Comments on the efficient coding hypothesis

Sarah Marzen

February 12, 2021

Simply stated, the efficient coding hypothesis states that the brain has adapted to represent the statistics of its environment. However, there are a number of formulations of this hypothesis, all in use, all with varying amounts of success in explaining experimentally observed receptive fields and the like. In an attempt to better understand which formulation is "best", I review some here.

In the back of my mind is Shannon's idea that information theory is delicate, that it has been formulated for very specific situations, and that haphazard applications of its ideas to biology and psychology can result in a mess.

To review, we imagine that the environment is represented by random variable X with realizations $x \in \mathcal{X}$ and fixed probability distribution $p(x)$, though one can certainly argue that $p(x)$ is highly dependent on the actions that result from computations in the brain. We imagine that the neural response in some brain region is represented by Y with realizations $y \in \mathcal{Y}$. For simplicity– and because it illustrates the main ideas– we will imagine that both X and Y are discrete random variables, with $|\mathcal{X}| = m < \infty$ and $|\mathcal{Y}| = n < \infty$. The entropy of a random variable X is given by

$$H[X] = \sum_x p(x) \log \frac{1}{p(x)}, \tag{1}$$

the conditional entropy given by

$$H[Y|X] = \sum_{x,y} p(x)p(y|x) \log \frac{1}{p(y|x)}, \tag{2}$$

and the mutual information given by

$$I[X;Y] = H[Y] - H[Y|X] = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{3}$$

For most of the formulations here, it is sufficient to view mutual information as a nonlinear correlation coefficient. This will not be true of some recent formulations using rate-distortion theory, for which there is an operational definition and a theorem behind.

In all cases, we are looking for experimentally testable predictions about $p(y|x)$ (the neural response to a stimulus) given $p(x)$ (natural scene statistics) from a normative information-theoretic principle. Unfortunately, it will turn out to be the case that many of these normative principles are underdetermined.

The first formulation we consider is that of

$$p^*(y|x) = \arg \max_{p(y|x): I[X;Y] \leq C} I[X;Y] \tag{4}$$

where C is thought of as a capacity constraint on how much information can be passed through the neurons. The interesting thing about this approach is that channel capacity is thought of as channel coding problem, in which the stimulus response is fixed and the stimuli are allowed to change. Regardless, one can easily see that

$$\max_{p(y|x): I[X;Y] \leq C} I[X;Y] \leq C, \tag{5}$$

and that therefore if there is a solution $p^*(y|x)$, it achieves $I[X;Y] = C$. Let's count the number of parameters and the number of constraints. There are mn parameters involved in specifying $p(y|x)$, but m constraints from normalization, and one additional constraint from the channel capacity maximization. One could add additional constraints, and this is often done. One would need to add $mn-m-1$ constraints on variance and the like to achieve a normative principle that is not underdetermined and that can therefore make testable predictions.

Sometimes, one specifies that the mapping must be deterministic. This turns the problem into one of combinatorial search between one of n^m possible mappings, with no guarantee that one will win. However, this approach seems to work in practice. Given that neural responses are noisy, I am not sure this is the right formulation.

The second formulation we consider is

$$p^*(y|x) = \max_{p(y|x)} H[Y], \tag{6}$$

and it is trivial to realize that $p^*(y) = \frac{1}{n}$. Sometimes additional constraints are added, turning this into a Maximum Entropy problem that is still easily solvable. The issue is that there are many mappings that result in that distribution. In practice, to make this formulation work, one must add constraints on the form of the mapping. I take the perhaps unpopular view that if the constraints are what makes the experimentally testable predictions, the normative principle might need fixing. Furthermore, there is no guarantee that relevant information about the stimulus is represented in an easily decodable way unless these constraints are properly deployed.

The third formulation to consider is perhaps the simplest:

$$p^*(y|x) = \max_{p(y|x)} I[X;Y]. \tag{7}$$

Unlike the previous formulations, constraints do not seem necessary to uniquely determine a solution, to the best of my knowledge. However, the likely result is determinism or near-determinism if $m < n$, and if one thinks that neurons are noisy– so that firing rate is not the carrier of the neural code– this may not accord well with data.

The final formulation relies upon the strength of the rate-distortion theorem. The actual statement of this theorem is a bit of a mouthful. We imagine that the environment has sent N symbols to the first brain region, which then sends one of M words to the next brain region. We define rate as $\frac{\log M}{N}$, nats per input symbol, and imagine that we would like to minimize this. Why can depend on exactly the brain region, but M is certainly related to the number of neurons, and more neurons is costly in terms of both materials and energy. The downstream brain region then decodes the word and tries to guess at what the environment has said, and there is some distortion between what was said and what was guessed, $d(x_{0:N}, \hat{x}_{0:N}) = \sum_i d(x_i, \hat{x}_i)$. The distortion measure depends greatly on the task at hand, and much work is spent finding good distortion measures. We define the rate-distortion function as

$$R(D) := \min_{\text{sensors}:\mathbb{E}[d(x_i,\hat{x}_i)]} \frac{\log M}{N}, \tag{8}$$

where the minimization is also performed over all N. Notice that we have minimized over all possible sensors. This minimization seems impossible, and it is, since it usually requires going to infinite block lengths $N \to \infty$. However, the rate-distortion theorem guarantees that for memoryless input and memoryless sensors,

$$R(D) = \min_{p(\hat{x}|x):\mathbb{E}[dx,\hat{x}]\leq D} I[X;Y]. \tag{9}$$

And so, mutual information has an operational definition. It is absolutely possible to find the $p(\hat{x}|x)$ that achieves this minimum, but it is not the sensor that the brain will be using. It is guaranteed to have the statistics of the sensor that the brain will be using.

After that mouthful, I would like to argue that early brain regions are best seen as lossy source coders, and that the rate-distortion function $R(D)$ can be used to validate their performance and also that the $p(\hat{x}|x)$ that optimizes Eq. 9 can be used to make predictions about the statistics of neural responses.