The efficient coding principle

Information theory

A bit is the most basic unit of information (I), or entropy (H). For example, given a probability distribution $P(a)$ over the allowed values of a, the amount of information needed to get the exact value of a is

$$H(a) = I(a) = \sum_{a} P(a) \{-\ log_2 P(a)\} \text{ bits} = -\sum_{a} P(a)\ log_2 P(a) \text{ bits}$$

Entropy H is a measure of uncertainty, or the amount of information missing before one knows the exact value of the variable. It is the highest when all possibilities are equally likely (the variable is uniformly distributed).

When a signal is transmitted through a noisy channel, the transmission process can be described by 3 components: (1) signal S, (2) noise N, and (3) output $O = S + N$. In a noisy channel, the output is unlikely to be equal to the signal itself. Rather, by observing the output, we learn additional information about the signal. This information is called the mutual information:

$$I(O, S) = I(S, O) = H(S) + H(O) - H(O, S) = H(S) - H(S|O) = H(O) - H(O|S)$$

In a Gaussian channel (S, N, and O are Gaussian random variables), larger standard deviation leads to higher entropy. Moreover, the mutual information between S and O depends on the signal-to-noise ratio σ_s^2/σ_n^2 :

$$I(O, S) = log_2 \frac{\sigma_O}{\sigma_n} = \frac{1}{2} log_2 (1 + \frac{\sigma_s^2}{\sigma_n^2})$$

A non-zero mutual information means some information is shared between the variables. Thus, they are redundant:

$$H(O) + H(S) \geq H(O, S)$$

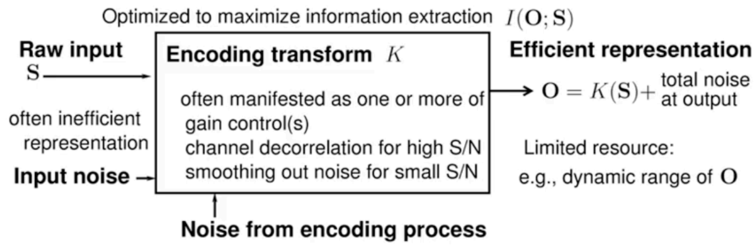
$$\text{Redundancy} = \frac{H(O)+H(S)}{H(O, S)} - 1 \geq 0$$

Redundant information requires more coding resources than non-redundant information. However, redundant information is not inherently bad, as it helps in error correction.

Formulation of the efficient coding principle

The primary bottleneck in the visual system is the optic nerve. The efficient coding principle proposes a hypothesis on how the information from the retina may be condensed and transmitted to V1 by maintaining as much visual information as possible while minimizing the neural connections. Mathematically, the goal of the efficient coding principle is to find the optimal K to minimize the neural cost required to transmit the information:

$$E(K) \equiv \text{neural cost} - \lambda I(O, S)$$



When the input noise is negligible, the transformation decorrelates the input channels to reduce information redundancy. When the noise is high, however, the input redundancy is maintained for error correction and the noise is smoothed.

Application on input sampling in contrast, color, and space

A few examples of phenomena explained by the efficient coding principle:

1. Contrast sampling in a fly's compound eye: the highest sensitivity to most probable inputs
2. Spatial sampling by photoreceptor distribution on the retina: cone density matches the distribution density of the objects
3. Optimal color sampling by the cones on the light spectrum

Formulation and solution of efficient coding by linear neural receptive fields

The transform K describes the properties of the neuron's receptive field, e.g., spatial or temporal receptive field, color tuning, ocular dominance properties, etc. The optimal encoding transform K that minimizes $E(K)$ is found by solving $\delta E / \delta K = 0$. The solution factors into:

$$K = UgK_0$$

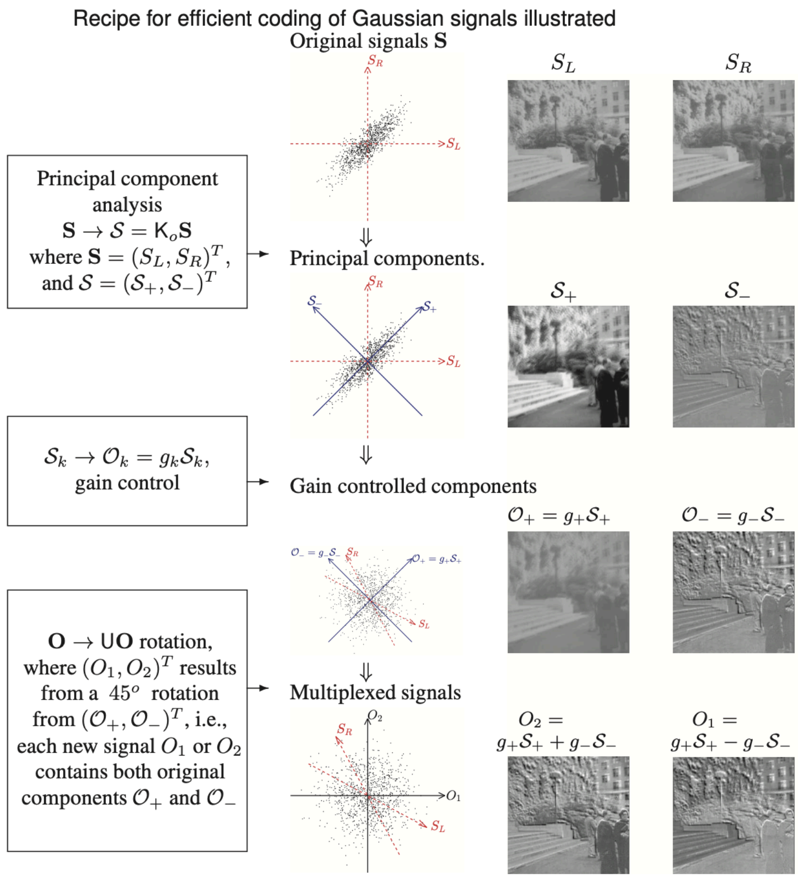
where K_0 – principal component decomposition, g – gain control according to signal-to-noise ratio (SNR) of each component, and U – multiplexing. In high SNR conditions, the gain amplifies the weaker signal to achieve decorrelation and redundancy reduction; in low SNR, the gain drops to smooth out the noise.

Case study: efficient coding for stereo vision

A neuron in V1 receives a weighted sum of left and right eye inputs. The preference towards one eye over the other can be observed in the ocular dominance columns.

Let's examine how the efficient coding principle can be applied for efficient coding in stereo vision:

1) The principle component decomposition transforms the inputs from both eyes into decorrelated signals of (1) ocular summation and (2) ocular opponency.
2) The gain control reshapes the probability distributions of the decorrelated signals (1) and (2) so that their variances are equal.
    a) When SNR is high, ocular contrast is enhanced (whitening).
    b) When SNR is low, ocular contrast is abandoned and the summation channel is enhanced (smoothing).
3) A special case of multiplexing, $U = K_0^{-1}$, is applied. This matrix gives the lowest distortion between the inputs and outputs, and minimizes required neural wiring.



The ocularity of V1 neurons depends on the input statistics, for example:

- In bright conditions, more neurons will be binocular than in a dim environment
- In animals with long interocular distances, there will be fewer binocular cells
- More binocular cells are tuned to horizontal orientation than to vertical orientation

Efficient coding by the receptive fields of the retinal ganglion cells

The efficient coding principle can be applied to 3 feature dimensions, space, time, and color:

(1) **Retinal spatial coding**. Input is a vector with as many components as there are spatial locations.

1. Decorrelation – Fourier transform of the spatial image.
2. Gain control gives a specific weight to each Fourier component determined by its signal power.
3. Multiplexing – inverse Fourier transform.

The shape of the receptive field has the center-surround receptive field, and is the same across all neurons with the center of the receptive field translated between the neurons. When the noise is high, the receptive field size is larger and the correlations between neighboring ganglion cells will be observed more easily.

(2) **Retinal temporal coding**. Input is a signal in time. The efficient coding principle algorithm is analogous to spatial coding.

(3) **Retinal color coding**. The algorithm is similar to the ocular coding. Performing it with three color dimensions results in 3 decorrelated channels: luminance, yellow-blue and red-green.

Coding in V1, the primary visual cortex

To apply the efficient coding principle to V1 neurons, two additional goals must be fulfilled: (1) translational invariance, and (2) scale invariance. This can be achieved by using a different K transform. However, the representation in V1 is overcomplete and redundant, which goes against the efficient coding principle as the redundant representation requires high neuronal cost.

V1 neurons can be tuned to many features, such as orientation, color, spatial scale or frequency, disparity, eye of origin, motion direction, temporal frequencies, etc. However, no single cell is tuned to all the feature dimensions at once. Considering the contrast between features requires high enough signal power. This can be achieved by integrating along some feature dimensions, while taking the contrast of other feature dimensions. That is the reason why a very low number of V1 neurons are tuned to more than two features at once.

Other formulations, and unsupervised learning of the efficient codes

Alternative formulations of the efficient coding principle have been proposed, such as sparse coding, maximum entropy coding, and unsupervised learning approaches. However, these alternatives do not explain the overrepresentation in V1 and lack predictive power. Despite its limitations and because of its predictive power, the efficient coding principle remains the dominating theory in the field.