Information Theory

Theoretical Neuroscience Ch 4 Notes (Dayan and Abbott)

Neha Peddinti

Jan 2021

Contents

1 Goal

1. **Previously:** Determine **what** a neural response says about a stimulus.

2. **Now:** Determine *how much* a neural response tells us about a stimulus.

2 Entropy and Mutual Information

1. Basic info theory applications

   (a) Symbols (code): neuron firing rates (simpler than using the whole spike train); using the rates, discrete or continuous, provides a lower bound on the actual amount of info carried by the spike train.

   (b) Data (info to encode): stimulus.

2. **Entropy:** measure of the level of information in a code, or how much uncertainty there is in the code's possible outcomes; measure of response variability.
   Ex: A coin flip can communicate fewer bits of information (so it has less entropy) than a word on a piece of paper, as the coin flip has fewer possible outcomes.

   (a) "Surprise" associated with recording a response rate r, as a function of the probability of getting that response: $h(P[r])$.
   Should follow these intuitive properties:

      i. Decreasing function of $P[r]$, so that low probability responses correspond to greater levels of "surprise."

      ii. Surprise measure for measuring r_1 and r_2 independently (which has probability $P[r_1]P[r_2]$) should be equal to the sum of their individual surprise measures, so that entropy can be additive for independent sources:

$$h(P[r_1]P[r_2]) = h(P[r_1]) + h(P[r_2]).$$

   (b) Only function h that works:

$$h(P[r]) = -\log_2 P[r],$$

   which is the number of bits required to express that value of r.
   Ex: If r can take on 4 different values (uniform probability), then $h(P[r]) = -\log_2(0.25) = 2$, since you need 2 bits to encode 4 options.

   (c) **Shannon's entropy:** Average the "surprise measure" over all responses.

$$H = -\sum_r P[r] \log_2 P[r].$$

   Ex: A set of identical responses has zero entropy because either $P[r] = 0$ or $\log_2 P[r] = \log_2 1 = 0$ for all terms in the above sum.

3. **Mutual information:** entropy-based measure of the correlation between response variability and stimulus variability, which quantifies how much info the response actually provides about the stimulus (as opposed to how much info it could *theoretically* provide, which is measured by entropy).

   (a) Entropy of the responses to a given stimulus s:

$$H_s = -\sum_r P[r|s] \log_2 P[r|s],$$

   where a nonzero H_s means that the same s can result in multiple different responses r, so this variation is just noise.

   (b) Noise entropy: average H_s over all stimuli.

$$H_{\text{noise}} = \sum_s P[s] H_s = -\sum_{s,r} P[s] P[r|s] \log_2 P[r|s]$$

   (c) **Mutual information:** Subtract noise entropy from full response entropy (because if all the entropy is noise, then the mutual info should be zero, so this difference measures the response variability that is actually the result of stimulus variability). Given that $P[r] = \sum_s P[s]P[r|s]$:

$$\begin{aligned} I_m = H - H_{\text{noise}} &= \sum_{s,r} P[s]P[r|s] \log_2 \left(\frac{P[r|s]}{P[r]} \right) \\ &= \sum_{s,r} P[r,s] \log_2 \left(\frac{P[r,s]}{P[r]P[s]} \right), \end{aligned}$$

   so mutual information is symmetric with respect to s and r, and also $I_m = 0$ if $P[r|s] = P[r]$, which occurs if r is independent of s.
   If each stimulus s produces a unique and distinct response r_s, then $P[r|s] = 1$ for $r = r_s$ and zero otherwise, so

$$I_m = \sum_s P[s] \log_2 \left(\frac{1}{P[r_s]} \right) = -\sum_s P[s] \log_2 P[s],$$

   which is the entropy of the stimulus, implying zero noise entropy.

4. **Kullback-Leibler (KL) divergence:** measure of the difference between probability distributions $P[r]$ and $Q[r]$,

$$D_{KL}(P,Q) = \sum_r P[r] \log_2 \left(\frac{P[r]}{Q[r]} \right).$$

   (a) $D_{KL}(P,Q) \geq 0$, with equality iff $P = Q$, a property characteristic of distance measurements (although D_{KL} is not symmetric with respect to P and Q so it's not exactly like a distance metric). This can be shown using Jensen's inequality.

   (b) Mutual information of r and s is the KL divergence between the distributions $P[r,s]$ and $P[r]P[s]$.

5. Continuous firing rates: $p[s]$ instead of $P[s]$, and r can take on a continuous range of values instead of a set of discrete values.

(a) Finding H:

$$H = -\sum p[r]\Delta r \log_2(p[r]\Delta r)$$
$$= -\sum p[r]\Delta r \log_2 p[r] - \log_2 \Delta r$$
$$\lim_{\Delta r \to 0}\{H + \log_2 \Delta r\} = -\int dr\, p[r] \log_2 p[r],$$

as H diverges when $\Delta r \to 0$, which intuitively suggests infinite entropy when r can have infinite precision. Δr in this case corresponds to the uncertainty with which firing rate can be measured.
This term usually doesn't matter because the above integral applies when you subtract two entropies computed with the same resolution Δr.

(b) **Differential entropy:** that integral above.

(c) **Noise entropy:** similar to the differential entropy.

$$\lim_{\Delta r \to 0}\{H_{\text{noise}} + \log_2 \Delta r\} = -\int ds \int dr\, p[s]p[r|s] \log_2 p[r|s]$$

(d) **Continuous mutual information:**

$$I_m = \int ds \int dr\, p[s]p[r|s] \log_2 \left(\frac{p[r|s]}{p[r]}\right),$$

where the $\log_2 \Delta r$ factor cancels completely since both entropy terms in the difference have the same resolution.

3 Information and Entropy Maximization

1. Goal: Determine whether neural responses to natural stimuli are optimized to convey as much information as possible.

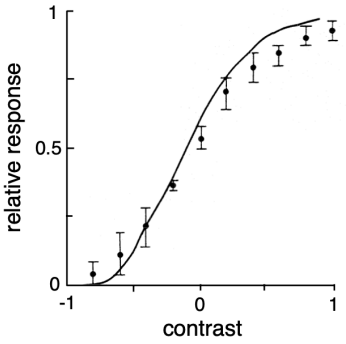
2. Method: Determine the response characteristics that maximize the mutual information conveyed, and compare with the measured neural responses.

3. Maximize response entropy (without worrying about how large noise entropy might become yet) for a single neuron.

   (a) Maximize
   $$-\int_0^{r_{\max}} dr\, p[r] \log_2 p[r],$$
   with the probability density constraint
   $$\int_0^{r_{\max}} dr\, p[r] = 1.$$
   Using Lagrange multipliers to show that the optimal probability density $p[r]$ is independent of r, so it must be uniform:
   $$p[r] = \frac{1}{r_{\max}},$$
   which gives entropy
   $$H = \log_2 \left(\frac{r_{\max}}{\Delta r}\right).$$

   (b) **Histogram equalization:** If the response probability density is the optimal uniform distribution above, $p[r] = r/r_{\max}$, then the firing rate $r = f(s)$ should map stimulus values s to responses r so that the fraction of stimuli that occur between s and $s_\Delta s$ equals the fraction of the full response range (0 to $r_{\max}$) that these stimulus values can result in:
   $$\frac{|f(s + \Delta s) - f(s)|}{r_{\max}} = p[s]\Delta s.$$
   Assuming f monotonically increases with s:
   $$\frac{\mathrm{d}f}{\mathrm{d}s} = r_{\max} p[s],$$
   so the final solution is
   $$f(s) = r_{\max} \int_{s_{\min}}^{s} ds' p[s'],$$
   where $s_{\min}$ is the minimum value of s and is assumed to correspond to the minimum value of r, $r = 0$.

   (c) Neurons that satisfy the entropy-maximizing condition: LMC (large monopolar cell) in the visual systems of flies (Laughlin, 1981).
   Graph below shows entropy-maximizing curve compared to the data points showing the response of the fly LMC produced by a light/dark image with a given level of contrast divided by the maximum response.



4. Populations of neurons: optimize population response rather than individual responses (which could result multiple neurons encoding redundant information); ensure that neurons convey independent pieces of information (they have different response selectivities).

   (a) Entropy for the entire population response with N neurons:
   $$H = -\int d\mathbf{r}\, p[\mathbf{r}] \log_2 p[\mathbf{r}] - N \log_2 \Delta r.$$

   (b) Distribution of response of neuron a:
   $$p[r_a] = \int \Pi_{b \neq a} dr_b p[\mathbf{r}.$$

Possible justification (I tried lol):

$$\int d\mathbf{r}\, p[\mathbf{r}] = 1$$
$$\int dr_a\, p[r_a] = \int dr_a \int \Pi_{b\neq a} dr_b\, p[\mathbf{r}]$$
$$= \int d\mathbf{r}\, p[\mathbf{r}] = 1.$$

(c) Entropy of neuron a:

$$H_a = -\int dr_a p[r_a] \log_2 p[r_a] - \log_2 \Delta r$$
$$= -\int d\mathbf{r} p[\mathbf{r}] \log_2 p[r_a] - \log_2 \Delta r.$$

(d) Deriving first condition (independent neuron responses/exact factorization):

$$\sum_a H_a - H = \int d\mathbf{r} p[\mathbf{r}] \log_2 \left(\frac{p[\mathbf{r}]}{\Pi_a p_[r_a]} \right) \geq 0,$$

since the integral is the KL divergence between $p[\mathbf{r}]$ and $\Pi_a p[r_a]$, and KL divergence is always non-negative.
Note: Sum of individual entropies equals full entropy when

$$p[\mathbf{r} = \Pi_a p[r_a],$$

which occurs when the neuron responses are statistically independent, and this is the maximum possible full entropy because

$$H \leq \sum_a H_a.$$

(e) Deriving second condition (probability equalization):
All neurons must individually optimize for imposed constraints on distribution shape, so all the distributions should end up looking the same if the same constraints are imposed: $p[r_a]$ is the same for all values of a.

(f) Finding response distributions that satisfy similar (less rigorous) conditions of variance equalization and decorrelation:
Due to probability equalization, $\langle r_a \rangle$ and $\langle (r_a - \langle r \rangle)^2 \rangle = \sigma_r^2$ for all a.
Covariance matrix should be proportional to the identity matrix if the neurons are statistically independent:

$$Q_{ab} = \int d\mathbf{r} p[\mathbf{r}](r_a - \langle r \rangle)(r_b - \langle r \rangle) = \sigma_r^2 \delta_{ab},$$

where δ_{ab} is the Kronecker delta.

(g) Application to retinal ganglion cell receptive fields (RFs):
Use the methods of Ch 2 to determined that the linear spatial response of each cell, with an RF centered at the point $\mathbf{a}$, is:

$$L_s(\mathbf{a}) \int d\mathbf{x} D_s(\mathbf{x} - \mathbf{a}) s_s(\mathbf{x},$$

where $s_s\mathbf{x}$ is the stimulus at point $\mathbf{x}$ and D_s is the spatial kernel.
Neuron is being specified by parameter $\mathbf{a}$ instead of an index a so that the population can be represented with a continuous integral over $\mathbf{a}$ (since the retina is densely populated with neurons that have closely intersecting RF fields).
Applying decorrelation condition to two cells with RFs centered at $\mathbf{a}$ and $\mathbf{b}$:

$$Q_{LL}(\mathbf{a}, \mathbf{b}) = \langle L_s(\mathbf{a}) L_s(\mathbf{b}) \rangle$$
$$= \int d\mathbf{x} d\mathbf{y} D_s(\mathbf{x} - \mathbf{a}) D_x(\mathbf{y} - \mathbf{b}) \langle s_s(\mathbf{x}) s_s(\mathbf{y}) \rangle$$
$$= \sigma_L^2 \delta(\mathbf{a} - \mathbf{b}),$$

where the correlation function of the stimulus, averaged over all possible natural scenes that the RF is believed to be optimized for, is

$$Q_{ss}(\mathbf{x} - \mathbf{y}) = \langle s_s(\mathbf{x}) s_s(\mathbf{y}) \rangle.$$

Solving (which requires Fourier transforms), gives us an optimal kernel/RF, known as the **whitening filter**, described by:

$$\left| \hat{D}_s(\kappa) \right| = \frac{\sigma_L}{\sqrt{\hat{Q}_{ss}(\kappa)}},$$

which can correspond to several functions $\hat{D}_s(\kappa)$. Note that κ is just the parameter for the Fourier transform of D_s and Q_{ss}, which is the spatial frequency of the stimulus.
This is called the whitening filter because the power spectrum of L is

$$\left| \hat{D}_s(\kappa) \hat{Q}_{ss}(\kappa) \right| = \sigma_L^2,$$

which is independent of spatial frequency **kappa** and therefore similar to white noise. This means that decorrelation and variance equalization require different spatial frequencies to be encoded at equal signal strength.
Noise filter: similar calculations, but to get a filter that suppresses signals with very low stimulus power, while the whitening filter boosts signals with only fairly low stimulus power.
Test if cat LGN cells acted as temporal whitening filters: Cats were shown a movie, which caused a roughly even power distribution of different temporal frequencies, and then a white-noise stimulus, which had large variations in power distributions over different frequencies (see diagram below).

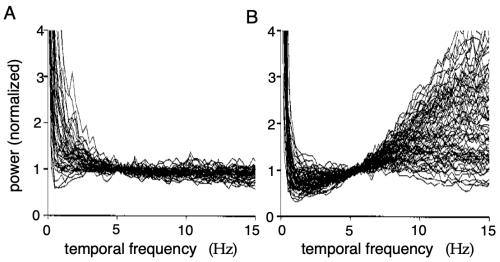


Figure 4.5 (A) Power spectra of the spike trains of 51 cat LGN cells in response to presentation of the movie *Casablanca*, normalized to their own values between 5 and 6 Hz. B) Equivalently normalized power spectra of the spike trains of 75 LGN cells in response to white-noise stimuli. (Adapted from Dan et al., 1996.)

4 Entropy and Information for Spike Trains

1. Goal: Use spike trains instead of firing rates to determine the information content of a neuronal response, in order to avoid large underestimates of entropy.

2. **Entropy/information rate $\dot{\mathbf{H}}$:** total entropy divided by spike train duration, since the entropy typically grows linearly with the length of the spike train.

3. Entropy of independent interspike intervals (amount of time between spikes), where $p[\tau]$ is the ISI distribution, and $p[\tau]\Delta\tau$ is the probability of an ISI measured to be between τ and $\tau + \Delta\tau$, with measurement resolution $\Delta\tau$:

$$\dot{H} = -\frac{N}{T}\int p[\tau]\log_2(p[\tau]\Delta\tau).$$

   This is an upper bound, since correlations between ISIs reduce the total amount of information they encode.
   $N = \langle r\rangle T$ is the number of intervals, so in general:

$$\dot{H} \leq -\langle r\rangle \int_0^\infty p[\tau]\log_2(p[\tau]\Delta\tau).$$

4. If the spike train can be modeled by a homogeneous Poisson process with rate $\langle r\rangle$: $p[\tau] = \langle r\rangle \exp(-\langle r\rangle\tau)$ and the ISIs are statistically independent, so:

$$\dot{H} = \frac{\langle r\rangle}{\ln(2)}(1 - \ln(\langle r\rangle\Delta\tau)),$$

   so in general, neural responses with higher average firing rates that can be measured with lower temporal resolutions seem to have higher entropy.

5 Summary

1. Shannon's information theory: provides measures of entropy and mutual information to determine how much a neural response actually encodes about a stimulus.

2. Mutual information is related to the KL divergence between two probability distributions.

3. Determining if cat LGN cells optimized the mutual information that a population encoded:

   (a) It can be shown that optimizing information from a population is roughly similar to using a whitening filter for the RFs of the retinal cells, where optimal information encoding occurs when the power distribution of the signals across different stimuli is constant.

   (b) When the cat was shown a movie (a set of natural stimuli), the recorded power distribution of the spike trains of neurons that responded to different frequencies was roughly constant, suggesting that the model of the whitening filter (which implies maximum entropy techniques) is a decent fit.

4. Entropy can be extracted from spike-trains (if you collect large amounts of data with long trial durations T so that the correlation between trials is minimized) by finding the probability distribution of various spike trains and calculating entropy with the relevant equation.
   Statistical mechanics arguments suggest that the true entropy for $T \to \infty$ is proportional to $1/T$ for large finite T, so true entropy can be estimated by plotting the calculated entropy rate (from experiments) vs $1/T$. This seems to align with data from an H1 visual neuron of a fly responding to a randomly moving visual stimulus.