posterior entropy, one with continuous and one with discrete encoding.

4.1 Continuous example: 2D Gaussian

To illustrate the role played by the loss function in Bayesian efficient codes, we first consider a simple example with two noisy neurons encoding a bivariate Gaussian stimulus (Fig. 2A). Suppose that a 2D stimulus $\mathbf{x} = (x_1, x_2)$ has an independent Gaussian distribution with standard deviation 10 in both directions. Then consider three possible noisy encoders, each of which corresponds to making a measurement of $\mathbf{x}$ corrupted by additive Gaussian noise:

- **encoder 1**: $y_1 = x_1 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 100/99)$.
- **encoder 2**: $y_1 = x_1 + \epsilon_1, \quad y_2 = x_2 + \epsilon_2, \quad \epsilon_1 \sim \mathcal{N}(0, 100/49), \; \epsilon_2 \sim \mathcal{N}(0, 700/93)$.
- **encoder 3**: $y_1 = x_1 + \epsilon_1, \quad y_2 = x_2 + \epsilon_2, \quad \epsilon_1 \sim \mathcal{N}(0, 100/19), \; \epsilon_2 \sim \mathcal{N}(0, 100/19)$.

The first encoder makes a low-noise measurement of x_1 and ignores x_2, whereas the other two encoders distribute noise more or less evenly across x_1 and x_2, resulting in posteriors with standard deviations $(\sigma_1 = 1, \sigma_2 = 10)$, $(\sigma_1 = 2, \sigma_2 = 7)$, and $(\sigma_1 = 5, \sigma_2 = 5)$, respectively, as depicted in Fig. 2A.

It should be obvious from this example that there is no clear sense in which any of these posteriors can be declared *better in general*. The three posteriors differ in how uncertainty is distributed across the two stimulus dimensions; which is better depends entirely on what the organism cares about. To make this concrete, we consider three loss functions: the posterior entropy, $L_1 = -\log(2\pi e\,\sigma_1\sigma_2)$, (equivalent to maximizing mutual information), the total standard deviation, $L_2 = \sigma_1 + \sigma_2$, and the total variance, $L_3 = \sigma_1^2 + \sigma_2^2$. Each loss function gives a different best encoder. The first encoder achieves the smallest entropy, and therefore achieves the highest mutual information between stimulus and response, even though it entirely ignores x_2. The second encoder achieves minimal total-deviation loss, because 2+7=9 (encoder 2) is less than 1+10=11 (encoder 1) or 5+5=10 (encoder 3). The third encoder minimizes total-variance loss L_3, because $25 + 25 = 50$ is smaller than either $1 + 100 = 101$ (encoder 1) or $4 + 49 = 53$ (encoder 2).

This simple example illustrates the manner in which different loss functions give rise to different notions of optimal coding. The loss functions differ in how they penalize the allocation of uncertainty across stimulus dimensions. Entropy is only sensitive to the product $\sigma_1\sigma_2$, which corresponds to the volume of the posterior. We could stretch the posterior by an arbitrary constant a in one dimension and by $1/a$ in the other, and we would not affect entropy. The other two loss functions, on the other hand, seek to minimize the summed uncertainty (standard deviation or variance) along the two axes. Compared to the entropy, they disfavor posteriors with large uncertainty in one direction, an effect that is larger for variance than for standard deviation. For Gaussian posteriors, they can also be interpreted in terms of minimizing error in an optimal decoder. Minimizing total standard deviation is equivalent to finding an encoder that minimizes the summed absolute error of a decoded estimate:

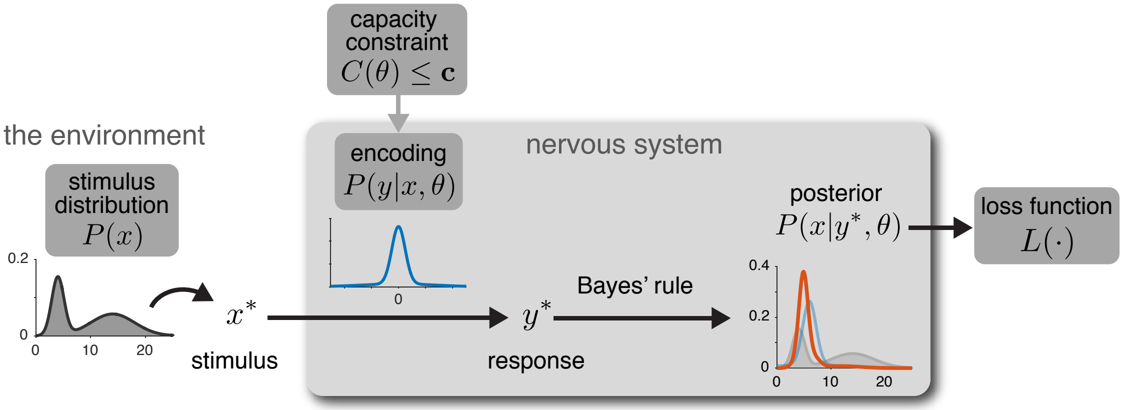

Figure 1: Bayesian efficient coding schematic. The theory is governed by four basic ingredients, highlighted in dark gray boxes. During any perceptual interval, a stimulus $\mathbf{x}^*$ is drawn from the prior $P(\mathbf{x})$ and presented to the organism. The nervous system encodes this stimulus with a sample $\mathbf{y}^*$ from the encoding distribution $P(\mathbf{y}|\mathbf{x}^*,\theta)$, which is governed by parameters θ (e.g., defining a neuron's receptive field, nonlinearity, noise, etc). An application of Bayes' rule leads to the posterior distribution $P(\mathbf{x}|\mathbf{y}^*,\theta)$, which captures all information available to downstream brain areas about the stimulus given the sensory response $\mathbf{y}^*$. The loss function $L(\cdot)$ characterizes the desirability of this posterior. A Bayesian efficient code is one for which parameters θ are set to minimize the average loss over stimuli and responses drawn from the prior and encoding model, subject to the resource constraint on the encoder, $C(\theta) < \mathbf{c}$.

$\mathbb{E}[||\hat{\mathbf{x}} - \mathbf{x}||_1] = \mathbb{E}[|\hat{x}_1 - x_1| + |\hat{x}_2 - x_2|]$, where $\hat{\mathbf{x}}$ is an estimate that minimizes the ℓ_1 error $||\hat{\mathbf{x}} - \mathbf{x}||_1$. Similarly, minimizing total variance is equivalent to minimizing the mean squared error of an optimal decoder: $\mathbb{E}[||\hat{\mathbf{x}} - \mathbf{x}||_2^2] = \mathbb{E}[(\hat{x}_1 - x_1)^2 + (\hat{x}_2 - x_2)^2]$, where $\hat{\mathbf{x}}$ is the posterior mean[1], also known as the Bayes' least squares (BLS) or minimum mean squared error (MMSE) estimator.

We could of course have considered other loss functions that would have given us different reasons for preferring any of these three encoders. For example, a "minimax" loss function $L = \max(\sigma_1, \sigma_2)$, which cares only about minimizing the worst possible performance in any dimension, would also favor the third encoder. Thus, optimality is in the eye of the loss function, and for any set of encoders there may be multiple ways to regard them as optimal in a Bayesian sense.

4.2 Discrete example: multiple choice exam

As second example, consider the case of a discrete stimulus that takes on one of four possible values $\{a, b, c, d\}$, each with prior probability 0.25, and a noisy neuron that can respond with 1, 2, 3, or 4 spikes. We will consider the following two possible encoding rules (See Table 1 and Fig. 2B).

$P(y|x)$

y \ x	a	b	c	d
1	0.5	0.5	0	0
2	0.5	0.5	0	0
3	0	0	0.5	0.5
4	0	0	0.5	0.5

(a) encoder 1

$P(y|x)$

y \ x	a	b	c	d
1	0.8	0.2/3	0.2/3	0.2/3
2	0.2/3	0.8	0.2/3	0.2/3
3	0.2/3	0.2/3	0.8	0.2/3
4	0.2/3	0.2/3	0.2/3	0.8

(b) encoder 2

Table 1: Discrete encoding for multiple choice test

The first encoder maps stimuli a and b to responses 1 and 2 randomly with equal probability, and similarly maps stimuli c and d to responses 3 and 4. The second encoder, on the other hand, maps $a \to 1$, $b \to 2$, $c \to 3$, $d \to 4$ with probability 0.8, and with probability 0.2 maps to one of the other three responses selected at random. Fig. 2B shows the kinds of posteriors that arise under these two encoders—these are in fact equal to the rows of the encoding tables given above, since the prior is uniform and each row already sums to 1. For the first encoder, the posterior assigns probability 0.5 to two stimuli and rules out the other two. For the second encoder, the posterior concentrates on one stimulus with probability 0.8, and spreads the remaining 0.2 evenly across the other three stimuli.

It is now interesting to consider which of these two encoders is best according to different loss functions. First, let us consider posterior entropy, which corresponds to maximizing information as in Barlow's classic definition. The prior has an entropy $H(x) = -4 \times 0.25 \log_2 0.25 = 2$ bits. The mean posterior entropy of the first encoder $H(x|y,\theta_1) = -2 \times 0.5 \log_2 0.5 = 1$ bit, so the mutual information between stimulus and response is 1 bit. By contrast, the posterior entropy of the second encoder is $H(x|y,\theta_2) = -0.8 \log_2 0.8 - 0.2 \log_2(0.2/3) = 1.04$ bits, which gives mutual information of only $2 - 1.04 = 0.96$ bits. This means that the second encoder preserves strictly less Shannon information about the stimulus than the first, so the first encoder is more efficient[2]. A second natural choice of loss function, however, is the "percent correct", defined as the percent of the time a *maximum a posteriori* decoder chooses the correct stimulus.

[1] Note that this connection between sum of moments and minimization of error does not hold in general. See supplemental information for detail.

[2] Strictly speaking the two channels have different capacity, the definition of redundancy doesn't apply directly.

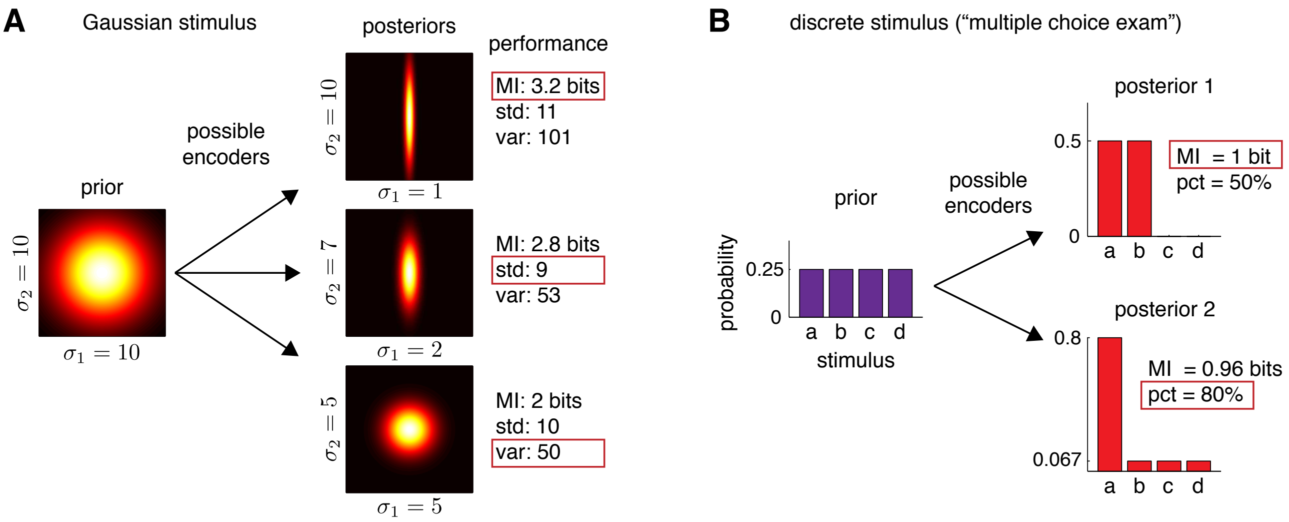

Figure 2: Optimal encoding depends critically on choice of loss function. (A) Illustration showing three possible encoders of a stimulus with an independent bivariate Gaussian prior distribution. The encoders produce three different Gaussian posteriors, shown at right. For axis-aligned Gaussian distributions, the entropy depends on the product of standard deviations $\sigma_1\sigma_2$, whereas the "total deviation" depends on the sum of standard deviations $\sigma_1 + \sigma_2$, and total variance depends on the sum of variances $\sigma_1^2 + \sigma_2^2$. According to entropy loss, the top encoder is best (achieving a mutual information of $\log_2(10 \cdot 10) - \log_2(10 \cdot 1) = 3.2$ bits), but for total-deviation loss, the middle encoder is best (achieving $\sigma_1 + \sigma_2 = 9$), and for total-variance loss, the bottom encoder is best (achieving $\sigma_1^2 + \sigma_2^2 = 50$). **(B)** Discrete encoding example, showing two possible encoders for a multiple-choice exam. The prior is an equal $p = 1/4$ probability for each of the four possible stimuli $\{a, b, c, d\}$. The top encoder eliminates two possibilities so that the posterior assigns probability $p_i = 1/2$ to two stimuli and 0 to the other two. The bottom encoder gives a posterior distribution with probability $p_i = 0.8$ for one stimulus and the remaining 0.2 probability spread evenly among the other three. The top encoder is optimal for information theoretic loss, since it achieves a mutual information of 1 bit, vs. only 0.96 bits for the bottom encoder. However, for "percent correct" loss, which is sensitive only to $\max(\{p_i\})$, the bottom encoder is clearly better. It identifies the correct stimulus 80% of the time (good for a grade of B-), whereas the top encoder achieves only 50% (an F).

According to this loss function, the second encoder is clearly superior, since 80% of the time it concentrates on the correct stimulus, whereas the best possible decoding of responses from the first encoder can only answer correctly 50% of the time.

This example has perhaps greater cultural and psychological salience if we reframe it terms of students studying for a multiple choice exam. Each question on the exam will have four possible choices (a, b, c, and d), which occur equally often. Student 1 adopts a study strategy that allows her to rule out two of the four choices with absolute certainty, but to have total uncertainty about which of the two remaining options is correct for each exam question. Student 2, on the other hand, adopts a study strategy that allows her to know the correct answer 80% of the time, but has uniform uncertainty about the remaining three options, which are correct the remaining 20% of the time. Which student's strategy is better? If we judge them according to mutual information, the first student has clearly learned more; her brain has stored 0.04 more bits about the subject matter than the second student. However, if we judge them according to the number of questions they can answer correctly on the exam, the second student's strategy is clearly better: her expected grade is a B-, with a score of 80%, whereas the second student is expected to fail with only half the questions answered correctly.

An interesting corollary of this example, therefore, is that although information-theoretic learning (*cf.* [12, 46]) is optimal for True/False exams, it can be substantially sub-optimal for multiple-choice exams.

5 Linear receptive fields

Now we turn to an application motivated by biology. Neurons in the early visual system are often described as performing an approximately linear transformation of the light pattern falling on the retina. A large body of previous work has examined the optimality of these linear weighting functions under the efficient coding paradigm and its variants [10, 15, 16, 18, 33, 47–52].

Here we re-examine this problem through the lens of Bayesian efficient coding. We consider the following simplified model for linear encoding of sensory stimuli:

$$\text{stimulus distribution: } \mathbf{x} \sim \mathcal{N}(0, Q); \tag{7}$$

$$\text{(noisy) encoding model: } \mathbf{y} = W\mathbf{x} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, R) \tag{8}$$

$$\text{s.t. } \mathbb{E}[\mathbf{y}^\top \mathbf{y}] \leq c. \quad \text{("power constraint")} \tag{9}$$

In this setup, we assume that the stimulus, an image consisting of n pixels, is encoded into a population of n neurons via a weight matrix W. Each row of W corresponds to the linear receptive field of a single neuron. For tractability, we assume that the stimulus has a Gaussian distribution (with covariance Q defining the correlations between pixels), and the neural population response is corrupted by additive Gaussian noise with covariance R. We impose a power constraint on the response, corresponding to a bound on the sum of squared responses $\mathbf{y}$. This is a common choice of constraint in the efficient coding and engineering literature due to differentiability and analytic tractability [29, 47]. In practice, the power constraint restricts the model from achieving infinite signal-to-noise ratio by growing W without bound.

For this model, the marginal response distribution $P(\mathbf{y})$ is Gaussian, $\mathbf{y} \sim \mathcal{N}(0, WQW^\top + R)$, so the power constraint can be written in terms of the trace of the marginal covariance:

$$\mathbb{E}[\mathbf{y}^\top \mathbf{y}] = \mathbb{E}[\mathrm{Tr}[\mathbf{y}\mathbf{y}^\top]] = \mathrm{Tr}[WQW^\top + R] \leq c. \tag{10}$$

The posterior distribution $P(\mathbf{x}|\mathbf{y})$ is also Gaussian, $\mathcal{N}(\mu, \Sigma)$, with mean and covariance

$$\mu = \Sigma W^\top R^{-1}\mathbf{y}, \quad \Sigma = (W^\top R^{-1} W + Q^{-1})^{-1}. \tag{11}$$

In this setup, the coding question of interest is: given stimulus covariance Q, noise covariance R, and power constraint c, what is the optimal linear encoding matrix W? That is, what receptive fields are optimal for encoding of the stimuli from $P(x)$ in the face of noise and a constraint on response variance? Here we consider two possible forms for the loss functional:

$$L_{\text{entropy}} = H(\mathbf{x}|\mathbf{y}) = \tfrac{1}{2}\log|2\pi e\Sigma| = \tfrac{1}{2}\sum_i \log\sigma_i + \boldsymbol{c} \tag{12}$$

$$L_{\text{covtropy}} = \mathrm{Tr}[\Sigma^{\frac{p}{2}}] = \sum_i \sigma_i^p \tag{13}$$

where σ_i^2 are the eigenvalues of posterior covariance Σ, or the variances of the posterior along its principal axes.

The first loss function, posterior entropy, corresponds to classic *infomax* efficient coding, since minimizing posterior entropy corresponds to maximizing mutual information between stimulus and response. The second loss function, which we term *covtropy* due to its similarity to entropy, is the summed p'th powers of posterior standard deviation along each principal axis. Like entropy, the covtropy depends only on the eigenvalues of the posterior covariance matrix, and is thus invariant to rotations (See Supplementary Information). For $p = 2$, covtropy corresponds to the sum of posterior variances along each axis; minimizing it is equivalent to minimizing the mean squared error [15–17] (see Appendix for proof).

For $p = 1$, covtropy corresponds to a sum of standard deviations; this penalizes a posterior with large variance in one direction less severely than covtropy with $p = 2$. In the limit $p \to 0$, minimizing covtropy is identical to maximizing mutual information, since $\log\sigma = \lim_{p\to 0}\frac{1}{p}\sigma^p$. In this limit, the optimal code minimizes the sum of log standard deviations, $\sum_i \log\sigma_i$, which is equivalent to minimizing the product $\prod_i \sigma_i$. The other interesting regime to consider is the limit $p \to \infty$. In this regime, minimizing covtropy is equivalent to minimizing $\max_i\{\sigma_i\}$, the maximal posterior standard deviation, a form of “mini-max” encoding. The optimal code is therefore the one that achieves smallest maximal posterior standard deviation in any direction.

Note that for the linear Gaussian model considered here, the posterior covariance Σ is independent of the response $\mathbf{y}$. This makes the problem easier to analyze because there is no need to compute average loss over the response distribution $P(\mathbf{y})$ (eq. 5). We show here that there is an analytic solution for the optimal linear encoding matrix W in this setting, for both infomax loss and covtropy loss (with any choice of $p > 0$), if we assume that Q, R, and W have a common diagonalization. This condition arises naturally for a convolutional code, that is, W contains a single receptive field shape that is circularly shifted by one pixel for each neuron in the population, and noise is spatially shift invariant (i.e., R is a circulant matrix). In the following, we examine the properties of Bayesian efficient linear receptive field codes for both information-theoretic and covtropy loss functions (derivation in the Supplementary Information).

5.1 Infomax encoding

The optimal weight matrix W for infomax coding is given by:

$$W_{\text{MI}} = \left(\tfrac{c}{d}I - R\right)^{\frac{1}{2}} Q^{-\frac{1}{2}}, \tag{14}$$

where c is the power constraint and d is the number of stimulus dimensions. For this encoder, the responses $\mathbf{y}$ are perfectly whitened, meaning the response covariance is proportional to the identity, $\mathbb{E}[\mathbf{y}\mathbf{y}^\top] \propto I$, and the posterior covariance of $\mathbf{x}|\mathbf{y}$ is proportional to the product of signal and noise covariances, $\Sigma \propto (QR)$.

5.2 Minimum covtropy encoding

For covtropy loss with exponent p, the optimal encoding weights W are given by

$$W_p = \left(\alpha(RQ)^{\frac{p}{p+2}} - R\right)^{\frac{1}{2}} Q^{-\frac{1}{2}}, \tag{15}$$

where constant $\alpha = c / \left(\mathrm{Tr}\left[(RQ)^{\frac{p}{p+2}}\right]\right)$ simply enforces the power constraint. For these weights, the marginal covariance is

$$\mathbb{E}[\mathbf{y}\mathbf{y}^\top] \propto (QR)^{\frac{p}{p+2}} \tag{16}$$

which implies that optimal responses can be more or less correlated than the stimulus. For example, when noise covariance R is proportional to stimulus covariance Q, then for $p = 2$ (minimum mean-square error encoding), the optimal receptive fields perfectly preserve the correlations in the stimulus, that is $\mathrm{cov}[Y] \propto Q$. For $p > 2$, the optimal receptive fields *increase* correlations so that responses are *more* correlated than the stimuli (See Fig. 3).

Although the responses become more correlated with increasing p, the posterior covariance, given by

$$\Sigma \propto (QR)^{\frac{1}{p+1}} \tag{17}$$

becomes increasingly whitened (i.e., closer to identity) with increasing p, due to the fact that $\lim_{a\to 0} M^a = I$ for any non-singular matrix M. This results in a posterior with the same

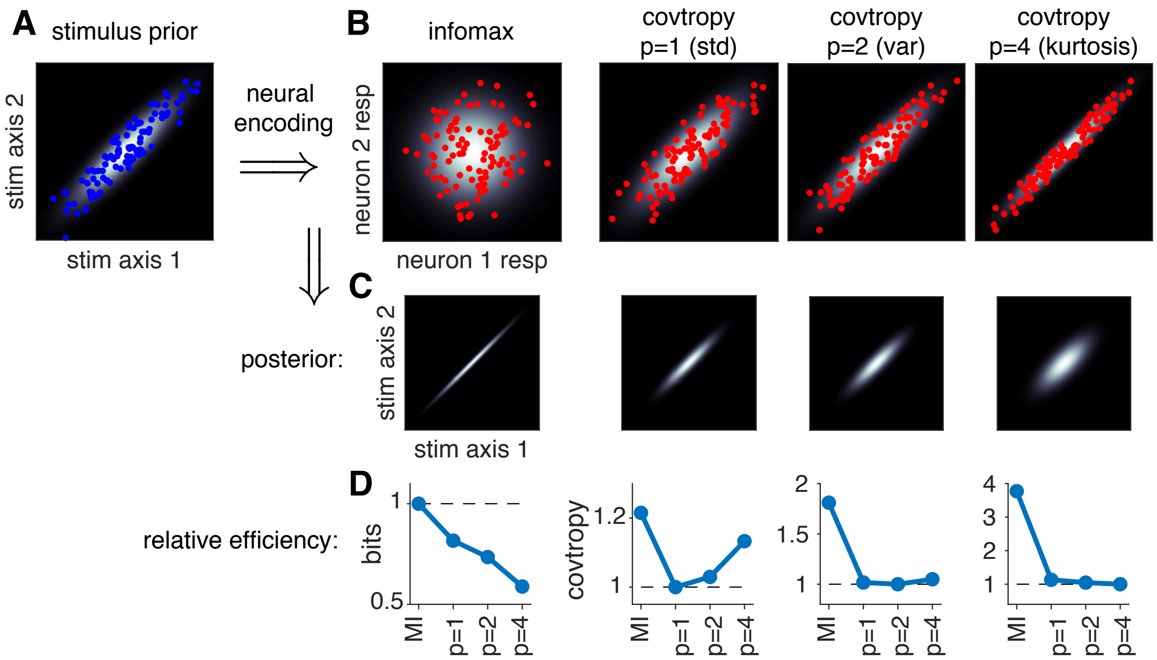

Figure 3: Bayesian efficient codes for a linear receptive field model with Gaussian noise. (A) Gaussian stimulus distribution with strong correlations ($\rho = 0.9$), with 50 samples shown (blue dots). **(B)** Neural responses (red dots) corresponding to optimal representations under infomax and covtropy loss functions, under infinitesimal Gaussian noise with covariance proportional to the stimulus distribution. Responses are uncorrelated ("white") for the infomax encoder, but unchanged for the $p = 2$ (minimum variance) encoder and *stronger* for $p = 4$. **(C)** Posterior distribution shape for each optimal encoder, showing that the infomax encoder achieves small posterior entropy using a long narrow posterior, whereas optimal covtropy posteriors grow more spherical with increasing p. **(D)** Relative performance of each encoder according to the loss function considered, normalized so the optimum is 1. Each encoder does best according to its own loss function, and the fall-off in efficiency between infomax and covtropy encoders is substantial; the infomax coder captures almost twice as much information as the $p = 4$ covtropy encoder in terms of bits (left plot), while the infomax coder exhibits nearly 4 times higher $p = 4$ covtropy than the optimal $p = 4$ encoder (right plot).

uncertainty in all directions, regardless of the prior or noise covariance, in accordance with the "minimax" property for $p = \infty$ noted above.

6 Nonlinear response functions

So far we have focused on encoders that linearly transform the stimulus. However, many neurons exhibit rectifying or saturating nonlinearities that map the raw stimulus intensities so as to make optimal use of a neuron's limited response range. Barlow's efficient coding theory states that the nonlinearity should map the stimulus distribution so as to maximize information between stimulus and response [2, 31], which naturally depends on the prior distribution over stimuli, the conditional response distribution, and the particular constraint on neural responses. For the case of noiseless responses and a simple constraint over the range of allowed responses, the shape of the information-maximizing nonlinearity is proportional to the cumulative distribution function (CDF) of the stimulus distribution, producing a uniform marginal distribution over responses [53].

In a groundbreaking paper that is widely considered the first experimental validation of Barlow's theory, Simon Laughlin [27] measured graded responses from blowfly large monopolar cells (LMC) to contrast levels measured empirically in natural scenes. Laughlin found that the LMC response nonlinearity, measured with responses to sudden contrast increments and decrements, exhibited a striking resemblance to the shape of the empirically measured CDF of contrast levels in natural scenes. Fig. 4A shows a reproduction of the original figure, showing that the nonlinear response is closely matched with the stimulus distribution expected by the infomax solution.

Here we reexamine Laughlin's conclusions by analyzing the same dataset through the enlarged framework of Bayesian efficient coding. We can formalize the coding problem as follows:

$$\text{stimulus distribution: } x \sim P(x) \tag{18}$$

$$\text{encoding model: } y = g(x); \tag{19}$$

$$\text{s.t. } y \in \{0, \ldots, y_{max}\} \tag{20}$$

Here x is a scalar stimulus with prior distribution $P(x)$, and the encoding model is described by $g(x)$, a noiseless, quantizing transformation from stimulus to discrete response levels. Note that some information about the stimulus is lost due to quantization error. For classical infomax encoding, we know that optimal g performs histogram equalization by taking on the quantiles of $P(x)$, resulting in a uniform marginal response distribution $P(y)$ [27, 53].

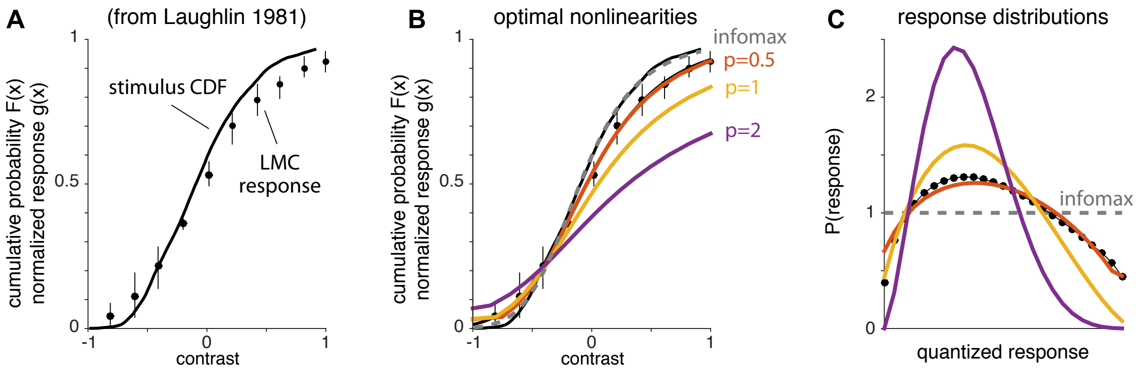

Figure 4: Bayesian efficient coding for nonlinear transformations. (A): Measured large monopolar cell (LMC) response from Laughlin [27] and the cumulative distribution function of the stimulus statistic. The two lines should be identical for optimal encoding under the infomax criterion. **(B)** Sigmoid fits for curves in panel A, and four predicted nonlinear encoders based on the sigmoid fit of the stimulus distribution. **(C)** Predicted response marginals under four different optimal codes using $P(x)$ in (B), and marginal estimated from empirically measured g.

We extracted the stimulus distribution from (Laughlin [27], Fig. 2) in order to determine the optimal nonlinearity under Bayesian efficient coding paradigms for several different choices of loss function. In particular, we considered the p-th power loss function of the form:

$$L = \mathbb{E}\left[|x - \hat{x}(y)|^p\right] = \int |x - \hat{x}(y)|^p \, P(x|y)\, \mathrm{d}x \qquad (21)$$

where $\hat{x}(y)$ is the Bayesian optimal decoder for x from y [54]. For $p = 2$, this loss is equal to mean squared error, and estimate $\hat{x}(y)$ is equal to the mean of the posterior over x given y. Smaller values of p correspond to reducing the relative penalty on large errors and increasing the relative penalty on small errors; in the limit of $p \to 0$, the loss converges to posterior entropy, making it equivalent to the infomax setting [24, 54].

We used numerical optimization to find the optimal nonlinearity $g(x)$ for $p = \frac{1}{2}, 1, 2$ (see Methods). We plotted these Bayes optimal nonlinearities against the infomax nonlinearity from Laughlin, as well as the empirically measured LMC response nonlinearity (Fig. 4B). Although the infomax nonlinearity resembles the true nonlinearity by eye, close inspection reveals the BEC nonlinearity with $p = \frac{1}{2}$ provides a much closer fit. This means that the LMC response function is more accurately described as minimizing average decoding error raised to the one-half power than as maximizing information.

The difference between infomax and $p = \frac{1}{2}$ loss appears slight when viewed in terms of the optimal nonlinearity $g(x)$, but is more dramatic when we consider the predicted marginal distribution over responses $P(y)$ (Fig. 4C). We fit the neuron's true response nonlinearity with a sigmoid function (Naka-Rushton CDF; See Methods), and computed the predicted marginal response distribution $P(y)$ for each nonlinearity $g(x)$. As expected, the infomax nonlinearity (dashed trace) produces a flat (uniform) distribution over response levels. However, both the $p = \frac{1}{2}$ nonlinearity and the predicted LMC response distribution exhibit a noticeable peak around intermediate response levels, with intermediate response levels used more often, and responses at the extreme left and right tails used less often. We note that such peaking is expected under loss functions that penalize the magnitude of decoding errors. In fact, the optimal nonlinearities for $p = 1$ and $p = 2$ generate responses distributions that are even more peaked than the real data, since they assign less probability mass to outermost response levels, where decoding errors are largest. The infomax loss function, by contrast, ignores the size of errors and simply seeks to match quantiles of the stimulus distribution.

7 Discussion

We have synthesized Barlow's efficient coding hypothesis with the Bayesian brain hypothesis in order to formulate a Bayesian theory of efficient neural coding. In this theory, an efficient neural code corresponds to an encoding model that produces optimal posteriors over stimuli under a capacity constraint on the neural response. The optimality of the posterior is determined by a loss function L, and the choice of loss function can have major effects on the optimal code. We have shown that Barlow's original theory corresponds to a special case of this theory in which the loss function is the Shannon entropy of the posterior, corresponding to codes that maximize mutual information between stimulus and response. However, such codes may be inefficient with respect to loss functions that are sensitive to the shape of the posterior, or the size of different decoding errors.

To illustrate our framework, we have derived Bayesian efficient codes for two canonical neural coding problems: (1) population encoding of high-dimensional stimuli with linear receptive fields; and (2) single-neuron encoding of low-dimensional stimuli with a nonlinear response function. In the first case, we showed that the "whitening" solution favored by classic efficient coding can be sub-optimal for other loss functions, and that even in the high-SNR (signal-to-noise ratio) regime, optimal codes may *increase* the correlations between neurons. In the second case, we re-examined the classical results of Laughlin, and showed that the nonlinear response functions of the blowfly LMC neurons are in fact more consistent with a Bayesian efficient code that minimizes the average square root of the decoding error than a code that maximizes mutual information. While these two examples are both highly simplified, they illustrate the power of this general formulation of efficient coding, and show surprising and non-intuitive results that contrast with previous findings.

7.1 Relationship to previous work

In the years since the seminal papers of Attneave and Barlow, a large literature has taken up the problem of optimal neural coding, spanning a wide range of both information and non-information-theoretic frameworks. Barlow's original paper considered only noiseless encoding, where each stimulus maps to a deterministic response [2]. Atick and Redlich extended this framework to incorporate noise, deriving the remarkable result that optimal receptive field shapes change with SNR, consistent with observed changes in retinal ganglion receptive fields [10]. They showed that whitening is optimal only at high SNR, and that optimal responses become correlated at lower SNR—note that this differs from our result showing that whitening can be sub-optimal even at high SNR for other loss functions (Figs. 3 & 5).

Several alternate versions of efficient coding based on information theory were advanced in the early years. Barlow first proposed that neurons optimize a quantity he called *redundancy*, given by $1 - I(\mathbf{x}, \mathbf{y})/C$, or one minus the mutual information divided by the channel capacity C [2]. (In fact, Barlow referred to his theory as the "redundancy reduction hypothesis", and the more common "efficient coding hypothesis" label has only appeared more recently, e.g., [55]) Atick and Redlich modified Barlow's theory to replace the C in the denominator by C_{out}, a modified notion of channel capacity related to the system's total power, which they held to be more biologically plausible [10, 48]. Neither theory therefore amounted to a pure "infomax" hypothesis, such as that advanced by [8], since optimality could be increased either by increasing mutual information or decreasing the channel capacity C or C_{out} in the denominator.

It bears mentioning that Barlow's definition of redundancy has no relationship to the concepts of redundancy and synergy later introduced by Brenner and colleagues, which quantify whether groups of neurons encode more or less information jointly than separately [56–59]. An efficient code according to Barlow or Atick & Redlich may be perfectly efficient in the sense of maximizing the ratio of information to capacity, while being either redundant or synergistic in the (more widely used) sense defined by Brenner *et al* [60].

Theories of efficient coding based on information-theory have been applied to a wide variety of different sensory systems and brain areas. These can be roughly grouped into those focused on linear receptive fields [32, 35, 61, 62], those focused on tuning curves [23–26, 63], and those addressing some aspect of population coding, such coupling strengths between neurons [64, 65] or use of multiple pathways [51, 66–68]. A substantial portion of this literature work has approached the problem of optimal coding through the lens of Fisher information [23, 24, 26, 63, 67, 69], although Fisher information may not accurately quantify coding performance in low SNR settings (e.g., short time windows or low firing rates) [17, 22, 25, 70].

A substantial literature has also considered optimal coding under alternate loss functions, and recent work has shown that codes optimized for mutual information may perform poorly for non-information-theoretic loss functions [71]. The most well-studied alternate loss function is the squared loss, $\mathbb{E}[(x-\hat{x}(y))^2]$, which results in so-called "minimum mean squared error" codes; such codes achieve minimum expected posterior variance (as opposed to minimum posterior entropy). These codes have received substantial attention in both engineering [72, 73] and neuroscience [15, 16, 18, 22, 25, 74]. Optimal codes for a wide variety of other losses or optimality criteria have also been proposed, including: minimization of motor errors [14], maximization of accuracy [75–77], optimal coding for control applications [78], and optimal future prediction [12, 79–83], and loss based on natural selection [6]. Other recent work has considered codes that maximize metabolic efficiency [11, 13], which in our framework corresponds more naturally to the constraint (which is concerned with use of finite resources) than the loss function (which is concerned with the posterior distribution).

In the statistics and decision-making literature, our work is closely related to Bayesian statistical decision theory [84–86], although such work has tended not to consider the problem of optimal sensory coding. One noteworthy difference between our framework and classical Bayesian decision-making theory is that the loss functions in decision theory are typically defined in terms of an expected cost of making a categorical decision or a point estimate. In our framework, by contrast, the loss is defined as a functional of the posterior; this allows us to consider a wider class of loss functions such as the posterior entropy or the average posterior standard deviation, which cannot in general be written as an expectation of a cost function involving the true stimulus and its estimate. Thus, our loss function merely specifies what counts as a good or desirable posterior distribution, without assuming how the posterior will be used (e.g., to generate a point estimate).

7.2 Relevance of Shannon information theory

Shannon's information theory holds a special status in engineering, signal processing, and other fields due to its universal implications for communication, compression, and coding in settings of perfect signal recovery. However, it is unclear whether information-theoretic quantities like entropy and mutual information are necessarily relevant to information processing and transmission in biological systems [87]. In particular, Shannon's theory (1) requires complex computation and long delays to encode and decode in ways that achieve optimality [5]; (2) ignores biological constraints (e.g., neurons cannot implement arbitrary nonlinear functions, and are inherently noisy); (3) applies to settings of perfect signal recovery, which may not be possible or even desirable in biological settings.

In the Bayesian efficient coding, the encoder is selected from a biologically relevant parametric family. And although we consider the (negative) mutual information as one possible choice of loss

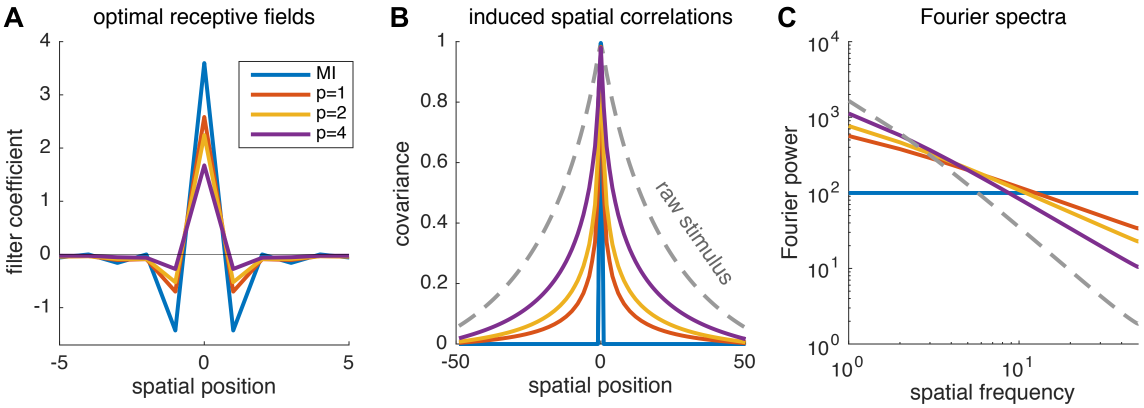

Figure 5: Optimal linear receptive fields for different loss functions. **(A)** Optimal 1D linear receptive fields for a neuron population with (infinitesimal) independent Gaussian noise and correlated Gaussian stimuli with power spectrum proportional to $1/f^2$. Plots show central 10 elements of optimal 100-element RFs optimized for mutual information (MI) and covtropy with $p = 1, 2, 4$. **(B)** Auto-correlation of the stimulus (dashed) and the 100-neuron population response generated by each encoder. The infomax filter induces perfect whitening so that the autocorrelation is a delta function (blue), while optimal covtropy encoders induce only partial whitening (even in the regime of high SNR). **(C)** Fourier power spectrum of stimulus and population responses, showing that optimal covtropy populations induce partial whitening but preserve more power at low frequencies.

function, there is no *a priori* reason for preferring it to other loss functions; as we have shown, optimizing MI it will not necessarily give good performance for other losses (e.g., Figs. 2 & 3). We therefore find no justification for claims (commonly found in the neuroscience literature) that, in the absence of knowledge about the brain's true loss function, it is somehow best to maximize mutual information.

A framework more relevant to sensory coding in the nervous system is Shannon's rate distortion theory for lossy encoding [28]. Given an allowed maximum distortion (as a measure of encoding-decoding error) L^*, the rate distortion function $R(L^*)$ describes the minimum necessary mutual information of the channel (implemented by the encoder-decoder pair) to achieve it: $R(L^*) = \inf_{L(\theta)\le L^*} I(\mathbf{x}; \hat{\mathbf{x}}(\theta))$. One can prove that the more mutual information is allowed, the smaller the achievable distortion; in fact $R(L^*)$ is a non-increasing convex function [29]. This allows an interesting overlap between the two theories: a subclass of BEC theory with mutual information as constraint can be reformulated as a special case of rate distortion theory. Consider the following dual formulations:

$$\theta_{RD} = \arg\min_{\theta} I(\mathbf{x}; \mathbf{y}|\theta) \qquad \text{subject to } \bar{L}(\theta) \le L^* \quad (22)$$

$$\theta_{BEC} = \arg\min_{\theta} \bar{L}(\theta) \qquad \text{subject to } I(\mathbf{x}; \mathbf{y}|\theta) \le R(L^*) \quad (23)$$

where $\bar{L}(\theta)$ is the loss given in eq. 5. The rate distortion solution θ_{RD} coincides with the BEC solution θ_{BEC}, because the distortion function is monotonically decreasing. Although this provides a insightful connection, such relations do not always exist, and more generally it is not clear why a bound on mutual information would be a biologically relevant constraint on a sensory encoder.

7.3 Limitations and future directions

What cost or loss function is the nervous system optimizing for? We emphasize that BEC alone cannot serve as a normative theory to answer this question; each problem and environment should dictate the loss functional and prior. Rather, BEC should be used as a guiding principle that frees the theory of efficient coding from its traditional reliance on information theoretic principles, and to cover an appropriately broad range of theories of optimal neural encoding.

BEC on its own has many degrees of freedom and is likely under-constrained by current neural and behavioral observations. Moreover, different combinations of constraints and loss functions may be consistent with a single encoding model, providing multiple "optimal" explanations for a single encoder. The priors, encoding models, constraints, and loss functions we have considered here were guided primarily by tractability as opposed to neural or biological plausibility, but some of these components (e.g., prior and noise) can be quantified with measurements [36, 88, 89]. We can start to make reasonable inferences about constraints and loss functions through careful study of the brain's resource consumption [90, 91] or the behavioral consequences of different kinds of errors [92, 93].

A variety of other issues relating to normatively optimal neural coding remain to be addressed by our theory:

Computational demands: the BEC framework does not consider constraints on the brain's computational capabilities. In particular, our theory—like with the Bayesian brain hypothesis itself—assumes that the brain can compute the desired posterior distribution over stimuli given the response, or at least extract the posterior statistics needed for optimal behavior. This is almost certainly not the case for many of the complex inference problems the brain faces. It seems more likely that the brain relies on shortcuts, heuristics, and approximate inference methods that result in sub-optimal use of the stimulus information contained in neural responses relative to a Bayesian ideal observer [94]. The BEC paradigm could therefore be extended to incorporate computational constraints: for example, linear readout of neural activity is linear, or the use of particular approximate Bayesian inference methods [95]).

The role of time: we have focused our analyses on coding problems that follow the sequential nature of Bayesian inference (stimulus → response → posterior distribution) and have ignored the continuous-time nature of stimuli, responses, and behavior. However, there is nothing about our theory that precludes application to continuous-time problems, and previous work has formulated population codes capable of updating a posterior distribution in continuous time [96–100]. In many cases optimal coding depends on the timescale over which information is integrated: for example, previous work has shown that optimal tuning curve width depends on the time over which spikes are generated, with longer time windows necessitating narrower tuning [17, 25]. Recent work has discovered limits on fast timescale transmission of information in physical and biological systems [101], but there is still a need for a theory of optimal coding in settings where there is uncertainty about the time (or times) at which the posterior distribution will be evaluated or "read out".

Latent variables: we have not so far discussed latent variable models, in which there are additional unobserved stochastic components affecting either the stimulus or the neural response. In stimulus latent variables models, the quantity of interest is a latent variable $\mathbf{z}$ that may be regarded as underlying or generating the sensory stimulus $\mathbf{x}$. (For example, $\mathbf{z}$ might the velocity of an object or the identity of a face, and $\mathbf{x}$ is the resulting image or image sequence presented to the retina [102]). In such settings, the posterior over the latent variable $\mathbf{z}$ might be the quantity of interest for an optimal code; BEC can handle this case by defining a loss function sensitive to the posterior $p(\mathbf{z}|\mathbf{y})$. This can be seen as a valid instance of BEC because this distribution can be written as a functional of the stimulus posterior: $p(\mathbf{z}|\mathbf{y}) = \int p(\mathbf{z}|\mathbf{x})p(\mathbf{x}|\mathbf{y})\,\mathrm{d}\mathbf{x}$, where $p(\mathbf{z}|\mathbf{x})$ is the posterior over the latent given the stimulus under the generative model. Latent variables also arise in models of neural activity with shared underlying variability [103–106]. In such cases, it is natural to write the encoding model itself as a joint distribution over activity and latents, $p(\mathbf{y}, \mathbf{z}|\mathbf{x})$; the BEC paradigm can once again handle this case by marginalizing over the latent to obtain $p(\mathbf{x}|\mathbf{y})$. Thus, BEC can accommodate both kinds of latent variable models typically used in neuroscience, and the consideration of stimulus-related latent variables may motivate the design of loss functions are sensitive to particular latent variables while discarding other aspects of the stimulus (e.g., as in the information bottleneck [81] or accuracy maximization analysis [75, 102]).

Although we have focused on two canonical problems that arise repeatedly in the neural coding literature, namely optimal linear receptive fields and optimal nonlinear input-output functions, we hope that future work will address other coding problems relevant to information processing in the nervous system (e.g., multilayer cascades, dynamics, correlations), and will be extended to non-Bayesian frameworks (e.g., that take into account computational costs or constraints). We believe that the BEC paradigm provides provides a rigorous theoretical framework for neuroscientists to evaluate neural systems, synthesizing the Bayesian brain and efficient coding hypothesis into a formalism that can incorporate diverse optimality criteria beyond classic information-theoretic costs.

Methods

Blowfly data from Laughlin [27]

We extracted the data points from the figures in the original paper [27] to fit the models in section 6 and for the plots in Fig. 4. The data are available as supplementary data files.

To fit the stimulus CDF and the response nonlinearity of the blowfly LMC, we fit a 3-parameter Naka-Rushton function $|(x-a)^b|/(|x-a|^b+c)$ (plotted as grey dotted line in Fig. 4B) to estimate the stimulus CDF (Fig. 4A). The parameters for the stimulus CDF were $a = -1.52, b = 5.80, c = 7.55$, and $a = -1.90, b = 5.84, c = 37.17$ for the LMC response.

To compute the optimal nonlinearities plotted in Fig. 4B, we parametrized the nonlinearity as a piecewise constant function defined on 25 bins, and numerically minimized the loss (eq. 21) using MATLAB's $\mathrm{fminunc}$ to optimize the location of the bin edges for each value of p. To compute the marginal response distributions for different nonlinearities (Fig. 4C), we transformed the quantization bin edges through the prior CDF.

Acknowledgments

We would like to thank Jacob Yates and Ozan Koyluoglu for helpful discussions. JWP was supported by grants from the McKnight Foundation, Simons Collaboration on the Global Brain (SCGB AWD1004351) and the NSF CAREER Award (IIS-1150186). IMP was partially supported by the Thomas Hartman Center for Parkinson's Research (64249).

References

[1] Attneave, F. Some informational aspects of visual perception. *Psychological review*, 61(3):183–193, 1954.

[2] Barlow, H. B. Possible principles underlying the transformation of sensory messages. In Rosenblith, W., editor, *Sensory Communication*, pages 217–234. Cambridge, MA: MIT Press, 1961.

[3] Helmholtz, H. v. *Physiological optics*, volume 3. Optical Society of America, 1925.

[4] Knill, D. & Richards, W. *Perception as Bayesian Inference*. Cambridge University Press, 1996.

[5] Barlow, H. Redundancy reduction revisited. *Network: Computation in Neural Systems*, 12(3):241–253, 2001.

[6] Geisler, W. S. & Kersten, D. Illusions, perception and bayes. *nature neuroscience*, 5(6):508–510, 2002.

[7] Kersten, D., Mamassian, P. & Yuille, A. Object perception as bayesian inference. *Annual Review of Psychology*, 55: 271–304, 2004.

[8] Linsker, R. How to generate ordered maps by maximizing the mutual information between input and output signals. *Neural computation*, 1(3):402–411, 1989.

[9] Barlow, H. & Földiák, P. Adaptation and decorrelation in the cortex. In Durbin, R., Miall, C. & Mitchison, G., editors, *The Computing Neuron*, pages 54–72. Wokingham, England: Addison-Wesley, 1989.

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

[11] Laughlin, S. B. Energy as a constraint on the coding and processing of sensory information. *Current opinion in neurobiology*, 11(4):475–480, 2001.

[12] Bialek, W., Nemenman, I. & Tishby, N. Predictability, complexity, and learning. *Neural Comput*, 13(11):2409–2463, Nov 2001.

[13] Balasubramanian, V., Kimber, D. & Berry II, M. J. Metabolically efficient information processing. *Neural Computation*, 13(4):799–815, 2001.

[14] Salinas, E. How behavioral constraints may determine optimal sensory representations. *PLoS Biol*, 4(12):e387, 11 2006.

[15] Doi, E., Balcan, D. C. & Lewicki, M. S. Robust coding over noisy overcomplete channels. *Image Processing, IEEE Transactions on*, 16(2):442–452, 2007.

[16] Doi, E. & Lewicki, M. S. Characterization of minimum error linear coding with sensory and neural noise. *Neural Computation*, 23(10):2498, 2011.

[17] Berens, P., Ecker, A. S., Gerwinn, S., Tolias, A. S. & Bethge, M. Reassessing optimal neural population codes with neurometric functions. *Proceedings of the National Academy of Sciences*, 108(11):4423, 2011.

[18] Doi, E. & Lewicki, M. S. A simple model of optimal population coding for sensory systems. *PLoS Comput Biol*, 10 (8):e1003761, 08 2014.

[19] Rao, R. P. Bayesian computation in recurrent neural circuits. *Neural computation*, 16(1):1–38, 2004.

[20] Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9:1432–1438, 2006.

[21] Jazayeri, M. & Movshon, J. A. Optimal representation of sensory information by neural populations. *Nature neuroscience*, 9:690–696, 2006.

[22] Yaeli, S. & Meir, R. Error-based analysis of optimal tuning functions explains phenomena observed in sensory neurons. *Frontiers in computational neuroscience*, 4, 2010.

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

[24] Wang, Z., Stocker, A. & Lee, D. Optimal neural tuning curves for arbitrary stimulus distributions: Discrimax, infomax and minimum l_p loss. In Bartlett, P., Pereira, F. C. N., Burges, C. J. C., Bottou, L. & Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 2177–2185, 2012.

[25] Grabska-Barwinska, A. & Pillow, J. W. Optimal prior-dependent neural population codes under shared input noise. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. & Weinberger, K., editors, *Advances in Neural Information Processing Systems 27*, pages 1880–1888. Curran Associates, Inc., 2014.

[26] Wei, X. & Stocker, A. A. A bayesian model constrained by efficient coding explains anti-bayesian percepts. *Nat Neurosci*, 2015.

[27] Laughlin, S. B. A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch*, 36:910–912, 1981.

[28] Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, 1948.

[29] Cover, T. M. & Thomas, J. A. *Elements of Information Theory*. Wiley-Interscience, August 1991. ISBN 0471062596.

[30] Brenner, N., Bialek, W. & de Ruyter van Steveninck, R. Adaptive rescaling maximizes information transmission. *Neuron*, 26(3):695–702, August 2017.

[31] Fairhall, A. L., Lewen, G. D., Bialek, W. & Ruyter Van Steveninck, R. R.de . Efficiency and ambiguity in an adaptive neural code. *Nature*, 412(6849):787–792, Aug 2001.

[32] Dan, Y., Atick, J. J. & Reid, R. C. Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *J Neurosci*, 16(10):3351–3362, May 1996.

[33] Bell, A. J. & Sejnowski, T. J. The "independent components" of natural scenes are edge filters. *Vision Res*, 37 (23):3327–3338, Dec 1997.

[34] Schwartz, O. & Simoncelli, E. P. Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8):819–825, August 2001.

[35] Pitkow, X. & Meister, M. Decorrelation and efficient coding by retinal ganglion cells. *Nature Neuroscience*, 15(4):628–635, March 2012.

[36] Geisler, W. S., Perry, J. S., Super, B. J. & Gallogly, D. P. Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, 41(6):711–724, 2001.

[37] Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.

[38] Weiss, Y., Simoncelli, E. P. & Adelson, E. Motion illusions as optimal percepts. *Nature Neuroscience*, 5:598–604, 2002.

[39] Körding, K. & Wolpert, D. Bayesian integration in sensorimotor learning. *Nature*, 427:244–247, 2004.

[40] Stocker, A. A. & Simoncelli, E. P. Noise characteristics and prior expectations in human visual speed perception. *Nature neuroscience*, 9(4):578, 2006.

[41] Faisal, A. A. & Wolpert, D. M. Near optimal combination of sensory and motor uncertainty in time during a naturalistic perception-action task. *J Neurophysiol*, 101(4):1901–1912, Apr 2009.

[42] Girshick, A. R., Landy, M. S. & Simoncelli, E. P. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nat Neurosci*, 14(7):926–932, 2011.

[43] Dayan, P., Hinton, G. E., Neal, R. M. & Zemel, R. S. The helmholtz machine. *Neural computation*, 7(5):889–904, 1995.

[44] Zemel, R. S., Dayan, P. & Pouget, A. Probabilistic interpretation of population codes. *Neural Comput*, 10(2):403–430, Feb 1998.

[45] Deneve, S., Latham, P. E. & Pouget, A. Efficient computation and cue integration with noisy population codes. *Nat Neurosci*, 4(8):826–831, Aug 2001.

[46] Príncipe, J. C. *Information Theoretic Learning*. Springer, 2010. ISBN 1441915699.

[47] Atick, J. J. & Redlich, A. N. What does the retina know about natural scenes? *Neural Computation*, 4(2):196–210, 1992.

[48] Atick, J. J. Could information theory provide an ecological theory of sensory processing? *Network: Computation in neural systems*, 3(2):213–251, 1992.

[49] Linsker, R. Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural Computation*, 4(5):691–702, 1992.

[50] Liu, Y. S., Stevens, C. F. & Sharpee, T. O. Predictable irregularities in retinal receptive fields. *Proceedings of the National Academy of Sciences*, 106(38):16499–16504, 2009.

[51] Gjorgjieva, J., Sompolinsky, H. & Meister, M. Benefits of pathway splitting in sensory coding. *The Journal of Neuroscience*, 34(36):12127–12144, 2014.

[52] Kastner, D. B., Baccus, S. A. & Sharpee, T. O. Critical and maximally informative encoding between neural populations in the retina. *Proceedings of the National Academy of Sciences*, 112(8):2533–2538, 2015.

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

[54] Wang, Z., Wei, X.-X., Stocker, A. A. & Lee, D. D. Efficient neural codes under metabolic constraints. In *Advances in Neural Information Processing Systems*, pages 4619–4627, 2016.

[55] Simoncelli, E. P. & Olshausen, B. A. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

[56] Brenner, N., Strong, S. P., Koberle, R., Bialek, W. & de Ruyter van Steveninck, R. R. Synergy in a neural code. *Neural Comput*, 12(7):1531–1552, Jul 2000.

[57] Schneidman, E., Bialek, W. & Berry, M. J. Synergy, redundancy, and independence in population codes. *J Neurosci*, 23(37):11539–11553, Dec 2003.

[58] Puchalla, J. L., Schneidman, E., Harris, R. A. & Berry, M. J. Redundancy in the population code of the retina. *Neuron*, 46(3):493–504, 2005.

[59] Schneidman, E., Puchalla, J. L., Segev, R., Harris, R. A., Bialek, W. & Berry, M. J. Synergy from silence in a combinatorial neural code. *The Journal of Neuroscience*, 31 (44):15732–15741, 2011.

[60] Latham, P. & Nirenberg, S. Synergy, redundancy, and independence in population codes, revisited. *J. Neurosci.*, 25:5195–5206, 2005.

[61] Dong, D. W. & Atick, J. J. Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network: Computation in Neural Systems*, 6 (2):159–178, 1995.

[62] Karklin, Y. & Simoncelli, E. P. Efficient coding of natural images with a population of noisy linear-nonlinear neurons. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F. & Weinberger, K., editors, *Advances in Neural Information Processing Systems (NIPS*11)*, volume 24. MIT Press, 2011.

[63] Pouget, A., Deneve, S., Ducom, J.-C. & Latham, P. E. Narrow versus wide tuning curves: What's best for a population code? *Neural Computation*, 11(1):85–90, 1999.

[64] Tkacik, G., Prentice, J. S., Balasubramanian, V. & Schneidman, E. Optimal population coding by noisy spiking neurons. *Proc Natl Acad Sci U S A*, 107(32):14419–14424, Aug 2010.

[65] Weber, F., Machens, C. K. & Borst, A. Disentangling the functional consequences of the connectivity between optic-flow processing neurons. *Nature neuroscience*, 15 (3):441–448, 2012.

[66] Gjorgjieva, J., Meister, M. & Sompolinsky, H. Optimal sensory coding by populations of on and off neurons. *bioRxiv*, pages 1–13, 2017.

[67] Wang, Z., Wei, X.-X., Stocker, A. A. & Lee, D. D. Efficient neural codes under metabolic constraints. In *Advances in Neural Information Processing Systems*, pages 4619–4627, 2016.

[68] Brinkman, B. A. W., Weber, A. I., Rieke, F. & Shea-Brown, E. How do efficient coding strategies depend on origins of noise in neural circuits? *PLOS Computational Biology*, 12 (10):e1005150+, October 2016.

[69] Wang, Z., Stocker, A. A. & Lee, D. D. Efficient neural codes that minimize lp reconstruction error. *Neural Computation*, 28(12):2656–2686, 2017/05/01 2016.

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

[71] Zhao, M.-J., Edakunni, N., Pocock, A. & Brown, G. Beyond Fano's inequality: Bounds on the optimal F-score, BER, and cost-sensitive risk and their implications. *Journal of Machine Learning Research*, 14:1033–1090, 2013.

[72] Wolf, J. & Ziv, J. Transmission of noisy information to a noisy receiver with minimum distortion. *Information Theory, IEEE Transactions on*, 16(4):406–411, 1970.

[73] Seidler, J. Bounds on the mean-square error and the quality of domain decisions based on mutual information. *Information Theory, IEEE Transactions on*, 17(6):655–665, 1971.

[74] Ruderman, D. L. Designing receptive fields for highest fidelity. *Network: Computation in neural systems*, 5(2): 147–155, 1994.

[75] Geisler, W. S., Najemnik, J. & Ing, A. D. Optimal stimulus encoders for natural tasks. *Journal of vision*, 9(13), 2009.

[76] Burge, J. & Geisler, W. S. Optimal defocus estimation in individual natural images. *Proceedings of the National Academy of Sciences*, 108(40):16849–16854, 2011.

[77] Burge, J. & Jaini, P. Accuracy maximization analysis for sensory-perceptual tasks: Computational improvements, filter robustness, and coding advantages for scaled additive noise. *PLoS comp. biol.*, 13(2):e1005281, 2017.

[78] Susemihl, A. K., Meir, R. & Opper, M. Optimal neural codes for control and estimation. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. & Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2987–2995. Curran Associates, Inc., 2014.

[79] Srinivasan, M. V., Laughlin, S. B. & Dubs, A. Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 216(1205):427–459, 1982.

[80] Hosoya, T., Baccus, S. A. & Meister, M. Dynamic predictive coding by the retina. *Nature*, 436:71–77, 2005.

[81] Tishby, N., Pereira, F. & Bialek, W. The information bottleneck method. In *Proceedings 37th Allerton Conference on Communication, Control, and Computing*, 2000.

[82] Schneidman, E., Slonim, N., Tishby, N., Steveninck, R.deRuyter van & Bialek, W. Analyzing neural codes using the information bottleneck method. *Proc. of Advances in Neural Information Processing System (NIPS-13)*, 2002.

[83] Palmer, S. E., Marre, O., Berry, M. J. & Bialek, W. Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, 112(22):6908–6913, 2015.

[84] Bulthoff, H. H. Bayesian decision theory and psychophysics. *Perception as Bayesian inference*, page 123, 1996.

[85] Körding, K. P. & Wolpert, D. M. Bayesian decision theory in sensorimotor control. *Trends in cognitive sciences*, 10 (7):319–326, 2006.

[86] Berger, J. O. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[87] Shannon, C. The bandwagon. *Information Theory, IRE Transactions on*, 2(1):3–3, 1956.

[88] Ruderman, D. & Bialek, W. Statistics of natural images: scaling in the woods. *Physics Review Letters*, 73:814–817, 1994.

[89] Geisler, W. S. Visual perception and the statistical properties of natural scenes. *Annu Rev Psychol*, 59:167–192, 2008.

[90] Lennie, P. The cost of cortical computation. *Current biology*, 13(6):493–497, 2003.

[91] Niven, J. E. & Laughlin, S. B. Energy limitation as a selective pressure on the evolution of sensory systems. *Journal of Experimental Biology*, 211(11):1792–1804, 2008.

[92] Körding, K. P. & Wolpert, D. M. The loss function of sensorimotor learning. *Proceedings of the National Academy of Sciences of the United States of America*, 101(26):9839–9842, 2004.

[93] Körding, K. P., Fukunaga, I., Howard, I. S., Ingram, J. N. & Wolpert, D. M. A neuroeconomics approach to inferring utility functions in sensorimotor control. *PLoS biology*, 2 (10):e330, 2004.

[94] Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E. & Pouget, A. Not noisy, just wrong: The role of suboptimal inference in behavioral variability. *Neuron*, 74(1):30 – 39, 2012.

[95] Grabska-Barwińska, A., Barthelmé, S., Beck, J., Mainen, Z. F., Pouget, A. & Latham, P. E. A probabilistic approach to demixing odors. *Nature neuroscience*, 20(1):98–106, 2017.

[96] Deneve, S. Bayesian inference in spiking neurons. In *Advances in neural information processing systems*, pages 353–360, 2005.

[97] Deneve, S., Duhamel, J.-R. & Pouget, A. Optimal sensorimotor integration in recurrent cortical networks: a neural implementation of kalman filters. *Journal of Neuroscience*, 27(21):5744–5756, 2007.

[98] Huys, Q. J. M., Zemel, R. S., Natarajan, R. & Dayan, P. Fast population coding. *Neural computation*, 19(2):404–441, 2007.

[99] Natarajan, R., Huys, Q. J. M., Dayan, P. & Zemel, R. S. Encoding and decoding spikes for dynamic stimuli. *Neural Computation*, 20(9):2325–2360, 2008.

[100] Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E. & Pouget, A. Probabilistic population codes for bayesian decision making. *Neuron*, 60(6):1142–1152, 2008.

[101] Lahiri, S., Sohl-Dickstein, J. & Ganguli, S. A universal tradeoff between power, precision and speed in physical communication. *arXiv preprint arXiv:1603.07758*, 2016.

[102] Burge, J. & Geisler, W. S. Optimal speed estimation in natural image movies predicts human performance. *Nat Commun*, 6, 08 2015.

[103] Macke, J. H., Buesing, L., Cunningham, J. P., Byron, M. Y., Shenoy, K. V. & Sahani, M. Empirical models of spiking in neural populations. In *Advances in neural information processing systems*, volume 24, pages 1350–1358, 2011.

[104] Vidne, M., Ahmadian, Y., Shlens, J., Pillow, J. W., Kulkarni, J., Litke, A. M., Chichilnisky, E. J., Simoncelli, E. P. & Paninski, L. Modeling the impact of common noise inputs on the network activity of retinal ganglion cells. *J. Computational Neuroscience*, 33(1):97–121, 2012.

[105] Rabinowitz, N. C., Goris, R. L., Cohen, M. & Simoncelli, E. Attention stabilizes the shared gain of v4 populations. *eLife*, 2015.

[106] Ecker, A. S., Denfield, G. H., Bethge, M. & Tolias, A. S. On the structure of neuronal population activity under fluctuations in attentional state. *The Journal of Neuroscience*, 36 (5):1775–1789, 2016.

Supplementary Information

Appendix A: Loss functionals

Loss functionals quantifies the goodness of the posterior distribution $P(\mathbf{x}\,|\,\mathbf{y},\theta)$. There are broadly two classes of loss functionals: **entropic loss** or **reconstruction loss**. Entropic losses quantify the average uncertainty of the posterior: More concentrated the posterior, the better.

$$L(P(\mathbf{x}\,|\,\mathbf{y})) = \begin{cases} \mathbb{E}_{\mathbf{x}|\mathbf{y}}\left[-\log P(\mathbf{x}|\mathbf{y})\right] & \text{(1a)} \\ \qquad\text{(Shannon entropy)} & \\ \dfrac{1}{1-\alpha}\log \mathbb{E}_{\mathbf{x}|\mathbf{y}}\left[P(\mathbf{x}|\mathbf{y})^{\alpha-1}\right] & \text{(1b)} \\ \qquad\text{(Rényi } \alpha\text{-entropy)} & \\ \dfrac{1}{1-\alpha}\mathbb{E}_{\mathbf{x}|\mathbf{y}}\left[P(\mathbf{x}|\mathbf{y})^{\alpha-1}-1\right] & \text{(1c)} \\ \qquad\text{(Tsallis entropy)} & \\ \operatorname{tr}\left(\operatorname{cov}\left[\mathbf{x}\,|\,\mathbf{y}\right]^{\frac{p}{2}}\right) & \text{(1d)} \\ \qquad(p\text{-covtropy}) & \end{cases}$$

Both eq. 1b and eq. 1c converges to eq. 1a in the limit of $\alpha \to 1$.

Traditional entropic losses do not discriminate points in the stimulus domain, that is, any error is treated equally. However, when stimulus space is $\mathbb{R}^n$, it makes sense to prefer locally tight posterior distributions. One such measure is the p-covtropy which measures the posterior concentration around the posterior mean (only defined for distributions in $\mathbb{R}^d$). Note that $\frac{p}{2}$'th power of the posterior covariance matrix corresponds to taking p'th power of the standard deviations along each principal axis through diagonalization.

The second class of loss functionals depend on a specific readout (a.k.a. estimator or reconstruction) $\hat{\mathbf{x}}(\mathbf{y})$ that maps the neural response $\mathbf{y}$ back to the stimulus domain. We consider a general average reconstruction error loss functional of the form: $L(P(\mathbf{x}\,|\,\mathbf{y})) = \mathbb{E}_{\mathbf{x}|\mathbf{y}}\left[d(\mathbf{x},\hat{\mathbf{x}}(\mathbf{y}))\right]$ where $d(\cdot,\cdot)$ is a distortion measure.

$$L(P(\mathbf{x}\,|\,\mathbf{y})) = \begin{cases} \mathbb{E}_{\mathbf{x}|\mathbf{y}}\|\mathbf{x}-\hat{\mathbf{x}}(\mathbf{y})\|_p & \text{(2a)} \\ \qquad(p\text{-norm error}) & \\ \mathbb{E}_{\mathbf{x}|\mathbf{y}}\|\mathbf{x}-\hat{\mathbf{x}}(\mathbf{y})\|_2^2 & \text{(2b)} \\ \qquad\text{(mean squared error)} & \\ \mathbb{E}_{\mathbf{x}|\mathbf{y}}\left[1-\delta(\mathbf{x}-\hat{\mathbf{x}}(\mathbf{y}))\right] & \text{(2c)} \\ \quad = P(\mathbf{x}\neq\hat{\mathbf{x}}(\mathbf{y})|\mathbf{y}) \qquad \text{(0-1 loss)} & \end{cases}$$

Mean squared error and the 0-1 loss are widely used in communication theory, machine learning, and statistics in general.

There are loose bounds that connect the posterior entropy with Bayesian error rate (e.g. Fano bound for classification) [73]. However, these bounds do not justify infomax strategy because better Bayesian error rate can be achieved with a system that directly optimizes for the target error measure.

Note the similarity between p-norm error and the p-covtropy when posterior mean is used for decoding. They coincide when $p = 2$, however for $p \neq 2$, unlike the p-norm error, p-covtropy is invariant under unitary transformation of the stimulus space (which is the key difference between the 2D Gaussian and the linear receptive field examples). This is an important distinction if axes in $\mathbf{x}$ do not have special meaning, and rotated posteriors are considered equally good. We can make this connection rigorous for a Gaussian posterior, $\mathbf{x}\,|\,\mathbf{y} \sim \mathcal{N}(\mu, C)$. Let $C = UDU^{-1}$ be the eigendecomposition of the covariance matrix. Let $Z \sim \mathcal{N}(0, D)$ be aligned on the principal axes, the p-th power of the p-norm is given by,

$$\mathbb{E}_Z\left[\sum_i |z_i|^p\right] = \sum_i \mathbb{E}_Z\left[|z_i|^p\right] = \sum_i \kappa(p)\sigma_i^p$$
$$= \kappa(p)\operatorname{tr}\left(D^{\frac{p}{2}}\right) = \kappa(p)\operatorname{tr}\left(C^{\frac{p}{2}}\right)$$

where $\kappa(p) = \frac{1}{\sqrt{\pi}}2^{\frac{p}{2}}\Gamma\left(\frac{p+1}{2}\right)$.

In section 5, we discuss the loss functions only for Gaussian posteriors. However, covtropy is well defined for any distribution with a valid covariance matrix. Note that minimizing the limiting case of covtropy for $p \to 0$ and maximizing mutual information do not coincide in general since posterior entropy is not a sole function of the covariance in general.

Appendix B: Equivalence of MSE and covtropy for $p = 2$

Here we provide a simple proof showing that minimizing covtropy for $p = 2$ corresponds to minimizing mean-squared error (MSE). Let $\mathbb{E}[\|\mathbf{x}-\hat{\mathbf{x}}(\mathbf{y})\|_2^2]$ denote the MSE for any estimator $\hat{\mathbf{x}}(\mathbf{y})$, where expectation is taken with respect to the posterior $P(\mathbf{x}|\mathbf{y})$. It is well known that the posterior mean or "Bayes least squares" estimator $\hat{\mathbf{x}}_{BLS} = \mathbb{E}[\mathbf{x}|\mathbf{y}]$ achieves the minimum of the MSE, which is then given by $\mathbb{E}[\|\mathbf{x}-\hat{\mathbf{x}}_{BLS}\|_2^2] = \mathbb{E}[(\mathbf{x}-\hat{\mathbf{x}}_{BLS})^\top(\mathbf{x}-\hat{\mathbf{x}}_{BLS})]$. We can use identities involving trace and the definition of covariance to show that MSE is equal to $\mathbb{E}[\mathrm{Tr}[(\mathbf{x}-\hat{\mathbf{x}}_{BLS})^\top(\mathbf{x}-\hat{\mathbf{x}}_{BLS})]] = \mathbb{E}[\mathrm{Tr}[(\mathbf{x}-\hat{\mathbf{x}}_{BLS})(\mathbf{x}-\hat{\mathbf{x}}_{BLS})^\top]] = \mathrm{Tr}[\mathbb{E}[(\mathbf{x}-\hat{\mathbf{x}}_{BLS})(\mathbf{x}-\hat{\mathbf{x}}_{BLS})^\top]] = \mathrm{Tr}[\Sigma]$, which is the covtropy with $p = 2$. Thus, the code that achieves minimum $p = 2$ covtropy corresponds to the code that achieves minimum MSE point estimation of the stimulus.

Appendix C: Linear receptive fields

Here we derive the optimal receptive fields for linear encoding under Gaussian noise (Sec. 5). Recall the model (eqs. 7-9):

$$\mathbf{x} \sim \mathcal{N}(0, Q); \tag{3}$$
$$\mathbf{y} = W\mathbf{x} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, R), \tag{4}$$

and we wish to find the optimal W subject to the power constraint

$$\mathbb{E}[\mathbf{y}^\top\mathbf{y}] = \mathrm{Tr}[WQW^\top + R] \leq c. \tag{5}$$

We make the assumption that Q,R, and W matrices commute, which means they share a common eigenbasis or can be diagonalized by the same orthogonal matrix. This occurs, for example, if all three are circulant matrices, so that W consists of shifted copies of a single RF shape.

C.1 Infomax

Maximizing information is equivalent to minimizing the posterior entropy:

$$\begin{aligned} L = H(\mathbf{x}|\mathbf{y},\theta) &= \tfrac{1}{2}\log|2\pi e\Sigma| \\ &= -\tfrac{1}{2}\log|W^\top R^{-1}W + Q^{-1}| + const. \end{aligned} \tag{6}$$

We can find the optimal W subject to the power constraint above using the method of Lagrange multipliers:

$$\begin{aligned} &\frac{\partial}{\partial W}(L + \lambda \mathrm{Tr}[WQW^\top + R]) \\ &\quad = -\tfrac{1}{2}(W^2R^{-1} + Q^{-1})^{-1}2WR^{-1} + 2\lambda WQ = 0, \end{aligned} \tag{7}$$

where λ is a Lagrange multiplier. This implies

$$(W^2R^{-1} + Q^{-1})RQ = \tfrac{1}{2\lambda}I, \tag{8}$$

giving

$$W^2Q + R = \tfrac{1}{2\lambda}I, \tag{9}$$

and finally

$$W = (\tfrac{1}{2\lambda}I - R)^{\frac{1}{2}}Q^{-\frac{1}{2}}. \tag{10}$$

Plugging his solution into the constraint (eq. 5) gives

$$\mathrm{Tr}[\tfrac{1}{2\lambda}I] \leq c \tag{11}$$

which implies $1/(2\lambda) = c/n$, where $n = \dim(\mathbf{y})$ is the number of neurons. Substituting for λ gives the desired expression:

$$\hat{W}_{\text{infomax}} = (\tfrac{c}{n}I - R)^{\frac{1}{2}}Q^{-\frac{1}{2}}. \tag{12}$$

C.2 Minimum p-Covtropy

We can take a similar approach for the loss function we have called p-covtropy, which is effectively the mean p'th power error in the eigenbasis of the posterior distribution. This is given by

$$L_p = \mathrm{Tr}[\Sigma^{\frac{p}{2}}], \tag{13}$$

which involves the posterior covariance matrix Σ to the matrix-power $p/2$. It turns out this loss function has the same optimum as infomax loss in the limit $p \to 0$.

Once again, the method of Lagrange multipliers allows us to solve for W explicitly in the case that W, R, and Q commute. Taking the derivative of the Lagrangian with respect to W and setting to zero yields

$$\begin{aligned} \frac{\partial}{\partial W}\left(L + \mathrm{Tr}\big[\mathrm{cov}(\mathbf{y})\big]\right) &= \frac{\partial}{\partial W}(\mathrm{Tr}[\Sigma^{\frac{p}{2}}] + \lambda\mathrm{Tr}[WQW^\top + R]) \\ &= -p(W^2R^{-1} + Q^{-1})^{-\frac{p}{2}-1}WR^{-1} + 2\lambda WQ = 0. \end{aligned}$$

This can be simplified to

$$(W^2R^{-1} + Q^{-1})^{-\frac{p+2}{2}} = \tfrac{2\lambda}{p}QR$$

and so

$$\begin{aligned} W^2 &= R\left(\tfrac{2\lambda}{p}QR\right)^{\frac{-2}{p+2}} - RQ^{-1} && (14) \\ &= \left((\tfrac{p}{2\lambda})^{\frac{2}{p+2}}(QR)^{\frac{p}{p+2}} - R\right)Q^{-1} && (15) \end{aligned}$$

and finally

$$\hat{W}_p = \left(\alpha(QR)^{\frac{p}{p+2}} - R\right)^{\frac{1}{2}}Q^{-\frac{1}{2}}, \tag{16}$$

with $\alpha = (\frac{p}{2\lambda})^{\frac{2}{p+2}}$. Substituting $\hat{W}_p$ into the power constraint (eq. 5) gives:

$$\mathrm{Tr}[\alpha(QR)^{\frac{p}{p+2}}] \leq c, \tag{17}$$

which achieves the constraint with equality when

$$\alpha = \frac{c}{\mathrm{Tr}[(QR)^{\frac{p}{p+2}}]}, \tag{18}$$

which completes the derivation.