Efficient Coding and Fisher Information

Daksh Mehta

February 2026

The Homunculus (Part I)

You open a textbook on neuroscience and see a picture of the "sensory homunculus." It's a little man stretched over the surface of the brain. His hands and lips are gigantic, while his legs and torso are tiny. This map tells us that the brain allocates more neurons to processing signals from your thumb than from your kneecap.

Why? The standard hypothesis is Efficient Coding. Your brain has finite resources (neurons/energy). To maximise the information transmitted about the world, it should allocate resources based on the probability of stimuli. You touch things with your hands constantly (high probability), so you need high resolution there. You rarely touch things with your knee, so it gets low resolution.

Wei and Stocker (2015) formalised this. They suggested that the brain performs a coordinate transformation on the input space to "flatten" the probability distribution, effectively equalising the information density. Let's define a stimulus vector $x \in \mathbb{R}^n$ (e.g., the position of a touch) distributed according to a probability density $p(x)$. The brain encodes this via a function $y = F(x)$, where $y \in \mathbb{R}^n$ is the internal neural representation.

1. Let's start with a 1 dimensional case. $x \in \mathbb{R}$ is a scalar. The "local resolution" or "sensitivity" of the neural representation is determined by the slope of the mapping function, $F'(x)$. If the slope is steep, a small change in x yields a large change in y (this is formalised by discriminability).

   The "Efficient Coding Hypothesis" states that to maximise information capacity, the sensitivity should be proportional to the cumulative probability density. Show that if we require the distribution of the output y to be uniform (maximum entropy), then the slope of the function must satisfy:

   $$|F'(x)| \propto p(x)$$

   (*Hint: Use the change of variables formula for probability distributions:* $p_y(y) = p_x(x)/|F'(x)|$).

2. Now consider the homunculus, imagine you peeled off all the skin and laid it out in a 2D plane (sorry for the mental image). $x \in \mathbb{R}^2$ represents a point on the body surface. The map $F : \mathbb{R}^2 \to \mathbb{R}^2$ maps skin locations to cortical locations. Since F is most likely non-linear, we linearise it locally. Around a point x_0, $F(x) \approx F(x_0) + J(x_0)(x - x_0)$, where J is the Jacobian matrix.

   Recall that the determinant measures the factor by which a linear map expands volume. If the brain is efficiently allocating "cortical volume" to match the "probability volume" of the input, write down the relationship between $\det(J(x))$ and the prior probability $p(x)$.

3. Imagine x represents 2D gaze direction. Humans look at the horizon ($x_2 = 0$) much more than the sky or ground. Let the prior be $p(x_1, x_2) \propto e^{-x_2^2}$. Consider a linear approximation of the encoding at the horizon, represented by matrix J_{horiz}, and one looking up at the sky, J_{sky}.

   Which matrix has a larger determinant? Draw what a unit square of "gaze space" looks like after being mapped by J_{horiz} versus J_{sky}.

Solution.

1. We want the output distribution $p_y(y)$ to be uniform, i.e., equal to a constant C. Using the change of variables formula:

   $$C = \frac{p_x(x)}{|F'(x)|} \implies |F'(x)| = \frac{1}{C} p_x(x)$$

Thus, the slope (sensitivity) is proportional to the prior probability.

2. The determinant $\det(J(x))$ represents the local change in volume (how much sensory space is stretched into neural space). For efficient coding, we want to allocate cortical volume dy proportional to the probability mass of the input $p(x)dx$.

$$\det(J(x)) \propto p(x)$$

High probability regions get expanded (more neural territory); low probability regions get compressed.

3. Since the prior is higher at the horizon, $\det(J_{horiz}) > \det(J_{sky})$. So at the horizon, a unit square is stretched into a large shape with high resolution. At the sky, a unit square is squashed into a tiny shape with low resolution.

Attraction vs Repulsion (Part II)

We established that efficient codes warp space based on probability. This leads to a conflict. Standard Bayesian theory says our perception should always be biased towards the prior (attraction). However, Wei and Stocker (2015) showed that efficient coding often causes perception to be biased away from the prior (repulsion). Let's model this mathematically using Weber's Law.

Setup:

- **The Stimulus:** A magnitude variable $x > 0$ (e.g., light intensity, weight, or numerosity).
- **The Prior:** $p(x) \propto 1/x$. This is the standard prior for magnitude variables (Jeffrey's prior), reflecting that smaller values are more common in nature.
- **The Efficient Code:** From part 1, we know $F'(x) \propto p(x)$.
- **The Noise:** $\tilde{y} = y + \eta$, with $\eta \sim \mathcal{N}(0, \sigma^2)$.

1. First, let's see what happens *without* efficient coding. Imagine the neurons measured x directly with noise: $\tilde{x} = x + \eta$. We estimate $\hat{x}$ by minimising Mean Squared Error (the Mean of the Posterior).

   The posterior is $P(x|\tilde{x}) \propto P(\tilde{x}|x)p(x)$. Sketch the shape of the likelihood (Gaussian centered at $\tilde{x}$) and the prior $(1/x)$. Without doing an integral, argue why the mean of the posterior must be smaller than the observation $\tilde{x}$. *(This represents the standard Bayesian "attraction" toward smaller values, as they are more probable).*

2. Now let's bring back the efficient code to see how it warps the noise.

   First, determine the mapping $F(x)$. Since we know $F'(x) = 1/x$, integrate this to find the relationship $y = F(x)$. (You should find that the internal code is logarithmic).

   Now, consider the likelihood. We observe a noisy neural value $\tilde{y}$. Since the noise is Gaussian in the neural space, we can effectively treat the true log-stimulus as being normally distributed around our observation: $\ln(x) \sim \mathcal{N}(\tilde{y}, \sigma^2)$.

   By definition, if $\ln(x)$ is Gaussian, then x follows a Log-Normal distribution.

   The Log-Normal is an asymmetric distribution with a heavy tail. Because of this tail, its Mean (center of mass) is higher than its Mode (peak).

   (a) Write out the Likelihood function $L(x)$ and use the standard formula for the mean of a Log-Normal distribution, $\mathbb{E}[x] = e^{\mu + \frac{\sigma^2}{2}}$, to write down the mean of this likelihood in terms of $\tilde{y}$ and σ.

   (b) Let $\tilde{x} = e^{\tilde{y}}$ be the "naive" decoded stimulus. Using the approximation $e^z \approx 1 + z$ for small z, show that the mean of the likelihood is roughly:

$$\tilde{x} \cdot (1 + \frac{\sigma^2}{2})$$

Does this noise-dependent term pull the estimate towards 0 (attraction) or push it away towards infinity (repulsion)?

3. Finally, we combine the Prior and the Likelihood using Bayes' Rule to find the actual estimate. We assume the brain performs Bayesian inference in the neural space y.

   (a) Transform the prior $p(x) \propto 1/x$ into the neural space y using the change of variables rule $p(y) = p(x)/|F'(x)|$. What kind of distribution is the prior in y-space?

   (b) Given a Gaussian likelihood $P(\tilde{y}|y)$ and the prior $p(y)$ you just calculated, what is the Posterior distribution $P(y|\tilde{y})$?

   (c) We decode the estimate by taking the mean of the posterior in y-space (which is just $\tilde{y}$) and mapping it back to x using F^{-1}: $\hat{x} = e^{\tilde{y}}$. However, $\tilde{y}$ itself is noisy. Calculate the expected value of our estimate, $\mathbb{E}[\hat{x}]$, given the true stimulus x_{true}. (Recall: $\tilde{y} \sim \mathcal{N}(\ln x_{true}, \sigma^2)$ and $\mathbb{E}[e^Z] = e^{\mu+\sigma^2/2}$).

   (d) Is the final result $\mathbb{E}[\hat{x}]$ equal to x_{true} (unbiased), less than x_{true} (attraction), or greater than x_{true} (repulsion)?

4. We have seen that under "perfect" efficient coding, the prior is flattened, and a geometric skew (repulsion) emerges. However, in experiments, if you make a stimulus very blurry or noisy, the brain stops repelling and starts attracting the estimate back to the prior mean.

   Let's examine why the broadness of the likelihood (determined by σ^2) dictates which force wins the tug-of-war.

   We assumed that the efficient code "flattened" the prior. However, this flattening is only local. Contrast two regimes of noise (σ^2):

   (a) The likelihood is a narrow spike. It only "sees" the prior in a tiny local neighborhood where the slope has been flattened to nearly zero.

   (b) The likelihood is a broad plateau. It "sees" the global structure of the prior (e.g., the fact that the prior must eventually drop to zero for very large or very small stimuli).

   Using the logic of Bayes' Rule (*Posterior* $\propto$ *Likelihood* $\times$ *Prior*), explain which force (attractive or repulsive) dominates in each regime. What does this imply about the types of errors made by a "reliable" sensory system versus an "unreliable" one?

Extension. This theory has been used to explain lots of repulsive biases classically seen in human perception. The most prominent validation comes from the perception of visual orientation. Natural scene statistics show a strong prevalence of cardinal orientations (vertical and horizontal) compared to oblique angles. Consistent with Bayesian Efficient Coding predictions, humans show higher orientation discrimination sensitivity at cardinal axes (the "Oblique Effect"). Crucially, perceptual estimates of orientation show systematic repulsion away from the cardinal axes, a phenomenon that standard Bayesian models failed to explain without ad-hoc assumptions, but which arises naturally from the asymmetric likelihoods of efficient coding.

Solution.

1. The likelihood is a bell curve. The prior $1/x$ is high on the left and low on the right. Multiplying them suppresses the right side of the bell curve and lifts the left. The center of mass (Mean) shifts left. $\hat{x} < \tilde{x}$. The estimate is attracted to 0.

2. $F(x) = \ln(x)$ (Weber-Fechner Law). The likelihood is $L(x) \propto \exp(-(\tilde{y} - \ln x)^2/2\sigma^2)$. The mean of a Log-Normal variable is $e^{\tilde{y}+\sigma^2/2}$. Since $e^{\tilde{y}}$ is our decoded value $\tilde{x}$ (using $F^{-1}(x)$) , the mean is $\tilde{x} \cdot e^{\sigma^2/2}$. Since $e^{\text{positive}} > 1$ (because σ is positive), this shifts the mean to the right (away from 0).

3. (a) Change of variables: $p(y) = \frac{p(x)}{|F'(x)|} = \frac{1/x}{1/x} = 1$. The prior in the neural space is Uniform. The efficient coding transformation has "flattened" the prior!

   (b) Since the prior $p(y)$ is constant, the Posterior is proportional to the Likelihood. $P(y|\tilde{y}) \propto \mathcal{N}(\tilde{y}, \sigma^2) \cdot 1$.

(c) We estimate $\hat{x} = e^{\tilde{y}}$. To check for bias, we average over the noise. $\mathbb{E}[\hat{x}] = \mathbb{E}[e^{\tilde{y}}]$. Since $\tilde{y}$ is Gaussian with mean $\ln x_{true}$, this is the expectation of a Log-Normal variable:

$$\mathbb{E}[\hat{x}] = e^{\ln x_{true} + \sigma^2/2} = x_{true} \cdot e^{\sigma^2/2}$$

(d) Since $\sigma^2 > 0$, $e^{\sigma^2/2} > 1$. Therefore $\mathbb{E}[\hat{x}] > x_{true}$. Conclusion: The Repulsion wins. The efficient code effectively "absorbed" the prior to make the neural firing histogram flat. As a result, the attractive force of the prior vanished, leaving only the geometric skew of the non-linear mapping (repulsion).

4. In the low-noise regime, the likelihood is so narrow that the prior looks effectively flat (constant) within its window. Since a constant prior exerts no "pull," the Bayesian attractive force vanishes, leaving the geometric skew of the logarithmic mapping to dominate, resulting in repulsion.

   Conversely, in the high-noise regime, the likelihood becomes so broad that it provides almost no information (*Likelihood* $\rightarrow$ *constant*). In this limit, the Posterior becomes identical to the Prior. The brain ignores the sensory signal and defaults to its expectations, resulting in total attraction to the prior peak.

   Conclusion: Perceptual bias is a dynamic balance determined by the noise level (σ^2). A reliable sensory system (low noise) is dominated by the "repulsive" skew of its own efficient code, while an unreliable system (high noise) is dominated by "attractive" Bayesian priors.

Fisher Information (Part III)

In the previous section, we saw that efficient coding changes the variance of our perceptual estimates. But how do we quantify "variance" when we are dealing with high-dimensional vectors? We need a generalised measure of precision. In neuroscience (and statistics), this is the Fisher Information Matrix.

Intuitively, Fisher Information measures how well a population of noisy neurons can distinguish between two similar stimuli. If the neural response changes a lot when the stimulus changes slightly, information is high (it's easy to tell them apart). If the response stays the same, information is low.

The Setup: A population of N neurons encodes a k-dimensional stimulus vector $x \in \mathbb{R}^k$. We assume a linear tuning model:

$$r = Mx + \eta$$

where:

- $r \in \mathbb{R}^N$ is the vector of firing rates.
- $M \in \mathbb{R}^{N \times k}$ is the "tuning matrix" (columns are tuning curves).
- $\eta \sim \mathcal{N}(0, I_N)$ is independent Gaussian noise with variance 1.

1. You are the homunculus. You observe a specific firing pattern r. You want to guess what x caused it. The probability of observing r given a hypothetical x is the likelihood $P(r|x)$.

   Since the noise is Gaussian, r follows a distribution centered at Mx. Write down the expression for the log-likelihood function, $\mathcal{L}(x) = \log P(r|x)$, ignoring any constants that don't depend on x.

   Hint: The Gaussian PDF involves $\exp(-\frac{1}{2}\|r - \mu\|^2)$.

2. Imagine plotting $\mathcal{L}(x)$.

   - Scenario A: The plot is a sharp, narrow peak.
   - Scenario B: The plot is a flat, broad plateau.

   In which scenario are you more confident about the value of x?

   Mathematically, the "sharpness" of a peak is measured by its curvature (the second derivative). Calculate the Gradient ($\nabla_x \mathcal{L}$) and the Hessian matrix ($H = \nabla_x^2 \mathcal{L}$) of your log-likelihood with respect to x.

   Hint: Recall vector calculus: $\nabla_x (x^T A x) = 2Ax$ *if* A *is symmetric.*

3. The Fisher Information Matrix, denoted $\mathcal{I}(x)$, is defined as the negative expectation of the curvature across all possible noise patterns:

$$\mathcal{I}(x) = -\mathbb{E}_r\left[H(x)\right]$$

Substitute your Hessian from Q2 into this definition. You should find that the answer is a constant matrix that looks very familiar.

$$\mathcal{I}(x) = M^T M$$

4. Why go through all this trouble? The Cramer-Rao Bound states that for any unbiased decoder, the covariance of the estimate $\hat{x}$ is bounded by the inverse of the Fisher Information:

$$\text{Cov}(\hat{x}) \geq \mathcal{I}(x)^{-1}$$

If $\mathcal{I}(x) = M^T M$, what happens to the variance of our estimate if we double the gain of the neurons (i.e., multiply M by 2)?

5. In Q3, we assumed every neuron was independent. But real neurons share noise correlations. If Neuron A fires too much, Neuron B likely will too.

Let the noise η have a covariance matrix Σ (a symmetric, positive definite matrix). The probability density now depends on the Mahalanobis distance rather than the standard Euclidean distance:

$$P(r|x) \propto \exp\left(-\frac{1}{2}(r - Mx)^T \Sigma^{-1}(r - Mx)\right)$$

(a) Repeat the derivation from Q2/Q3 for this new likelihood. Show that the Fisher Information Matrix is now:

$$\mathcal{I}(x) = M^T \Sigma^{-1} M$$

(b) In the chapter on Least Squares, you learned that $M^T M$ appears when minimising the squared error. This new matrix $M^T \Sigma^{-1} M$ appears in Weighted Least Squares. Why does multiplying by Σ^{-1} make sense if we want to "down-weight" noisy directions? *(Hint: If Σ is large in a certain direction, what happens to Σ^{-1}?)*

6. Let's apply Singular Value Decomposition to the tuning matrix: $M = U\Sigma V^T$.

(a) Substitute this SVD into your formula $\mathcal{I} = M^T M$ and simplify it using the property that U is orthogonal ($U^T U = I$).

(b) You should end up with an expression involving only V and Σ.

(c) **Interpretation:** The singular values σ_i represent the "sensitivity" or "gain" of the population along specific directions in stimulus space (given by the columns of V). Why does U (the left singular vectors) not appear in the information matrix? *(Hint: U represents a rotation in the N-dimensional neural firing space. Does rotating the coordinate system of the neurons change how much information they carry?)*

Solution.

1. The likelihood is $P(r|x) \propto \exp\left(-\frac{1}{2}\|r - Mx\|^2\right)$. Taking the log gives:

$$\mathcal{L}(x) = -\frac{1}{2}(r - Mx)^T(r - Mx) + \text{const}$$

Expanding this: $\mathcal{L}(x) = -\frac{1}{2}(r^T r - 2x^T M^T r + x^T M^T M x)$.

2. Scenario A (Sharp Peak) implies high confidence (high curvature). To find the Hessian, we differentiate $\mathcal{L}(x)$ step-by-step.

First, expand the quadratic term inside the log-likelihood:

$$\begin{aligned}(r - Mx)^T(r - Mx) &= r^T r - r^T(Mx) - (Mx)^T r + (Mx)^T(Mx) \\ &= r^T r - 2(M^T r)^T x + x^T(M^T M)x\end{aligned}$$

Now, take the Gradient ∇_x of the log-likelihood $\mathcal{L}(x) = -\frac{1}{2}[\dots]$:

- $\nabla_x(r^T r) = 0$ (constant w.r.t x)
- $\nabla_x(-2(M^T r)^T x) = -2M^T r$
- $\nabla_x(x^T M^T M x) = 2M^T M x$

Summing these and multiplying by $-\frac{1}{2}$:

$$\nabla_x \mathcal{L} = -\frac{1}{2}(-2M^T r + 2M^T M x) = M^T r - M^T M x$$

Finally, take the derivative again to get the Hessian:

$$H = \nabla_x^2 \mathcal{L} = \nabla_x(M^T r - M^T M x) = -M^T M$$

3.

$$\mathcal{I}(x) = -\mathbb{E}[-M^T M] = M^T M$$

This is the Gram matrix of the tuning curves! It quantifies the total "signal power" of the population.

4. If $\mathcal{I} = M^T M$, then the variance is proportional to $(M^T M)^{-1}$. If we double the gain ($M \to 2M$), the Information becomes $(2M)^T(2M) = 4M^T M$. The Information quadruples. Consequently, the variance scales by $(4)^{-1} = 1/4$. Therefore, linearly scaling up neural firing rates quadratically reduces uncertainty.

5. (a) The log-likelihood is given by the Mahalanobis distance:

$$\mathcal{L}(x) = -\frac{1}{2}(r - Mx)^T \Sigma^{-1}(r - Mx)$$

To differentiate, we first expand the terms inside the parentheses. Since Σ is a covariance matrix, it (and its inverse Σ^{-1}) is symmetric. This allows us to combine the cross-terms:

$$\begin{aligned}(r - Mx)^T \Sigma^{-1}(r - Mx) &= r^T\Sigma^{-1}r - (Mx)^T\Sigma^{-1}r - r^T\Sigma^{-1}(Mx) + (Mx)^T\Sigma^{-1}(Mx) \\ &= \underbrace{r^T\Sigma^{-1}r}_{\text{constant}} - \underbrace{2x^T M^T\Sigma^{-1}r}_{\text{linear in } x} + \underbrace{x^T(M^T\Sigma^{-1}M)x}_{\text{quadratic in } x}\end{aligned}$$

Now, take the Gradient ∇_x of $\mathcal{L}(x) = -\frac{1}{2}[\dots]$:

$$\nabla_x \mathcal{L} = -\frac{1}{2}\left(-2M^T\Sigma^{-1}r + 2M^T\Sigma^{-1}Mx\right) = M^T\Sigma^{-1}(r - Mx)$$

Finally, differentiate again to get the Hessian:

$$H = \nabla_x\left[M^T\Sigma^{-1}r - (M^T\Sigma^{-1}M)x\right] = -M^T\Sigma^{-1}M$$

Therefore, $\mathcal{I}(x) = -\mathbb{E}[H] = M^T\Sigma^{-1}M$.

(b) Σ^{-1} is the "precision matrix." If variance (Σ) is high in a particular direction, precision (Σ^{-1}) is low (close to 0). This formula mathematically projects the signal M onto the "clean" dimensions where noise is low, and ignores the noisy dimensions.

6.

$$\mathcal{I} = (U\Sigma V^T)^T(U\Sigma V^T) = V\Sigma^T U^T U\Sigma V^T$$

Since U is orthogonal, $U^T U = I$.

$$\mathcal{I} = V\Sigma^2 V^T$$

The information matrix is determined solely by the singular values (Σ) and the input-space directions (V). U disappears because it represents a rotation in the high-dimensional neural space. Rotating the readout of the neurons (e.g., swapping Neuron 1 and Neuron 2, or taking linear combinations of them) does not create or destroy information about the stimulus.