Information Theory

Quan Wen

17 Dec 2019

Shannon Information

Shannon information can be formulated intuitively by playing the game "sixty-three". If you have a number in your hand, and this integer is randomly picked from 0 and 63 with equal probability. What's the smallest number of yes/no questions needed to identify an integer x between 0 and 63?

Intuitively, Six questions suffice. One reasonable strategy ask the following questions:

- $x \geq 32$?
- is x mod $32 \geq 16$?
- is x mod $16 \geq 8$?
- is x mod $8 \geq 4$?
- is x mod $4 \geq 2$?
- is x mod $2 \geq 1$?

What are the Shannon information contents of the outcomes in this example? If we assume that all values of x are equally likely, then the answers to the questions are independent and each has Shannon information content $\log_2(1/0.5) = 1$bit; the total Shannon information gained is always $\log_2 64 = 6$ bits. Furthermore, the number x that we learn from these questions is a six-bit binary number. Our questioning strategy defines a way of encoding the random variable x as a binary file.

Now let us put everything in the context of sensory encoding. Given a sensory stimulus, such as flash light with given intensity, neurons in the retina, such as biopolar cells, can faithfully respond to the intensity change. Let us denote the response amplitude r, which could be membrane potential, or the firing rate of the neuron. Just by measuring the response of the neuron, how much information we could learn about the stimulus. The answer is just given by

$$h = -\log_2 P(r)$$

where $P(r)$ is the probability of a given response amplitude.

When we write down the above formula, we are assuming that all responses have equal probability. When responses have unequal probabilities, we modify the above formula, by computing the average over all different responses. This leads to the definition of Shannon information

$$I = -\sum_r P(r) \log_2 P(r) \tag{1}$$

When we treat the response of a neuron as a continuous variable, one should then define the probability density function $p(r)$, and the Shannon information becomes

$$I = -\sum_r p(r)\Delta r \log_2 p(r)\Delta r \tag{2}$$

Furthermore,

$$\lim_{\Delta r \to 0}(I + \log_2 \Delta r) = -\int dr p(r) \log_2 p(r) \tag{3}$$

We now ask, what is the input-output relationship of a neuron, $r = f(s)$, that would maximize the information (or entropy) of the response? Let's consider the simple case, by assuming that there is a maximum response r_{max} a neuron could achieve. This could be the maximum firing rate. And the minimum response is zero. It is now your homework to show that under this condition, different responses with equal probability would maximize the entropy. Now use the normalization condition

$$\int_0^{r_{max}} p(r)dr = 1,$$

we found that $p(r) = \frac{1}{r_{max}}$. Given the probability density distribution of inputs $p(s)$, One must have $p(s)ds = p(r)dr$. As a result, $\frac{dr}{ds} = r_{max}p(s)$, and the input-output function should be proportional to the cumulative probability distribution of stimulus

$$f(s) = \int_0^s r_{max}p(s')ds' \tag{4}$$

Mutual Information

As we discussed in previous lectures, the entropy of the neural response is given by

$$H_r = -\sum_r P(r) \log_2 P(r) \tag{5}$$

In early sensory processing, what we are really interested is how much useful information has been transmitted from the input to the output. Are all the bits

in H_r useful? Not really, some of the bits may reflect noise. Given the stimulus value s, the response can be variable, as defined by the conditional probability $P(r|s)$. Thus the noise entropy is given by

$$H_{noise} = -\sum_r P(r|s) \log_2 P(r|s)$$

When we averaged this over different stimulus, one has

$$\langle H_{noise} \rangle = -\sum_s \sum_r P(s)P(r|s) \log_2 P(r|s) \tag{6}$$

The useful information is the subtraction of noise entropy from the response entropy, called mutual information

$$\begin{aligned} I &= H_r - H_{noise} \\ &= -\sum_r P(r) \log_2 P(r) + \sum_s \sum_r P(s)P(r|s) \log_2 P(r|s) \\ &= -\sum_r \sum_s P(r|s)P(s) \log_2 P(r) + \sum_s \sum_r P(s)P(r|s) \log_2 P(r|s) \\ &= \sum_r \sum_s P(r|s)P(s) \log_2 \frac{P(r|s)}{P(r)} \end{aligned} \tag{7}$$

Recall that joint probability

$$P(r, s) = P(r|s)P(s),$$

the mutual information can then be expressed in a symmetric form between stimulus and response,

$$I = \sum_r \sum_s P(r, s) \log_2 \frac{P(r, s)}{P(r)P(s)} \tag{8}$$

Mutual information also provides a distance measure between two distributions, called K-L divergence, formally defined as

$$D_{KL}(P, Q) = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)} \tag{9}$$

It is your homework to show that $D_{KL} \geq 0$ and $D_{KL} = 0$ only when $P = Q$.

Mutual information of Gaussian variable

Consider a gaussian random variable x drawn from the distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x - \langle x \rangle)^2}{2\sigma^2}\right] \tag{10}$$

Let us compute its entropy

$$S = -\int_{-\infty}^{\infty} p(x) \log_2 p(x) \tag{11}$$

To compute it explicitly, we have

$$S = \frac{1}{\ln 2}\left[\ln\sqrt{2\pi\sigma^2} + \int p(x)\frac{(x- < x >)^2}{2\sigma^2}\right] \tag{12}$$

$$= \frac{1}{\ln 2}\left[\frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2}\right] \tag{13}$$

$$= \frac{1}{\ln 2}\frac{1}{2}\left[\ln(2\pi\sigma^2) + 1\right] \tag{14}$$

$$= \frac{1}{\ln 2}\frac{1}{2}\ln(2\pi e\sigma^2) \tag{15}$$

$$= \frac{1}{2}\log_2(2\pi e\sigma^2) \tag{16}$$

Now let us consider a very simple linear model

$$y = wx + \xi \tag{17}$$

Here ξ is noise with mean $\langle x \rangle$ zero and variance σ_ξ^2.

$$I(x,y) = S(y) - \langle S(y|x)\rangle_x \tag{18}$$

$$= S(y) - \langle S(wx+\xi|x)\rangle_x \tag{19}$$

$$= S(y) - \langle S(\xi|x)\rangle_x \tag{20}$$

$$= S(y) - S(\xi) \tag{21}$$

$$= \frac{1}{2}\log_2\frac{\sigma_y^2}{\sigma_\xi^2} \tag{22}$$

$$= \log_2\frac{\sigma_y}{\sigma_\xi} \tag{23}$$

In addition, we know that

$$\sigma_y^2 = w^2\sigma_x^2 + \sigma_\xi^2 \tag{24}$$

Therefore,

$$I(x,y) = \frac{1}{2}\log_2\left(1 + \frac{w^2\sigma_x^2}{\sigma_\xi^2}\right) \tag{25}$$

Since we can rewrite

$$y = w(x + \xi_{eff}) \tag{26}$$

where $\xi_{eff} = \xi/w$. We thus have $\langle \xi_{eff}^2 \rangle = \sigma_\xi^2/w^2$. Define the signal to noise ratio as

$$\mathbf{SNR} = \frac{\langle x^2 \rangle}{\langle \xi_{eff}^2 \rangle} \tag{27}$$

The mutual information

$$I(x,y) = \frac{1}{2}\log_2(1+\mathbf{SNR}) \tag{28}$$

We can also generalize the above case to multivariate Gaussian distribution. Consider the model

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \xi \tag{29}$$

Let us define the covariance matrices for the output and the noise as

$$\mathbf{C_Y} = \langle \mathbf{y}\mathbf{y}^T \rangle \tag{30}$$

$$\mathbf{C}_\xi = \langle \xi\xi^T \rangle \tag{31}$$

The mutual information is now given by

$$I(x,y) = \frac{1}{2}\log_2 \frac{|\mathbf{C_Y}|}{|\mathbf{C}_\xi|}, \tag{32}$$

where $|\mathbf{C}|$ is the determinant of the covariance matrix.

Sensory encoding in multiple neurons

With all the above preparations, we could now consider a very interesting feed-forward model. The output response in neuron y_i is given by

$$y_i = \sum_j W_{ij}(x_j + z_j) + \eta_i \tag{33}$$

Here x may be viewed as the voltage signal from the photoreceptor, and we are measuring the response in the biopolar cell in the retina. $\mathbf{x} = [x_1, x_2, ..., x_N]^T$ with zero means drawn from a multivariate gaussian distribution

$$p(\mathbf{x}) = \frac{1}{Z}\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{C^{-1}}\mathbf{x}\right) \tag{34}$$

where Z is the normalization factor. $\mathbf{C}$ is the covariance matrix of $\mathbf{x}$, satisfying $\mathbf{C} = \langle \mathbf{x}\mathbf{x^T} \rangle$. η_i is noise drawn also from a gaussian distribution, whose covariance matrix is given by $\langle \eta_i \eta_j \rangle = \sigma_\eta^2 \delta_{ij}$. z_j is the input noise, and $\langle z_i z_j \rangle = \sigma_z^2 \delta_{ij}$.

By defining a new effective noise

$$\xi_i = \sum_j W_{ij} z_j + \eta_i, \tag{35}$$

we can now use the formula above, which gives us

$$I = \frac{1}{2}\log_2 \frac{|\mathbf{C_Y}|}{|\mathbf{C}_\xi|}. \tag{36}$$

Maximizing the mutual information may be achieved by boosting the signal. However, there is an energy cost. Here let us set a constraint on the total output signal, that is

$$\sum_i \langle y_i^2 \rangle = \mathrm{Tr} \langle \mathbf{y}\mathbf{y}^T \rangle \tag{37}$$

Thus our job is to maximize the following objective function

$$F = \frac{1}{2} \log_2 \frac{|\mathbf{C_Y}|}{|\mathbf{C}_\xi|} - \lambda \mathrm{Tr} \langle \mathbf{y}\mathbf{y}^T \rangle \tag{38}$$

Writing it down explicitly, we have

$$\mathbf{C_Y} = \langle (\mathbf{W}(\mathbf{x}+\mathbf{z}) + \eta)(\mathbf{W}(\mathbf{x}+\mathbf{z}) + \eta)^T \rangle \tag{39}$$

$$= \langle \mathbf{W}\mathbf{x}\mathbf{x}^\mathbf{T}\mathbf{W}^\mathbf{T} \rangle + \langle \mathbf{W}\mathbf{z}\mathbf{z}^\mathbf{T}\mathbf{W}^\mathbf{T} \rangle + \langle \eta\eta^\mathbf{T} \rangle \tag{40}$$

Likewise,

$$\mathbf{C}_\xi = \langle \mathbf{W}\mathbf{z}\mathbf{z}^\mathbf{T}\mathbf{W}^\mathbf{T} \rangle + \langle \eta\eta^\mathbf{T} \rangle \tag{41}$$

Now let us consider choosing an orthogonal matrix $\mathbf{U}\mathbf{U}^\mathbf{T} = \mathbf{U}^\mathbf{T}\mathbf{U} = \mathbf{I}$ that can diagonalize the input and output covariance matrix

$$\mathbf{U} \langle \mathbf{x}\mathbf{x}^\mathbf{T} \rangle \mathbf{U}^\mathbf{T} = \begin{bmatrix} \ddots & & \\ & D_i^2 & \\ & & \ddots \end{bmatrix} \tag{42}$$

Since both determinant and trace are self-invariant under a similarity transformation, we can take

$$\mathbf{U}\mathbf{C_Y}\mathbf{U}^\mathbf{T} = \hat{\mathbf{W}} \begin{bmatrix} \ddots & & \\ & D_i^2 & \\ & & \ddots \end{bmatrix} \hat{\mathbf{W}}^\mathbf{T} + \hat{\mathbf{W}} \begin{bmatrix} \ddots & & \\ & \sigma_z^2 & \\ & & \ddots \end{bmatrix} \hat{\mathbf{W}}^\mathbf{T} + \mathbf{U} \begin{bmatrix} \ddots & & \\ & \sigma_\eta^2 & \\ & & \ddots \end{bmatrix} \mathbf{U}^\mathbf{T}, \tag{43}$$

where

$$\hat{\mathbf{W}} = \mathbf{U}\mathbf{W}\mathbf{U}^\mathbf{T} = \begin{bmatrix} \ddots & & \\ & \psi_i & \\ & & \ddots \end{bmatrix} \tag{44}$$

The objective function can now be expressed as

$$F = \frac{1}{2} \sum_i \left[\log_2 \frac{D_i^2\psi_i^2 + \sigma_z^2\psi_i^2 + \sigma_\eta^2}{\sigma_z^2\psi_i^2 + \sigma_\eta^2} - 2\lambda(D_i^2\psi_i^2 + \sigma_z^2\psi_i^2 + \sigma_\eta^2) \right] \tag{45}$$

Let us first consider a special case $\sigma_z^2 = 0$ Then

$$F = \frac{1}{2} \sum_i \left[\log_2 \frac{D_i^2\psi_i^2 + \sigma_\eta^2}{\sigma_\eta^2} - 2\lambda(D_i^2\psi_i^2 + \sigma_\eta^2) \right] \tag{46}$$

By taking $\frac{\partial F}{\partial \psi_i^2} = 0$, we have

$$\begin{aligned} \frac{1}{1+\frac{D_i^2\psi_i^2}{\sigma_\eta^2}} &= 2\lambda \\ \text{or } D_i^2\psi_i^2 &= \text{const} \end{aligned} \tag{47}$$

Since the eigenvalue of the output covariance matrix is $D_i^2\psi_i^2 + \sigma_\eta^2$, efficient coding principle basically says that one would make all the eigenvalues equal, i.e., $\mathbf{C_Y} \sim \mathbf{I}$. This means efficient coding prefers decorrelation and whitening!

Now consider a more general case $\sigma_z^2 \neq 0$. By defining $\mathbf{SNR} = d_i^2 = \frac{D_i^2}{\sigma_z^2}$, $\hat{\psi_i^2} = \frac{\psi_i^2\sigma_z^2}{\sigma_\eta^2}$, $\hat{\lambda} = 2\lambda\sigma_\eta^2$, the objective function can be rewritten as

$$F = \frac{1}{2}\sum_i \left[\log_2 \frac{d_i^2\hat{\psi_i^2} + \hat{\psi_i^2} + 1}{\hat{\psi_i^2} + 1} - \hat{\lambda}(d_i^2\hat{\psi_i^2} + \hat{\psi_i^2} + 1)\right] \tag{48}$$

When the signal to noise level is small $d_i^2 \ll 1$, F is maximized when

$$\begin{aligned} \frac{d_i^2}{(\hat{\psi_i^2}+1)^2} &= \hat{\lambda} \\ \text{or } \hat{\psi_i^2} &= 0 \end{aligned} \tag{49}$$

In the opposite limit $d_i^2 \gg 1$, F is maximized when

$$\hat{\psi_i^2} \sim \frac{1}{d_i^2} \tag{50}$$

Stationary statistics and optimal coding

An important feature of many important natural stimuli is their stationarity. Consider grey level images with n pixels, the components of the input vector, x_i, which corresponds to the grey level light intensity (or contrast) of the image pixel indexed by i. Furthermore, since we are not interested in mean pixel values it is convenient to define the x_i to refer to the image pixel value relative to its mean. Natural images are well known to be redundant: obviously, neighboring pixels are highly correlated.

Spatial stationarity of the statistics of natural images means that the statistical properties are invariant to arbitrary global translations of pixels within the image. In other words, correlations between pixel values depend only on distances between the pixels, not on their absolute location. This is a very good approximation if we ignore the pixels at the boundaries of the image. A useful mathematical trick to get rid of boundary effects is impose a periodic boundary

condition: to imagine that the image is not on a two dimensional flat grid but on a torus, so that there are literally no boundaries to worry about. So we have

$$C_{ij} = \langle x_i x_j \rangle = f(|\mathbf{r_i} - \mathbf{r_j}|) \tag{51}$$

A matrix with this property is called the circulant matrix

$$C = \begin{bmatrix} c_0 & c_1 & \dots & c_{n-2} & c_{n-1} \\ c_{n-1} & c_0 & c_1 & & c_{n-2} \\ \vdots & c_{n-1} & c_0 & \ddots & \vdots \\ & & \ddots & \ddots & c_1 \\ c_2 & & & & \\ c_1 & c_2 & \dots & c_{n-1} & c_0 \end{bmatrix}.$$

Moreover, our covariance matrix is symmetric, meaning that $c_{n-i} = c_i$, and the covariant matrix is determined by $\lfloor n/2 \rfloor + 1$ elements. One important property of the circulant matrix is that its eigenvectors are Fourier basis

$$\begin{aligned} \mathbf{u_m} &= [1, \omega_m, \omega_m^2, \dots, \omega_m^{n-1}]^T \\ \omega_m &= \exp\left(\frac{i2\pi m}{n}\right) \\ m &= 0, 1, \dots, n-1 \end{aligned} \tag{52}$$

And the corresponding eigenvalue is the discrete Fourier transform of the matrix elements c_i.

$$\begin{aligned} \lambda_m = c_0 + c_1\omega_m + c_2\omega_m^2 + \dots + c_{n-1}\omega_m^{n-1} &= \sum_{j=0}^{n-1} c_j \exp(ikj) \\ k &= \frac{2\pi m}{n} \end{aligned} \tag{53}$$

Now go back to the vision problem. Let us imagine the image to be on a square lattice with $L \times L$ pixels. Then the Fourier basis vectors $\mathbf{u_m}$ are,

$$\begin{aligned} u_m^j &= \exp(i\mathbf{k_m} \cdot \mathbf{r^j}) \\ \mathbf{k} &= [k_x, k_y], \mathbf{r} = [r_x, r_y] \\ k_x, k_y = \frac{2\pi m_{x,y}}{L}, m_{x,y} &= 0, \pm 1, \dots, \pm\frac{L-1}{2} \\ r_x, r_y &= 0, \pm 1, \dots, \pm\frac{L-1}{2} \end{aligned} \tag{54}$$

The eigenvalue of the input covariance matrix $\mathbf{C_x}$ is now given by

$$\hat{c}(k_m) = D_m^2 = \sum_j c_j \exp(i\mathbf{k_m} \cdot \mathbf{r^j}) \tag{55}$$

Moreover, the eigenvalue of the covariant matrix is now, by definition the power spectra density of the image. To see this, we note that

$$\begin{aligned}\hat{c}(k) &= \sum_j c_j \exp(i\mathbf{k}\cdot\mathbf{r^j}) \\ &= \sum_i \sum_j \langle x(\mathbf{r_i})x(\mathbf{r_j})\rangle \exp(i\mathbf{k}\cdot(\mathbf{r_i}-\mathbf{r_j})) \\ &= \langle \sum_i x(\mathbf{r_i})\exp(i\mathbf{k}\cdot\mathbf{r_i}) \sum_j x(\mathbf{r_j})\exp(-i\mathbf{k}\cdot\mathbf{r_j})\rangle \\ &= \langle |\hat{x}(k)|^2\rangle\end{aligned}$$

We shall now go back to our original optimization problem. The synaptic weight matrix $\mathbf{W}$ can be viewed as the receptive field of the output neurons $\mathbf{y}$. Let us assume that all the output neurons have the same receptive field. And the only difference is that the receptive field are centered at different positions of the visual field $\mathbf{r_i}$. In other words, we shall assume the weight matrix is also translational invariant, that is $W_{ij} = w(|\mathbf{r_i} - \mathbf{r_j}|)$. Therefore, the $\mathbf{W}$ matrix is also a symmetric circulant matrix, and we can use the same Fourier basis to diagonalize $\mathbf{W}$ and $\mathbf{C_x}$.

As a result, the efficient coding principle would predict, under certain assumptions, that in the limit of high signal to noise level $d^2 = \frac{\hat{c}(k)}{\sigma_z^2} \gg 1$

$$\hat{w}(k) \sim \frac{1}{d} \sim \frac{1}{\sqrt{\hat{c}(k)}} \tag{56}$$

In the limit of low noise level, $d^2 \ll 1$,

$$\frac{\sigma_z^2}{\sigma_\eta^2}\hat{w}^2(k) = \frac{1}{\sqrt{\hat{\lambda}}}d - 1 \tag{57}$$

So depending on the tradeoff parameter λ, either $\hat{w}(k) = 0$, or

$$\hat{w}(k) \sim \frac{d^{1/2}}{\sigma_z} \sim \frac{\hat{c}^{1/4}(k)}{\sigma_z^{3/2}} \tag{58}$$

For natural images, it is well known that the power spectra density decays approximately as a power law, $\hat{c}(k) \sim \frac{1}{k^2}$, Thus we have that in the high SNR limit,

$$\begin{aligned}\hat{w}(k) &\sim k, \text{SNR} \gg 1 \\ \hat{w}(k) &\sim k^{-1/2}, \text{SNR} \ll 1\end{aligned} \tag{59}$$

Compare with experimental data...

These arguments - removing spatial correlations - can also be applied to the time domain. A time-dependent- signal has a power spectra density $S(\omega)$, and we can also ask what kind of temporal filters $K(\omega)$ would be optimal. Many naturally occuring signals have $1/f$ power spectra, this means that the optimal filter would obey $K^2(\omega) \sim \omega$ in the large SNR limit. When SNR $\ll 1$, we shall have $K^2(\omega) \sim 1/\sqrt{\omega}$.

These results can also be compared with experimental data...