Power-law efficient neural codes provide general link between perceptual bias and discriminability

Michael J. Morais & Jonathan W. Pillow
Princeton Neuroscience Institute & Department of Psychology
Princeton University
mjmorais, pillow@princeton.edu

Abstract

Recent work in theoretical neuroscience has shown that efficient neural codes, which allocate neural resources to maximize the mutual information between stimuli and neural responses, give rise to a lawful relationship between perceptual bias and discriminability in psychophysical measurements (Wei & Stocker 2017, [1]). Here we generalize these results to show that the same law arises under a much larger family of optimal neural codes, which we call power-law efficient codes. These codes provide a unifying framework for understanding the relationship between perceptual bias and discriminability, and how it depends on the allocation of neural resources. Specifically, we show that the same lawful relationship between bias and discriminability arises whenever Fisher information is allocated proportional to any power of the prior distribution. This family includes neural codes that are optimal for minimizing $L_p$ error for any $p$, indicating that the lawful relationship observed in human psychophysical data does not require information-theoretically optimal neural codes. Furthermore, we derive the exact constant of proportionality governing the relationship between bias and discriminability for different choices of power law exponent $q$, which includes information-theoretic ($q = 2$) as well as “discrimax” ($q = 1/2$) neural codes, and different choices of decoder. As a bonus, our framework provides new insights into “anti-Bayesian” perceptual biases, in which percepts are biased away from the center of mass of the prior. We derive an explicit formula that clarifies precisely which combinations of neural encoder and decoder can give rise to such biases.

1 Introduction

There are relatively few general laws governing perceptual inference, the two most prominent being the Weber-Fechner law [2] and Stevens’ law [3]. Recently, Wei and Stocker [1] proposed a new perceptual law governing the relationship between perceptual bias and discriminability, and showed that it holds across a wide variety of psychophysical tasks in human observers.

Perceptual bias, $b(x) = \mathbb{E}[\hat{x}|x] - x$, is the difference between the average stimulus estimate $\hat{x}$ and its true value $x$. Perceptual discriminability $D(x)$ characterizes the sensitivity with which stimuli close to $x$ can be discriminated, equivalently the just-noticeable difference (JND); this is formalized as the stimulus increment $D(x)$ such that the stimuli $x + \eta D(x)$ and $x - (1 - \eta) D(x)$ (for $\eta$ between 0 and 1) can be correctly distinguished with probability $\geq \delta$, for some value of $\delta$. Note that by this definition, lower discriminability $D(x)$ implies higher sensitivity to small changes in $x$, that is, improved ability to discriminate.

The law proposed by Wei and Stocker asserts that bias and discriminability are related according to:

$$b(x) \propto \frac{d}{dx} D(x)^2$$

(1)

where the right-hand-side is the derivative with respect to $x$ of the discriminability squared. The relationship is backed by remarkably diverse experimental support, crossing sensory modalities.

Wei and Stocker derived their law under the assumption of an information-theoretically optimal posterior when considered as a function of \( x \) which (according to Bayes’ rule) is proportional to the product of Fisher information and decoding (via its influence on the posterior).

The perceptual law proposed by Wei and Stocker can be seen to arise if perceptual judgments arise from other families of (non-information-theoretic) optimal codes? Here we provide a definitive answer to this question. We use a Bayesian observer model to generalize the Wei-Stocker law beyond information-theoretically optimal neural coding, or does the same bias-disriminability relationship require condition is necessary for the emergence of the perceptual law. Does the perceptual law arise from other families of (non-information-theoretic) optimal codes?

We first review the derivation of the Wei & Stocker result governing the relationship between bias and discriminability (Section 2). We then develop a generic variational objective for power-law efficient coding that reveals a many-to-one mapping from objective to resultant optimal code (Section 3). We use this objective to derive a nonlinear relationship between bias and discriminability that, in the limit of high signal-to-noise ratio (SNR), reproduces the Wei & Stocker result for all power-law efficient codes, with an analytic expression for the constant of proportionality (Section 4). In simulations, we explore a range of SNRs and power-law efficient codes to verify these results, and examine a variety of decoders including posterior mode, median, and mean estimators (Section 5), demonstrating the universality of the bias-discriminability relationship across a broad space of models.

2 The Wei & Stocker Law

The perceptual law proposed by Wei and Stocker can be seen to arise if perceptual judgments arise from a Bayesian ideal observer model with an appropriate allocation of neural resources. Perceptual inference in the Bayesian observer model (Fig. 1) consists of two stages: (1) encoding, in which an external stimulus \( x \) is mapped to a noisy internal representation \( y \) according to some encoding distribution \( p(y|x) \); and (2) decoding, in which the internal representation \( y \) is converted to a point estimate \( \hat{x} \) using the information available in the posterior distribution,

\[
p(x|y) \propto p(y|x)p(x),
\]

which (according to Bayes’ rule) is proportional to the product of \( p(y|x) \), known as the likelihood when considered as a function of \( x \), and a prior distribution \( p(x) \), which reflects the environmental stimulus statistics, and even task designs. At the heart of this experiment-unifying result is the Bayesian observer model, flexibly instantiating perception as Bayesian inference in an encoding and decoding cascade with a structure optimized to statistics in the natural environment [4, 5].

![Figure 1: (Left) Schematic of Bayesian observer model under power-law efficient coding.](image)

In each trial, a stimulus \( x^* \) is sampled from the prior distribution \( p(x) \), and encoded into a neural response \( y^* \) according to the encoding distribution \( p(y|x^*) \). Inference involves computing the posterior \( p(x|y^*) \propto p(y^*|x)p(x) \), and the optimal point estimate \( \hat{x} \) minimizes the expected loss \( \mathbb{E}_{p(x|y^*)}\mathcal{L}(\hat{x}, x) \). Power-law efficient coding stipulates that the encoding distribution \( p(y|x) \) has Fisher information proportional to \( p(x)^q \) for some power \( q \). Thus the prior influences both encoding (via the Fisher information) and decoding (via its influence on the posterior). (Right) Intuitive example of bias and discriminability: adjusting a crooked picture frame. The stimulus \( x \) represents the angle off of vertical. Discriminability \( D(x) \) measures the minimum adjustment needed for the observer to detect that it became better (or worse). Bias \( b(x) \) measures the offset of the estimated angle \( \hat{x} \) from the true angle, in this case the overestimation of the crookedness. Adapted with edits from [1].
stimulus statistics. Technically, the Bayes estimate is one that minimizes an expected loss under the posterior: 
\[ \hat{x}_{\text{Bayes}} = \arg \min_x \int dx p(x|y)L(x, \hat{x}), \] for some choice of loss function (e.g., \( L(x, \hat{x}) = (x - \hat{x})^2 \), which produces the “Bayes least squares estimator”).

Optimizing the encoding stage of such a model involves specifying the encoding distribution \( p(y|x) \). Intuitively, a good encoder is one that allocates neural resources such that stimuli that are common under the prior \( p(x) \) are encoded more faithfully than stimuli that are uncommon under the prior.

Recent work from several groups \([6–9]\) has shown that this allocation problem can be addressed tractably in the high-SNR regime using Fisher Information, which quantifies the local curvature of the log-likelihood at \( x \):

\[ J(x) = \mathbb{E}_{y|x} \left[ -\frac{\partial^2}{\partial x^2} \log p(y \mid x) \right]. \] (3)

In the high-SNR regime, Fisher information provides a well-known approximation to the mutual information between stimulus and response: \( I(x, y) \approx \frac{1}{2} \int dx p(x) \log J(x) + \text{const} \). This relationship arises from the fact that asymptotically, the maximum likelihood estimate \( \hat{x} \) behaves like a Gaussian random variable with variance \( \sigma^2 = 1/J(x) \) \([8, 7, 10]\). This relationship holds only in the high-SNR limit, which is also pivotal to the perceptual law. Previous work has shown that the allocation of Fisher information that maximizes mutual information between \( x \) and \( y \) is proportional to the square of the prior, such that

\[ J(x) \propto p(x)^2. \] (4)

The perceptual law of Wei & Stocker can be obtained by combining this formula with two other existing results relating Fisher information to bias and discriminability. First, Series, Stocker & Simoncelli 2009 \([11]\) showed that Fisher information placed a bound on discriminability. In the high SNR regime, this bound can be made tight resulting in the identity, \( D(x) \propto 1/\sqrt{J(x)} \), where the constant of proportionality depends on the desired threshold performance (e.g., 1 if the threshold \( \delta \approx 76\% \)). Second, the bias of a Bayesian ideal observer was shown in \([8, 1]\) to relate to the prior distribution via the relationship \( b(x) \propto \frac{d}{dx} \frac{1}{p(x)^2} \).

Combining these three proportionalities, we recover the perceptual law proposed by Wei & Stocker:

\[ b(x) \propto \frac{d}{dx} \frac{1}{p(x)^2} \propto \frac{d}{dx} \frac{1}{J(x)} \propto \frac{d}{dx} D(x)^2. \] (5)

Figure 2A illustrates the relationship between these quantities for a simulated example, highlighting its dependence on the high-SNR limit. In this paper, we will show that the condition \( J(x) \propto p(x)^2 \) is stronger than necessary, and that the same perceptual law arises from any allocation of Fisher information proportional to a power of the prior distribution, that is, \( J(x) \propto p(x)^q \) for any \( q > 0 \).
Before showing this result, we first revisit the normative setting in which such power-law allocations of Fisher information are optimal.

## 3 Power-law efficient coding

We first show from where this power-law relationship between Fisher information and prior can emerge in an efficient neural code, and what factors determine the choice of power $q$. Previous work on information-maximizing or “infomax” codes \[11, 8\] has started from the following constrained optimization problem:

$$\arg \max_{J(x)} \int dx \ p(x) \log J(x) \quad \text{subject to} \ C(x) = \int dx \ \sqrt{J(x)} \leq c,$$  \hspace{1cm} (6)

where $\log J(x)$ provides a well-known approximation to mutual information (up to an additive constant) as described above. Solving for the optimal Fisher information $J(x)$ using variational calculus and Lagrange multipliers produces (eq.4) with the equality $J(x) = c^2 p(x)^2$.

We can consider a more general method for defining normatively optimal codes by investigating Fisher information allocated according to

$$\arg \max_{J(x)} - \int dx \ p(x)J(x)^{-\alpha} \quad \text{subject to} \ C(x) = \int dx \ J(x)^{\beta} \leq c$$  \hspace{1cm} (7)

with parameters $\alpha \geq 0$ defining the coding objective and $\beta > 0$ specifying a resource constraint. Several canonical normatively optimal coding frameworks emerge from specific settings of the parameter $\alpha$, independent of the value of $\beta$:

1. In the limit $\alpha \to 0$, this is equivalent to maximizing mutual information, since $\lim_{\alpha \to 0} \frac{J(x)^{-\alpha} - 1}{-\alpha} \equiv \log J(x)$ [12].
2. If $\alpha = 1$, corresponds to minimizing the $L_2$ reconstruction error, sometimes called “discriminax” [6, 7] because it also optimizes squared discriminability.
3. For the the general case $\alpha = p/2$, for any $p > 0$, this optimization corresponds to minimizing the $L_p$ reconstruction error under the approximation $E_{x,y}\left(|\hat{x} - x|^p\right) \approx E_x (J(x)^{-p/2})$, [13].

Here we show that this third relationship arises under a more general setting. We prove a novel bound on the mean $L_p$ error of any estimator for any level of SNR (see Supplemental Materials for proof, which builds on results from \[13,14\]).

**Theorem** (Generalized Bayesian Cramer-Rao bound for $L_p$ error). For any point estimator $\hat{x}$ of $x$, the mean $L_p$ error averaged over $x \sim p(x), y|x \sim p(y|x)$, is bounded by

$$\int \int dx dy \ p(y,x) \left(\hat{x}(y) - x\right)^p \geq \int dx \ p(x) J(x)^{-p/2}$$  \hspace{1cm} (8)

for any $p > 0$, where $J(x)$ is the Fisher Information at $x$.

Thus, the objective given in (eq.7) captures a wide range of optimal neural codes via different settings of $\alpha$, including but not limited to classic efficient coding. We can solve this objective for any value of coding parameter $\alpha$ and constraint parameter $\beta > 0$ to obtain the optimal allocation of Fisher information. In all cases, the optimal Fisher information is proportional to the prior distribution raised to a power, which we therefore refer to as power-law efficient codes:

$$J_{\text{opt}}(x) = c^{1/\beta} \left( \frac{p(x)^{\gamma}}{\int dx \ p(x)^{\gamma}} \right)^{1/\beta} \triangleq k p(x)^{q},$$  \hspace{1cm} (9)

where $\gamma = \beta/(\beta + \alpha)$ and exponent $q = 1/(\beta + \alpha)$. (see Supplemental Materials for derivation).

The normalized power function of the prior in parentheses is known as the escort distribution with parameter $\gamma$ [15]. Escort distributions arise naturally in power-law generalizations of logarithmic quantities such as mutual information, and could offer a reinterpretation of efficient coding and neural
coding more generally in terms of key theories such as maximum entropy, source coding, and Fisher information in generalized geometries [16][17]. Here, we focus on the right-most expression, which characterizes a power-law efficient code in terms of the power \( q \) and constant of proportionality \( k = c^{1/\beta}(\int dx \, p(x)^\gamma)^{-1/\beta} \). One interesting feature of the power-law efficient coding framework is that the exponent \( q \), which determines how Fisher information is allocated relative to the prior, depends on both the coding parameter \( \alpha \) and the constraint parameter \( \beta \) via the relationship \( q = 1/(\beta + \alpha) \). This implies that the optimal allocation of Fisher information is multiply determined, and reveals an ambiguity between coding desideratum and constraint in any optimal code.

In the particular case of infomax coding, where \( \alpha = 0 \), we obtain \( q = 1/\beta \), meaning that the power law exponent \( q \) is determined entirely by the constraint, and the escort parameter \( \gamma \) equals 1. Previous work [7, 8, 12], therefore, could be interpreted to be implicitly or explicitly forcing the choice of \( \beta = 1/2 \). Any power-law efficient code with \( J(x) = kp(x)^q \) could be putatively “infomax” if we defined the constraint such that \( \beta = 1/q \). For example, the so-called discrimax encoder developed in [7] in which \( J(x) \propto p(x)^{1/2} \) could result from an infomax objective function (\( \alpha = 0 \)) if we only set the constraint \( \beta = 2 \). Rather than highlighting a pitfall of our procedure, this ambiguity instead highlights (i) the universality of the power-law generalization we present here, and (ii) the need to consider how other features of the observer model could further constrain the encoder to a uniquely infomax code.

4 Deriving linear and nonlinear bias-discriminability relationships

Next, we wish to go beyond proportionality and determine the precise relationship between bias and discriminability under the power-law efficient coding framework described above. However, any optimization of Fisher information, including ours, doesn’t prescribe a method for selecting a parametric encoding distribution \( p(y \mid x) \) associated with a particular power-law efficient code, that is, a distribution with Fisher information allocated according to \( J(x) = kp(x)^q \). For simplicity, we therefore consider a power-law efficient code that is parametrized as Gaussian in \( y \) with mean \( x \):

\[
p(y \mid x) = \mathcal{N}(x, \frac{1}{kp(x)^q}) = \sqrt{\frac{k p(x)^q}{2\pi}} \exp \left(-\frac{k p(x)^q}{2} (y - x)^2 \right),
\]

(10)

and we allocate Fisher information using a stimulus-dependent variance \( \sigma^2 = 1/kp(x)^q \). This is the only configuration with that allocates Fisher information appropriately and is also Gaussian in \( y \). The parametrization of this encoder differs from that used by Wei and Stocker [1][8], but critically has the same Fisher information. We can show that all key analytical results continue to hold in their parametrization when extended to power-law efficient codes, and that we ameliorate several issues in their models (see Supplemental Materials for comparisons and proofs). It also replicates the key results obtained with Wei and Stocker’s parametrization, namely repulsive “anti-Bayesian” biases, in which the average Bayes least squares estimate is biased away from prior relative to the true stimulus [8][18]. But we prefer this parametrization for its simplicity and interpretability in terms of its parameters \( k \) and \( q \).

At the decoding stage, Bayesian inference involves computing a posterior distribution over stimuli \( x \), using the encoding distribution (eq. [10]) as the likelihood:

\[
p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)} = \frac{p(x)}{p(y)} \sqrt{\frac{k p(x)^q}{2\pi}} \exp \left(-\frac{k p(x)^q}{2} (y - x)^2 \right).
\]

(11)

In the high-SNR limit, the likelihood narrows and the log-prior can be well-approximated with a quadratic about the true stimulus \( x_0 \), such that

\[
\log p(x) \approx a_0 + a_1 (x - x_0) + \frac{1}{2} a_2 (x - x_0)^2
\]

where the coefficients \( a_0 \), \( a_1 \), and \( a_2 \) are implicitly functions of \( x_0 \). For the MAP estimator \( \hat{x}_{MAP} \), the bias in response to the stimulus at \( x = x_0 \) can be expressed in this limit as (see Supplemental Materials for proof)

\[
b(x) = \frac{-\frac{(2+q)}{2q} \frac{1}{\sigma_1}}{1 - (qa_1 - \frac{a_2}{a_1}) \frac{(2+q)}{2q} \frac{1}{\sigma_1^2} \frac{d}{dx} D(x)^2} \frac{d}{dx} D(x)^2,
\]

(12)
where \( d_\delta^2 = \sqrt{2Z(\delta)} \) is the d-prime statistic for a fixed performance \( \delta \), and \( Z(\cdot) \) is the inverse normal CDF. We refer to this as our **nonlinear relationship** because it expresses bias \( b(x) \) as a nonlinear function of the squared discriminability \( D(x)^2 \). This relationship makes testable nonlinear predictions between bias and discriminability that depend on the shape of the prior at each value of \( x \) through the local prior curvature parameters \( a_1 \) and \( a_2 \).

We recover a linear relationship between bias and discriminability in the higher-SNR limit when \[ \left| \frac{d}{dx} D(x)^2 \right| \ll \left| \frac{(2 + q)}{2q^2} \left( qa_1 - \frac{a_2}{a_1} \right)^{-1} \right|, \] satisfied if the SNR \( k \gg \left| q(a_1 - \frac{a_2}{a_1}) e^{-q|\delta|} \right| \) for all \( x \). This specification of the high-SNR regime reveals that the likelihood must be so sharp around the stimulus that the prior, by comparison, becomes so broad that it is nearly flat. When satisfied, the final result is the following **linear relationship** between bias and discriminability:

\[
b(x) = \left( -\frac{(2 + q}{2q} \right) \left( -\frac{1}{d_\delta^2} \right) \frac{d}{dx} D(x)^2, \tag{13}\]

which indicates a negative constant of proportionality for all \( q \). There is no contribution of \( a_1 \) or \( a_2 \) to the coefficient of proportionality; only \( q \) matters. Thus, we confirm that for power-law efficient codes generally, the Wei-Stocke law \( b(x) \propto \frac{d}{dx} D(x)^2 \) holds in the limit of high SNR for all \( x \).

## 5 Simulating the model under different SNRs and power-law efficient codes

We used simulated data to test our derived nonlinear and linear relationships between bias and discriminability (eqs. \[12\] and \[13\]). We restricted these simulations to the high-SNR regimes in which the analytical predictions provide accurate descriptions of the simulated data, and examined the qualitative differences that emerge for different powers of the power-law efficient code. We consider a sweep of both of these parameters, \( k \) and \( q \), under different decoder loss functions, which yield different Bayesian estimators with very different implications for the resulting bias.

In all simulations, we propagate each stimulus \( x \sim p(x) \) on a finely tiled grid through a Bayesian observer model numerically, computing a posterior \( p(x|y) \propto p(x) \mathcal{N}(y; x, kp(x)^{-q}) \) for a power-law efficient code under many powers \( q \) and SNRs \( k \), and for each computed the Bayesian estimators associated with various loss functions of interest. We repeated this procedure for a large number of random smooth priors. The bias-discriminability relationship will be most clearly observed if our data can tile the space of discriminability and bias, achieved if the underlying priors are maximally random smooth priors. The bias-discriminability relationship will be most clearly observed if our simulations can tile the space of discriminability and bias, achieved if the underlying priors are maximally random smooth priors. The bias-discriminability relationship will be most clearly observed if our simulations can tile the space of discriminability and bias, achieved if the underlying priors are maximally random smooth priors.

In Figure 4, we explore how quality of predictions made by the nonlinear and linear relationships in (eqs. \[12\] and \[13\]) change as a function of the SNR \( k \) for various power-law efficient coding powers \( q \). In Figure 5, we observe how the slope of the [linear] relationship changes as a function of \( q \), to which we can compare our analytical predictions to simulated results.

### 5.1 Tests of prior-dependent nonlinear and prior-independent linear relationships

The nonlinear and linear bias-discriminability relationships together form a broad generalization of the perceptual law beyond Wei and Stocker’s prior work \[1\]. As SNR increases and the relationship converges onto a line (Figure 2D), the fluctuations along that line are captured by both relationships, but the nonlinear relationship captures some additional fluctuations orthogonal to the predicted line (Figure 5). Both nonlinear and linear relationships are exceptional approximations of the true bias
We quantify the quality of our nonlinear and linear predictions as a function of SNR by measuring the mean-squared error of the bias predictions under a model (nonlinear, linear) and the total mean-squared error, such that

\[
R = -\log \left( \frac{MSE_{\text{model}}}{MSE_{\text{null}}} \right) = -\log \left( \frac{\int dx \left( \hat{b}_{\text{model}} - b(x) \right)^2}{\int dx b(x)^2} \right)
\]  

(16)

where \( \hat{b}_{\text{model}} \), for clarity, represents a bias predicted under a given relationship (eqs. 12 or 13). We use the negative logarithm such that \( R > 0 \) imply model predictive performance better than null. This ratio is defined for each prior, which we then average over 200 random priors for all simulations.

The null model in all cases is 0 everywhere. We want each simulation’s mean-squared error to be normalized according to how much bias the underlying prior introduced – if the prior were flat, our Gaussian encoding model is unbiased and symmetric for all moments such that bias is 0 everywhere. For MAP estimation, we use our analytical nonlinear and linear relationships as the models (Figure 4A,B), further using the difference between the two \( \Delta R = -\log(MSE_{\text{nonlin}}/MSE_{\text{lin}}) \) to measure the relative performance of each model to the other (Figure 4C). We only highlight the regions where both models are making sensible predictions (\( R > 0 \)). For posterior median and mean computation, in the absence of analytical results, we use as the model a linear function regressed to the data. While by definition the estimated \( R > 0 \), the degree to which it’s positive makes it still a useful surrogate for measuring the relative linearity (Figure 4D,E).

The bias-discriminability relationship emerges from modest SNR \( k \) for any estimator (MAP, posterior median, posterior mean) and power-law efficient code with power \( q \), converging into the linear relationship as SNR \( k \) increases (Figure 4). The analytical results for the MAP estimator model the data well, as the linear and nonlinear error ratio measures cleanly cross 0 and peak (Figure 4A,B). The decrease after this peak is a numerical precision issue and isn’t informative of perceptual processing – both bias and discriminability measurements collapse into zero as \( k \) increases. The minimum SNR required for good predictions is lower for the nonlinear relationship, and this form makes better predictions than the linear relationship throughout, evidenced by the error ratio difference \( \Delta R \) being positive (Figure 4C). Moreover, the slope of the relationship as predicted from (eq. 13), as a function of \( q \), exactly matches simulations (Figure 5A).

### 5.2 Posterior median and mean estimators, anti-Bayesian repulsive biases

Analytical results for posterior median and posterior mean estimators are nontrivial, and beyond the scope of this work. However, they are likely tractable, and simulations offer interesting insight into potentially useful functional forms of an equivalent linear bias-discriminability relationship in the high-SNR limit. The posterior median could be asymptotically unbiased in \( q \) or unbiased at \( q = 2 \), as the bias tends to 0 rapidly, and the linear relationship erodes (Figure 4D, Figure 5B). The posterior
mean, on the other hand, is asymptotically unbiased for \( q = 1 \) and has repulsive biases away from the prior for \( q > 1 \) (Figure 4E), a hallmark of the Bayesian observer introduced by Wei and Stocker previously \([8]\). Although we have not developed a formal derivation, we propose the following simple relationship parametrizing the slope, after using curve-fitting to explore various functional forms:

\[
b(x) = \frac{\log(q)}{\sqrt{q} d_s} \frac{1}{d^2 D(x)} D(x)^2
\]

(17)

\( q = 1 \) is a natural transition point for these attractive-repulsive biases (see the zero-crossing in Figure 5C). Recalling (13), in this setting, the Fisher information is simply a scaling of the prior. For \( q < 1 \), low-probability events have boosted probability mass since \( p(x) < p(x)^q \). Meanwhile, for \( q > 1 \), these same events have compressed probability mass since \( p(x) > p(x)^q \). For a power-law efficient code, \( q \) is determining the weight of the tails of this likelihood. In this way, the specific infomax setting of \( q = 2 \) demonstrates repulsive biases not because it corresponds to a mutual information-maximizing encoder, but because of the tail behaviors it induces by being greater than 1.

6 Discussion

We have shown that a perceptual law governing the relationship between perceptual bias and discriminability arises under a wide range of Bayesian optimal encoding models. This extends previous work showing that the law arises from information-theoretically optimal codes \([11]\), which our work includes as a special case. Maximization of mutual information therefore does not provide a privileged explanation for the neural codes underlying human perceptual behavior, in the sense that the same lawful relationship emerges for all members of the more general family of power-law efficient codes. We have also extended the perceptual law put forth by Wei and Stocker by deriving the exact constant of proportionality between bias and derivative of squared discriminability for arbitrary choices of power-law exponent.
Simulation results
Analytical predictions
Efficient coding power
Slope at optimum of linearity index

Figure 5: Linear slope of the bias-discriminability relation as a function of the efficient coding power $q$. A. MAP estimator for analytical predictions (solid line) and simulations (dots). B. Posterior median estimator for simulations. C. Posterior mean estimator for simulations fit parsimoniously by a simple equation. Note that the slope changes sign after $q = 1$ (vertical line). Before this crossing, biases are prior-attractive ($q < 1$), and after are prior-repulsive, or “anti-Bayesian” ($q > 1$).

More generally, we have shown that power-law efficient codes arise under a general optimization program that trades off the cost of making errors against a constraint on the total Fisher information (eq. 7). Any particular allocation of Fisher information relative to the prior is therefore optimal under multiple settings of loss function and constraint, and information-theoretically optimal coding is consistent with a range of different power-law relationships between Fisher information and prior. This implies that the form of an optimal power-law efficient code depends on specifying a choice of constraint as well as a choice of loss function.

Although our work shows that Wei and Stocker’s perceptual law is equally consistent with multiple forms of optimal encoding, other recent work has suggested that information-maximization provides a better explanation of both perceptual and neural data than other loss functions [19]. One interesting direction for future work will be to determine whether other members of the power-law efficient coding family can provide equally accurate accounts of such data.

Another direction for future work will be to consider more general families of efficient neural codes. We hypothesize that, since power functions form a basis set for any function, we could show that Wei and Stocker’s law emerges whenever neural resources are allocated according to any strictly monotonic function of the prior (with positive support). Such an efficient coding principle could imply

$$J(x) \propto G(p(x)) \Rightarrow b(x) \propto \frac{d}{dx} D(x)^2$$

for strictly monotone $G : \{p(x) \mid x \in X\} \rightarrow \mathbb{R}^+$ (18)

Critically, various specialized neural circuits throughout the brain needn’t adopt the same power-law $q$, or function $G(\cdot)$. The end result is the same: biases nudge perceptual estimates towards stimuli that are more (or potentially less) discriminable (confer eq. 11; bias is a scaled step along the gradient of discriminability). Neural populations could therefore specialize computations by refining $q$ or $G(\cdot)$ to precisely privilege or discount representations of stimuli with different prior probabilities. Mutual information is one of many such specializations, and is likely sensible under some conditions, but not necessarily all. In this way, the bias-discriminability relationship could be the signature of a unifying organizational principle governing otherwise diverse neural populations that encode sensory information. It could be useful to reconceptualize “efficient codes” accordingly as a broad family of codes governed by this more general normative principle, within which an efficient code putatively allocates neural resources such that stimuli that are common under the prior are encoded more faithfully than stimuli that are uncommon under the prior. We note that this echoes our initial intuitions of a good encoder, and we’ve provided evidence to suggest that this simple condition could be sufficient.

**Acknowledgments**

We thank David Zoltowski and Nicholas Roy for helpful comments. MJM was supported by an NSF Graduate Research Fellowship; JWP was supported by grants from the McKnight Foundation, Simons Collaboration on the Global Brain (SCGB AWD1004351) and NSF CAREER Award (IIS-1150186).
References


1 Variational calculus derivation of power-law efficient coding

Optimizing (eq. 9) in the main text follows from a straightforward variational calculus optimization using Lagrange multipliers. Defining the functional as \( \mathcal{I} \), we can setup the optimization in (eq. 9) as

\[
\mathcal{I} = - \int dx \, p(x) J(x)^{-\alpha} + \lambda \left( \int dx \, J(x)^\beta - c \right)
\]  

(19)

for \( \alpha > 0 \) and \( \beta > 0 \) (recall that the specific case of infomax is \( \alpha \to 0 \)). Calculating the optimal Fisher information \( J(x) \) such that \( \nabla \mathcal{I} = 0 \), we find

\[
\partial \mathcal{I}/\partial J(x) = 0 = -\alpha p(x) J'(x) + \frac{\beta \lambda}{\alpha} J(x)^\beta - \frac{\beta}{\alpha + 1} \Rightarrow J(x) = \left( \frac{-\alpha}{\lambda \beta} p(x) \right)^{1/(\beta + \alpha)}
\]

(20)

We solve for the Lagrange multiplier \( \lambda \) using the other partial derivative:

\[
\partial \mathcal{I}/\partial \lambda = 0 \Rightarrow \int dx \, J(x)^\beta = c
\]

substituting (20) \( \Rightarrow \left( \frac{-\alpha}{\lambda \beta} \right)^{\beta/(\beta + \alpha)} \int dx \, p(x)^\beta = c \)

\[
\Rightarrow \lambda = -\frac{\alpha}{\beta} \frac{1}{c^{1/(\beta - \alpha)/\beta}} \left( \int dx \, p(x)^\beta \right)^{(\beta + \alpha)/\beta}
\]

If we define a surrogate variable \( \gamma = \beta/(\beta + \alpha) \), we can redefine

\[
\lambda = -\frac{\alpha}{\beta} \frac{1}{c^{1/\gamma}} \left( \int dx \, p(x)^\gamma \right)^{1/\gamma}
\]

(21)

\[
\Rightarrow J(x) = \left( \frac{\alpha \beta e^{1/\gamma}}{\alpha} \frac{p(x)^\gamma}{\int dx \, p(x)^\gamma} \right)^{1/(\beta + \alpha)}
\]

\[
\Rightarrow J(x) = e^{1/\beta} \left( \frac{p(x)^\gamma}{\int dx \, p(x)^\gamma} \right)^{1/\beta} \rightarrow kp(x)^q
\]

(22)

if \( k \triangleq e^{1/\beta} (\int dx \, p(x)^\gamma)^{-1/\beta} \) and \( q \triangleq 1/(\beta + \alpha) \), where we substitute exponents with some algebra for clarity, and report this result as equation (9) in the main text.

1.1 Generalized Bayesian Cramer-Rao bound for \( L_2 \) and \( L_p \) errors

The proof of our generalized Bayesian Cramer-Rao bound, a lower bound on the mean-\( L_p \) error of an estimator, follows from a hybrid of those of the van Trees inequality \[13\] and the Barakin-Vajda Cramer-Rao inequality \[20\], redeveloped for our purposes. To begin, the bias of any estimator \( \hat{x}(y) \) of a stimulus \( x \) is given by

\[
b(x) = \int dy \, p(y \mid x) (\hat{x}(y) - x)
\]
Multiplying the prior to each side and differentiating with respect to $x$, we get

$$
\frac{\partial}{\partial x} p(x) b(x) = \frac{\partial}{\partial x} \int dy \, p(y \mid x) p(x) (\hat{x}(y) - x)
$$

$$
= - \int dy \, p(y \mid x) p(x) + \int dy \, \frac{\partial}{\partial x} \left( p(y \mid x) p(x) \right) (\hat{x}(y) - x)
$$

Now integrating with respect to $x$ and substituting the joint distribution $p(y, x)$ where appropriate,

$$
p(x) b(x) \bigg|_x = - \int \int dx dy \, p(y, x) + \int \int dx dy \left( \frac{\partial}{\partial x} p(x) \right) (\hat{x}(y) - x)
$$

$$
\implies 0 = -1 + \int \int dx dy \, p(y, x) \left( \frac{\partial}{\partial x} \log p(y, x) \right) (\hat{x}(y) - x)
$$

where the zero holds if $b(x) = 0$ or $p(x) = 0$ at the endpoints of the domain of $x$ (or the two are equal at those endpoints). We can now apply Hölder’s inequality to the integral, a generalization of the Cauchy-Schwarz inequality stating that, for Hölder conjugates $p, u > 1$ such that $p^{-1} + u^{-1} = 1$, and functions $f(y, x)$ and $g(y, x)$, we have

$$
\left| \int \int dx dy \, p(y, x) f(y, x) g(y, x) \right| \leq \left( \int \int dx dy \, p(y, x) |f(y, x)|^p \right)^{1/p} \left( \int \int dx dy \, p(y, x) |g(y, x)|^u \right)^{1/u}
$$

Set $f(y, x) = (\hat{x}(y) - x)$ and $g(y, x) = \frac{\partial}{\partial x} \log p(y, x)$ such that

$$
1 \leq \left( \int \int dx dy \, p(y, x) \left| \frac{\partial}{\partial x} \log p(y, x) \right|^u \right)^{1/u} \left( \int \int dx dy \, p(y, x) \left| (\hat{x}(y) - x) \right|^p \right)^{1/p}
$$

Exponentiating everything to the $p$-th power and rearranging, we see that

$$
\int \int dx dy \, p(y, x) \left| (\hat{x}(y) - x) \right|^p \geq \left( \int \int dx dy \, p(y, x) \left| \frac{\partial}{\partial x} \log p(y, x) \right|^u \right)^{-p/u}
$$

(23)

$$
\geq \left( \int dx \, p(x) \left( \int dy \, p(y \mid x) \left| \frac{\partial}{\partial x} \log p(y, x) \right|^2 \right)^{u/2} \right)^{-p/u}
$$

$$
\geq \left( \int dx \, p(x) \left( \int dy \, p(y \mid x) \left| \frac{\partial}{\partial x} \log p(y, x) \right|^2 \right)^{p/2} \right)^{-1}
$$

$$
= \left( \int dx \, p(x) \left( - \int dy \, p(y \mid x) \left[ \frac{\partial^2}{\partial x^2} \log p(y \mid x) + \frac{\partial^2}{\partial x^2} \log p(x) \right] \right)^{p/2} \right)^{-1}
$$

after applying Jensen’s inequality twice. We note that equation (23) is a generalization of the Bayesian Cramer-Rao inequality, sometimes called the van Trees inequality after [13], to the mean-$L_p$ error of an estimator $\hat{x}$.

We can make several substitutions: the left-hand side is the mean-$L_p$ error of the estimator $\hat{x}$, and the right-hand side contains two partial derivatives – the first is the Fisher information, and the second is the curvature of the prior, a constant with respect to the efficient coding optimization problem. As Fisher information increases in the high-SNR limit, this term becomes negligible [14], and we see that

$$
\mathbb{E}_{x,y} \left( |\hat{x}(y) - x|^p \right) \geq \left( \mathbb{E}_x \left( J(x) - \frac{\partial^2}{\partial x^2} \log p(x) \right)^{p/2} \right)^{-1} \rightarrow \left( \mathbb{E}_x \left( J(x)^{p/2} \right) \right)^{-1}
$$

and

$$
- \mathbb{E}_{x,y} \left( |\hat{x}(y) - x|^p \right) \leq -\mathbb{E}_x \left( J(x)^{-p/2} \right)
$$

(24)
with equality under a chain of conditions, all satisfied when Hölder’s inequality holds with equality, occurring when, for a constant $A$,

$$\frac{\partial}{\partial x} \log p(y, x) = A(\hat{x}(y) - x) \implies \frac{\partial^2}{\partial x^2} \log p(y, x) = \frac{\partial^2}{\partial x^2} \log p(x \mid y) = -A$$

$$\implies p(x \mid y) = \exp(-Ax^2 + c_1x + c_2)$$

Equivalently, the bound is tight as the posterior becomes Gaussian, satisfied in the high-SNR limit we consider throughout. We use the negative of equation (24) to state the theorem in the main text.

2 Full derivation of nonlinear and linear bias-discriminability relations

We complete the derivation of the bias-discriminability relations in detail, in an attempt to highlight why each assumption or approximation is necessary to move the derivation forward. After some algebra pulling the functions of the prior into the exponent, we can equivalently express the posterior in (15) as

$$p(x \mid y) = \frac{1}{p(y)} \sqrt{\frac{k}{2\pi}} \exp \left( k \left[ -\frac{p(x)^q}{2} (y - x)^2 + \left( \frac{2 + q}{2q} \right) \frac{q}{k} \log p(x) \right] \right) \quad (25)$$

The maximum a posteriori estimator $\hat{x}_{MAP} \triangleq x^*$ will have a bias measurable from the argument of the exponential in the square brackets. The maximum of the log-posterior is that derivative:

$$\frac{\partial \log p(x \mid y)}{\partial x} \bigg|_{x=x^*} = 0 = -\frac{q}{2} p'(x^*) p(x^*)^q - 1 + (2 + q) \frac{q}{k} \frac{p'(x^*)^2}{p(x^*)^q}$$

(26)

and when we solve this quadratic equation for $y - x^*$,

$$y - x^* = \frac{-1}{q p'(x^*) p(x^*)^q - 1} \left( -p(x^*)^q \pm \sqrt{p(x^*)^q + (2 + q) \frac{q}{k} \frac{p'(x^*)^2}{p(x^*)^q}} \right)$$

$$= \frac{p(x^*)}{q p'(x^*)} \left( 1 \mp \sqrt{1 + (2 + q) \frac{q}{k} \frac{p'(x^*)^2}{p(x^*)^q+2}} \right)$$

and we can simplify using the binomial expansion, $(1 + x)^{\alpha} \approx 1 + \alpha x$ when $|\alpha x| \ll 1$ if we truncate terms of order $1/k^2$. Only the negative root that cancels the leading 1 is reasonable, leaving

$$y - x^* = \frac{p(x^*)}{q p'(x^*)} \left( 1 \mp \left[ 1 + \left( \frac{2 + q}{2} \right) \frac{q}{k} \frac{p'(x^*)^2}{p(x^*)^q+2} + O\left( \frac{1}{k^2} \right) \right] \right) \rightarrow -\left( \frac{2 + q}{2q} \right) \frac{q}{k} \frac{p'(x^*)}{p(x^*)} \frac{1}{p(x^*)^q}$$

when $k \gg \max_x \left( \frac{2 + q}{2q} \right) \left( \frac{q}{p(x)} \right)^2 \frac{1}{p(x)^q}$

the conditions on which can be satisfied in the high-SNR limit by setting $k$ sufficiently large as dictated here. Indeed, this inequality suggests a benchmark for measuring when the high-SNR limit begins to emerge in the model. We note that this is equivalent to truncating the quadratic term of (25) outright, but our derivation clarifies this truncation as a function of the prior and SNR.

As the SNR $k$ increases, the likelihood sharpens about the true stimulus $x_0$, such that the prior by comparison dilates and could be well-approximated locally by a log-linear or quadratic function about $x_0$. We consider the latter for generality, and find that

$$\log p(x^*) \approx a_0 + a_1(x^* - x_0) + \frac{1}{2} a_2(x^* - x_0)^2$$

and

$$p(x^*)^q = \exp(qa_0) \exp \left( q \left[ a_1(x^* - x_0) + \frac{1}{2} a_2(x^* - x_0)^2 \right] \right) \triangleq p(x_0)^q / \epsilon_{x_0}(y)$$

where we define $\epsilon_{x_0}(y) = \exp\left( -q \left[ a_1(x^* - x_0) + \frac{1}{2} a_2(x^* - x_0)^2 \right] \right)$ as the residual error in approximating $p(x^*)$ as $p(x_0)$. Since $a_1 + a_2(x^* - x_0) \equiv p'(x^*)/p(x^*)$ and $x^* - x_0$ will always
have opposite signs, we can enforce not only that $\epsilon_0(y) \in (0, 1]$, but also that $\epsilon_0(y) \to 1$ as the SNR increases in our asymptotic regime. We can solve for the bias readily, substituting these redefinitions of the prior:

$$y - x^* = (y - x_0) - (x^* - x_0) = \frac{(2 + q) \cdot qa_1}{2q} \cdot \frac{q a_1}{k p(x_0)^q} \cdot \epsilon_{x_0}(y) \left[ 1 + \frac{a_2}{a_1} (x^* - x_0) \right]$$

$$\Rightarrow b(x_0) = \mathbb{E}_{y|x_0} (x^* - x_0) = \mathbb{E}_{y|x_0} (y - x_0) + \frac{(2 + q) \cdot qa_1}{2q} \cdot \frac{q a_1}{k p(x_0)^q} \cdot \mathbb{E}_{y|x_0} \left[ \epsilon_{x_0}(y) \left[ 1 + \frac{a_2}{a_1} (x^* - x_0) \right] \right]$$

and $\mathbb{E}_{y|x_0} (y - x_0) = 0$ will follow as the first central moment of the Gaussian encoding distribution. Expanding the residual and again truncating $O(1/k^2)$ terms, we can resolve a more elaborate expression for the bias:

$$b(x_0) = \left( \frac{2 + q}{2q} \right) \cdot \frac{q a_1}{k p(x_0)^q} \left[ 1 + \left( \frac{a_2}{a_1} - qa_1 \right) b(x_0) \right]$$

We leverage prior results [11] to state the proportionality in (5) with equality, and the derivative of this relation is given by

$$D(x_0)^2 = \frac{d\sigma^2}{J(x_0)} = \frac{d\sigma^2}{k p(x_0)^q} = -\frac{\frac{d}{dx} D(x_0)^2}{\frac{d\sigma^2}{k p(x_0)^q}} = \frac{q a_1}{k p(x_0)^q}$$

We then solve for the bias:

$$b(x_0) = \frac{-\frac{(2 + q)}{2q} \cdot \frac{1}{a_2^2} \cdot \frac{d}{dx_0} D(x_0)^2}{1 - \left( \frac{q a_1}{a_2^2} \right) \cdot \frac{(2 + q)}{2q} \cdot \frac{1}{a_2^2} \cdot \frac{d}{dx_0} D(x_0)^2}$$

and report this as equation (27) in the main text.

### 3 Replication with Wei and Stocker’s alternative encoding model

However, Wei and Stocker do not build up their observer model under these noise assumptions. We frame the encoding problem in the high-SNR limit as a noise allocation problem, optimizing the Fisher information with low-noise priority for more frequently-observed stimuli. They frame the encoding problem in this limit as a stimulus remapping problem, optimizing a monotonic warping function $f(\cdot)$ to further separate adjacent stimuli in a "sensory space" with the same prior-driven prioritization. We define their observer model as having $f$-normally distributed noise since the sensory-space representation $\hat{y} = f(y)$ is normally distributed, by analogy to the log-normal distribution. Under $f$-normal noise assumptions, they posit a likelihood of the form

$$p(y | x) = \mathcal{N}(f(y); f(x), 1/k) = f'(y) \sqrt{\frac{k}{2\pi}} \exp \left( -\frac{k}{2} (f(y) - f(x))^2 \right)$$

where the $\sigma^2$ noise term in their framework is equivalent to our SNR term $1/k$, which emerges implicitly in $f$, since the Fisher information would become a simple function of $f(\cdot)$ and $k$,

$$J(x) = \frac{f'(x)^2}{\sigma^2} \to k f'(x)^2$$

Recalling our power-law efficient coding hypothesis, which stated $J(x) = k p(x)^q$, we could solve for $f(\cdot)$ and restate the constraints applied by Wei and Stocker,

$$f(x) = \frac{1}{\sqrt{k}} \int_{-\infty}^{x} dx' \sqrt{J(x')} = \int_{-\infty}^{x} dx' \sqrt{p(x')} = c$$

This constraint would not be particular to infomax efficient coding when stated in terms of the Fisher information, and would merely require that the support of $f(\cdot)$ be finite, or biologically speaking, that the neural resources allocated by the encoder are limited.
However, their $f$-normal model, as they present it, would not be a statistically viable alternative
model. Namely, the likelihood isn’t a proper distribution, and doesn’t integrate to 1 due to the finite
support of the warping function $f(\cdot)$, so restricted by their finite-resource constraint. If we integrate
the encoding distribution $p(y \mid x)$ with respect to $y$, we observe that, due to the resource constraint,
\[
\int dy \, p(y \mid x) = \Phi[\sqrt{k}(c - f(x))] - \Phi[-\sqrt{k}f(x)]
\]
where we use $\Phi[\cdot]$ to denote the cumulative standard normal distribution. The proper encoding
distribution would then take on a form like
\[
p(y \mid x) \approx \frac{2f'(y)}{\sqrt{k}[\Phi[\sqrt{k}(c - f(x))] - \Phi[-\sqrt{k}f(x)]^2} \exp\left(-\frac{k}{2} (f(y) - f(x))^2\right) \tag{31}
\]
While still semblant of a Gaussian, the posterior loses analytical tractability; moreover, the Fisher
information and therefore the warping function $f(\cdot)$ would have changed from the original forms
since we added a function of $x$, and we cannot compute either as a result. As the SNR $k \to \infty$,
the first $\Phi$ function tends to 1 while the second tends to zero, recovering their original form. But
it wouldn’t be desirable to select for an encoding model that’s only viable in the high-SNR regime,
since any comparison to that model’s behavior in a low-SNR state is ill-posed. It may be unclear
whether the emergent statistical properties are due to the transition into the high-SNR regime or the
transition into well-defined encoding distributions, among other potential pitfalls.

For these reasons, we elected to propose the Gaussian encoding model, noting it is the only encoding
distribution that could be Gaussian in $y$ satisfying the Fisher information constraints, and it remains
analytically tractable and interpretable in its parameters. That said, if we eschew concerns with the
viability of Wei and Stocker’s model, it is straightforward to replicate our core results of power-law
efficient coding and the bias-discriminability proportionality within their framework.

3.1 Generalizing biases of Wei and Stocker’s model for power-law efficient codes

Since the definition of discriminability follows only from the high-SNR limit through the Cramer-Rao
bound, and does not depend on efficient coding, they are the same under Normal and $f$-normal models.
We need only replicate the expressions they derive for the biases of Bayesian estimators under various
loss functions. We follow their proofs line-by-line as allowed, and adopt their prime-notation $(\cdot)'$
for derivatives for more immediate comparability. Note that at any stage, we could substitute $q = 2$ and
$1/k = \sigma^2$ and retrieve their results precisely. The posterior median derivation, hinging on the Taylor
expansion of $f^{-1}(\cdot)$, becomes a natural building block for all other analytical results they present. We
present this expansion equivalently in terms of the warping function $f(\cdot)$, as prematurely substituting
for the prior occludes the generality of the result. The relationship between the bias $b(x_0)$ and the
derivative of the inverse prior-squared $(1/p(x))^2_{x_0}$, as they present it, is only meaningful insofar as
that prior-squared term is equal to the Fisher information. The Fisher information is ultimately what
connects the bias to the discriminability, whether the power is 2 or a more general $q$.

$L_1$ norm-minimizing posterior median estimator We and Stocker observed that in their frame-
work, the posterior median $\tilde{x} = y$, such that the bias under this estimator becomes
\[
b(x_0) = \mathbb{E}_{y \mid x_0} (\tilde{x} - x_0) = \mathbb{E}_{y \mid x_0} (y) - x_0 \\
= \sqrt{\frac{k}{2\pi}} \int dy \, f'(y) \exp\left(-\frac{k}{2} (f(y) - f(x_0))^2\right) - x_0 \\
= \sqrt{\frac{k}{2\pi}} \int d\tilde{y} \, f^{-1}(\tilde{y}) \exp\left(-\frac{k}{2} (\tilde{y} - \tilde{x}_0)^2\right) - f^{-1}(\tilde{x}_0)
\]
Using the same second-order Taylor expansion of $f^{-1}(\cdot)$, it follows that
\[
f^{-1}(\tilde{y}) \approx f^{-1}(\tilde{x}_0) + (f^{-1}(\tilde{y}))'_{\tilde{y}=\tilde{x}_0} (\tilde{y} - \tilde{x}_0) + \frac{1}{2} (f^{-1}(\tilde{y}))''_{\tilde{y}=\tilde{x}_0} (\tilde{y} - \tilde{x}_0)^2
\]
where the derivative is with respect to \( \hat{y} \) evaluated at \( \hat{x}_0 \). Applying the expansion into the integral equation,

\[
E_{y|x_0}(y) - x_0 = \sqrt{\frac{k}{2\pi}} \int d\hat{y} \frac{1}{2} (f^{-1}(\hat{y}))''(\hat{y} - \hat{x}_0)^2 \exp \left( - \frac{k}{2} (\hat{y} - \hat{x}_0)^2 \right)
\]

\[
= \sqrt{\frac{k}{2\pi}} \int d\hat{y} \frac{1}{2} \left( \frac{1}{f'(f^{-1}(\hat{y}))} \right)' (\hat{y} - \hat{x}_0)^2 \exp \left( - \frac{k}{2} (\hat{y} - \hat{x}_0)^2 \right) - \frac{k'}{2} (\hat{y} - \hat{x}_0)^2 
\]

\[
= \frac{1}{2} \sqrt{\frac{k}{2\pi}} \int d\hat{y} \left( - \frac{f''(x_0)}{f'(x_0)^3} \right) (\hat{y} - \hat{x}_0)^2 \exp \left( - \frac{k}{2} (\hat{y} - \hat{x}_0)^2 \right) 
\]

\[
= \frac{1}{4} \left( \frac{1}{f'(x_0)^2} \right)' \sqrt{\frac{k}{2\pi}} \int d\hat{y} (\hat{y} - \hat{x}_0)^2 \exp \left( - \frac{k}{2} (\hat{y} - \hat{x}_0)^2 \right) 
\]

The integral term, with its preceding square root, evaluates the variance of the Gaussian encoding distribution in the sensory space, which is equal to \( 1/k \), leaving

\[
E_{y|x_0}(y) - x_0 = \frac{1}{4k} \left( \frac{1}{f'(x_0)^2} \right)' = \frac{1}{4k} \left( \frac{1}{p(x)^q} \right)' |_{x_0} 
\]

\[
\Rightarrow b(x_0) = E_{y|x_0}(\hat{x} - x_0) \approx \frac{1}{4k} \left( \frac{1}{p(x)^q} \right)' |_{x_0} = \frac{1}{4} \left( \frac{1}{J(x_0)} \right)' |_{x_0}
\]

(32)

As indicated above, the original result follows immediately for \( q = 2 \) and \( 1/k = \sigma^2 \).

**L_0 norm-minimizing MAP estimator** The mode of the \( f \)-normal posterior is equivalent to the mode of its log-posterior, thus

\[
\frac{\partial \log p(x | y)}{\partial x} = \frac{p'(x)}{p(x)} - k(f(y) - f(x)) f'(x)
\]

Applying the same first-order Taylor expansion of \( f(x) \) about \( y \) (sic, ultimately we replace \( f'(y) \) with \( f'(x) \), so it would have been more reasonable to perform the converse expansion), we get

\[
f(y) - f(x) \approx f'(y)(x - y)
\]

\[
\Rightarrow \frac{\partial \log p(x | y)}{\partial x} = \frac{p'(x)}{p(x)} - k(x - y)f'(y) f'(x)
\]

\[
= \frac{p'(x)}{p(x)} - k(x - y)p(y)^{q/2} p(x)^{q/2}
\]

Setting the lefthand side equal to zero, we isolate the difference between the Bayesian estimator \( \hat{x} \) and \( y \) to be

\[
\hat{x} - y = \frac{p'(\hat{x})}{p(\hat{x}) k p(y)^{q/2} p(\hat{x})^{q/2}} \approx \frac{p'(x_0)}{k p(x_0)^{q+1}} = - \frac{1}{qk} \left( \frac{1}{p(x_0)^q} \right)' |_{x_0}
\]

(33)

\[
\Rightarrow E_{y|x_0}(\hat{x} - y) = - \frac{1}{qk} \left( \frac{1}{p(x_0)^q} \right)' |_{x_0} = - \frac{1}{q} \left( \frac{1}{J(x_0)} \right)' |_{x_0}
\]

The bias under the MAP estimator becomes the composite sum of this equation with the posterior median above, such that

\[
b(x_0) = E_{y|x_0}(\hat{x} - x_0) = E_{y|x_0}(\hat{x} - y) + E_{y|x_0}(y - x_0)
\]

\[
= - \frac{1}{k} \left( \frac{1}{q} \right) \left( \frac{1}{p(x_0)^q} \right)' |_{x_0} - \left( \frac{1}{q} \right) \left( \frac{1}{J(x_0)} \right)' |_{x_0}
\]

(34)
Likewise, the original result with coefficient $-1/4$ follows immediately after resubstituting $q = 2$ and $1/k = \sigma^2$.

**Other $L_p$ Bayesian estimators** For any other $L_p$ norm-minimizing Bayesian risk minimization problem, we can recycle the Taylor expansion of $f^{-1}(\cdot)$ we developed above and repeat their derivations line-by-line to recover the same key relations wherein for some constant $A$,

$$b(x_0) = A \left( \frac{1}{J(x_0)} \right)_{x_0} \rightarrow A \left( \frac{1}{p(x_0)^q} \right)_{x_0} \rightarrow A \left( \frac{1}{p(x_0)^2} \right)_{x_0} \quad \text{when } q = 2 \quad (35)$$

Indeed, our concluding argument is that these derivations do not specially depend on infomax assumptions; they are a general property of an encoder constrained by some power function of the stimulus prior inferred from the environment, which we discuss in the following section.

### 3.2 Asymptotic equivalence of normal and $f$-normal encoding models

Although the assumptions and parameters seem different, it is straightforward to show that even this skewed $f$-normal noise model becomes a member of the same Gaussian class of encoding distributions in the high-SNR limit. Consider the first-order Taylor expansion of $f(y)$ about $f(x)$ employed when replicating of Wei and Stocker’s results, which would state that $f(y) \approx f(x) + f'(x)(y - x)$ and subsequently that $f'(y) \approx f'(x)$. Together with the definitions of Fisher information for this $f$-normal parametrization $J(x) = kf'(x)^2$,

$$p(y \mid x) = f'(x) \sqrt{\frac{k}{2\pi}} \exp \left( -\frac{kf'(x)^2}{2}(y - x)^2 \right) \rightarrow \sqrt{\frac{kp(x)^q}{2\pi}} \exp \left( -\frac{kp(x)^q}{2}(y - x)^2 \right) \quad (36)$$

where power-law efficient coding would dictate that $J(x) = kp(x)^q$. In this way, the Gaussian representation we present in our analysis of the Bayesian observer model in the high-SNR limit could supercede alternate parametrizations.