Why information theory

Understanding the neural code.
- Encoding and decoding. We imposed coding schemes, such as 2nd-order kernel, or NLP. We possibly lost information in doing so.
- Instead, use information:
  - Don’t need to impose encoding or decoding scheme (non-parametric).
  - In particular important for 1) spike timing codes, 2) higher areas.
  - Estimate how much information is coded in certain signal.

Caveats:
- No easy decoding scheme for organism (upper bound only)
- Requires more data and biases are tricky

Overview

- Entropy, Mutual Information
- Entropy Maximization for a Single Neuron
- Maximizing Mutual Information
- Estimating information
- Reading: Dayan and Abbott ch 4, Rieke

Definition

The entropy of a quantity is defined as
\[ H(X) = - \sum_x P(x) \log_2 P(x) \]
This is not ‘derived’, nor fully unique, but it fulfills these criteria:
- Continuous
- If \( p_i = \frac{1}{n} \), it increases monotonically with \( n \). \( H = \log_2 n \).
- Parallel independent channels add.

“Unit”: bits
Entropy can be thought of as physical entropy, “richness” of distribution [Shannon and Weaver, 1949, Cover and Thomas, 1991, Rieke et al., 1996]
Entropy

Discrete variable

\[ H(R) = - \sum_r p(r) \log_2 p(r) \]

Continuous variable at resolution \( \Delta r \)

\[ H(R) = - \sum_r p(r) \Delta r \log_2 (p(r) \Delta r) \]

letting \( \Delta r \to 0 \) we have

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

(Also called differential entropy)

Joint, Conditional entropy

Joint entropy:

\[ H(S, R) = - \sum_{r,s} P(S, R) \log_2 P(S, R) \]

Conditional entropy:

\[ H(S|R) = \sum_r P(R = r) H(S|R = r) \]

\[ = - \sum_r P(r) \sum_s P(s|r) \log_2 P(s|r) \]

\[ = H(S, R) - H(R) \]

If \( S, R \) are independent

\[ H(S, R) = H(S) + H(R) \]

Mutual information

Mutual information:

\[ I_m(R; S) = \sum_{r,s} p(r,s) \log_2 \frac{p(r,s)}{p(r)p(s)} \]

\[ = H(R) - H(R|S) = H(S) - H(S|R) \]

- Measures reduction in uncertainty of \( R \) by knowing \( S \) (or vice versa)
- \( I_m(R, S) \geq 0 \)
- The continuous version is the difference of two entropies, the \( \Delta r \) divergence cancels

Mutual Information

The joint histogram determines mutual information.

Given \( P(r, s) \Rightarrow I_m \).
Mutual Information: Examples

\[ I(Y_1; Y_2) = \int \int P(y_1, y_2) \log_2 \frac{P(y_1, y_2)}{P(y_1)P(y_2)} \, dy_1 \, dy_2 = -\frac{1}{2} \log_2 (1 - \rho^2) \]

\[ \rho \] is (Pearson-r) correlation coefficient.

Kullback-Leibler divergence

- KL-divergence measures distance between two probability distributions
- \( D_{KL}(P||Q) = \int P(x) \log_2 \frac{P(x)}{Q(x)} \, dx \), or \( D_{KL}(P||Q) = \sum_i P_i \log_2 \frac{P_i}{Q_i} \)
- Not symmetric, but can be symmetrized
- \( I_m(R; S) = D_{KL}(p(r, s)||p(r)p(s)) \).
- Often used as probabilistic cost function: \( D_{KL}(data||model) \).
- Other probability distances exist (e.g. earth-movers distance)
Given

\[ H(R) = -\int p(r) \log_2 p(r) dr - N \log_2 \Delta r \]
and

\[ H(R_i) = -\int p(r_i) \log_2 p(r_i) dr - \log_2 \Delta r \]

We have

\[ H(R) \leq \sum_i H(R_i) \]
(proof, consider KL divergence)

Redundancy can be defined as (compare to above)

\[ R = \sum_{i=1}^{n_r} I(r_i; s) - I(r; s) \]

Some codes have \( R > 0 \) (redundant code), others \( R < 0 \) (synergistic)

Example of synergistic code: \( P(r_1, r_2, s) \) with

\[ P(0,0,1) = P(0,1,0) = P(1,0,0) = P(1,1,1) = \frac{1}{4} \]

Entropy Maximization for a Single Neuron

\[ I_m(R; S) = H(R) - H(R|S) \]

- If noise entropy \( H(R|S) \) is independent of the transformation \( S \rightarrow R \), we can maximize mutual information by maximizing \( H(R) \) under given constraints
- Possible constraint: response \( r \) is \( 0 < r < r_{\text{max}} \). Maximal \( H(R) \) if \( p(r) \sim U(0, r_{\text{max}}) \) (\( U \) is uniform dist)
- If average firing rate is limited, and \( 0 < r < \infty \) : exponential distribution is optimal \( p(x) = 1/\bar{x} \exp(-x/\bar{x}) \). \( H = \log_2 e\bar{x} \)
- If variance is fixed and \( -\infty < r < \infty \): Gaussian distribution. \( H = \frac{1}{2} \log_2(2\pi e\sigma^2) \) (note funny units)

Let \( r = f(s) \) and \( s \sim p(s) \). Which \( f \) (assumed monotonic) maximizes \( H(R) \) using max firing rate constraint? Require:

\[ P(r) = \frac{1}{r_{\text{max}}} \]

\[ p(s) = p(r) \frac{dr}{ds} = \frac{1}{r_{\text{max}}} \frac{df}{ds} \]

Thus \( df/\text{ds} = r_{\text{max}} p(s) \) and

\[ f(s) = r_{\text{max}} \int_{s_{\text{min}}}^{s} p(s') ds' \]

This strategy is known as histogram equalization in signal processing
Fly retina

Evidence that the large monopolar cell in the fly visual system carries out histogram equalization

Contrast response for fly large monopolar cell (points) matches environment statistics (line) [Laughlin, 1981] (but changes in high noise conditions)

Information of time varying signals

Single analog channel with Gaussian signal \( s \) and Gaussian noise \( \eta \):

\[
I = \frac{1}{2} \log_2 \left( 1 + \frac{\sigma_s^2}{\sigma_\eta^2} \right) = \frac{1}{2} \log_2 (1 + \text{SNR})
\]

For time dependent signals \( I = \frac{1}{2} T \int \frac{d\omega}{2\pi} \log_2 (1 + \frac{s(\omega)}{n(\omega)}) \)

To maximize information, when variance of the signal is constrained, use all frequency bands such that signal+noise = constant.

Whitening. Water filling analog:

Information of graded synapses

Light - (photon noise) - photoreceptor - (synaptic noise) - LMC

At low light levels photon noise dominates, synaptic noise is negligible.

Information rate: 1500 bits/s

[de Ruyter van Steveninck and Laughlin, 1996]
Spiking neurons: maximal information

Spike train with \( N = T / \delta t \) bins [Mackay and McCulloch, 1952] \( \delta t \) “time-resolution”.

\[ pN = N_1 \text{ events, } \#\text{words} = \frac{N!}{N_1!(N-N_1)!} \]

Maximal entropy if all words are equally likely.

\[ H = \sum p_i \log_2 p_i = \log_2 N! - \log_2 N_1! - \log_2 (N-N_1)! \]

Use for large \( x \) that \( \log x! \approx x(\log x - 1) \)

\[ H = -\frac{T}{\delta t} \left[ p \log_2 p + (1-p) \log_2 (1-p) \right] \]

For low rates \( p \ll 1 \), setting \( \lambda = (\delta t)p \):

\[ H = T \lambda \log_2 \left( \frac{e \lambda \delta t}{\lambda} \right) \]

Calculation incorrect when multiple spikes per bin. Instead, for large bins maximal information for exponential distribution:

\[ P(n) = \frac{1}{Z} \exp[-n \log(1 + \frac{1}{\langle n \rangle})] \]

\[ H = \log_2(1 + \langle n \rangle) + \langle n \rangle \log_2(1 + \frac{1}{\langle n \rangle}) \approx \log_2(1 + \langle n \rangle) + 1 \]

---

Spiking neurons: rate code

[Stein, 1967]

Measure rate in window \( T \), during which stimulus is constant.

Periodic neuron can maximally encode \( 1 + (f_{max} - f_{min})T \) stimuli

\[ H \approx \log_2 [1 + (f_{max} - f_{min})T] \]. Note, only \( \propto \log(T) \)

---

[Stein, 1967]

Similar behaviour for Poisson: \( H \propto \log(T) \)
Spiking neurons: dynamic stimuli

[de Ruyter van Steveninck et al., 1997], but see [Warzecha and Egelhaaf, 1999].

Maximizing Information Transmission: single output

Single linear neuron with post-synaptic noise

\[ \mathbf{v} = \mathbf{w} \cdot \mathbf{u} + \eta \]

where \( \eta \) is an independent noise variable

\[ I_m(\mathbf{u}; \mathbf{v}) = H(\mathbf{v}) - H(\mathbf{v} | \mathbf{u}) \]

- Second term depends only on \( p(\eta) \)
- To maximize \( I_m \) need to maximize \( H(\mathbf{v}) \); sensible constraint is that \( ||\mathbf{w}||^2 = 1 \)
- If \( \mathbf{u} \sim N(0, Q) \) and \( \eta \sim N(0, \sigma^2_\eta) \) then \( \mathbf{v} \sim N(0, \mathbf{w}^T Q \mathbf{w} + \sigma^2_\eta) \)

Infomax

Infomax: maximize information in multiple outputs wrt weights [Linsker, 1988]

\[ \mathbf{v} = \mathbf{W} \mathbf{u} + \eta \]

\[ H(\mathbf{v}) = \frac{1}{2} \log \det(\langle \mathbf{v} \mathbf{v}^T \rangle) \]

Example: 2 inputs and 2 outputs. Input is correlated. \( w^2_{k1} + w^2_{k2} = 1 \).

For a Gaussian RV with variance \( \sigma^2 \) we have \( H = \frac{1}{2} \log 2\pi e \sigma^2 \).

To maximize \( H(\mathbf{v}) \) we need to maximize \( \mathbf{w}^T Q \mathbf{w} \) subject to the constraint \( ||\mathbf{w}||^2 = 1 \)

Thus \( \mathbf{w} \propto \mathbf{e}_1 \) so we obtain PCA

If \( \mathbf{v} \) is non-Gaussian then this calculation gives an upper bound on \( H(\mathbf{v}) \) (as the Gaussian distribution is the maximum entropy distribution for a given mean and covariance)

At low noise independent coding, at high noise joint coding.
Estimating information

Information estimation requires a lot of data. Most statistical quantities are unbiased (mean, var,...). But both entropy and noise entropy have bias.

Try to fit $1/N$ correction [Strong et al., 1998]

Common technique for $I_m$: shuffle correction [Panzeri et al., 2007]
See also: [Paninski, 2003, Nemenman et al., 2002]

Summary

- Information theory provides non parametric framework for coding
- Optimal coding schemes depend strongly on noise assumptions and optimization constraints
- In data analysis biases can be substantial
References I


References II


References III