Different equivalent views of Mutual Information:

- deviation from independence:

\[ I(X; Y) = D(P(X, Y) \| P(X)P(Y)) = \sum_x \sum_y P(x, y) \log \left( \frac{P(x, y)}{P(x)P(y)} \right) \]

- savings in encoding:

\[ I(X; Y) = H(X) + H(Y) - H(X|Y) \]

- why are these equivalent? (see homework)
Define conditional entropy $H(Y|X)$ as expected value of entropies of $Y$ averaged over fixed values of $X$:

$$H(Y|X) = \sum_{x} P(x) H(Y|X = x)$$

$$= - \sum_{x} P(x) \sum_{y} P(y|x) \log P(y|x)$$

$$= - \sum_{x} \sum_{y} P(x, y) \log P(y|x)$$

$$= - \mathbb{E}_{P(x,y)} [\log P(y|x)]$$
Chain rule for conditional entropy:

\[ H(X, Y) = H(X) + H(Y | X) \]

**Proof idea:** start with \( H(X, Y) \) and decompose it.

This relation allows another view of mutual information as the entropy of random variable \( X \) minus the conditional entropy of that random variable given another random variable \( Y \):

\[ I(X; Y) = H(X) - H(X | Y) \]

**Proof:** follows directly from definitions and laws of probability.
Application to Neuroscience

Consider recording from neuron (response $r$) while you present different stimuli $s$ to the animal.

The conditional entropy of the neuron’s response $r$ given the stimulus $s$ is sometimes called the noise entropy:

$$H_{\text{noise}} = H(r|s) = \sum_{s,r} P(s)P(r|s) \log P(r|s)$$

The mutual information between stimulus and response can then be expressed as:

$$I(r; s) = H(r) - H_{\text{noise}}$$
Differential Entropy

**Idea:** generalize to continuous random variables described by pdf:

\[ H(X) \equiv -\int_{-\infty}^{\infty} p(x) \log(p(x)) \, dx \]

**Notes:**
- differential entropy can be negative, in contrast to entropy of discrete random variable
- but still: the smaller differential entropy, the “less random” is \( X \)

**Example:** uniform distribution

\[ p(x) = \begin{cases} \frac{1}{a}, & \text{for } 0 \leq x \leq a \\ 0, & \text{otherwise} \end{cases} \]

\[ H(X) = -\int_{0}^{a} \frac{1}{a} \log\left(\frac{1}{a}\right) \, dx = \log(a) \]
Justification:
Probability for firing rate lying in the range between $r$ and $r+\Delta r$, for small $\Delta r$, is $p(r)\Delta r$
Thus, for the entropy we have

$$H = - \sum p(r) \Delta r \log_2 (p(r)\Delta r)$$
$$= - \sum p(r) \Delta r \log_2 p(r) - \log_2 \Delta r$$

Note: Diverges for $\Delta r \to 0$! A continuous variable measured with perfect accuracy has infinite entropy.
However, for reasonable $p$ we can write in this limit

Differential entropy:

$$\lim_{\Delta r \to 0} \{H + \log_2 \Delta r\} = - \int dr \; p(r) \log_2 p(r)$$
Continuous noise entropy:

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

Continuous mutual information:

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

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

Note that the factor of \( \log_2 \Delta r \) cancels because both entropies are evaluated at the same resolution.
4.2 Information and Entropy Maximization
Question: Consider single neuron with continuous response \( r \). When does distribution of neuron’s responses have the maximum (differential) entropy? When is this neuron most „informative“?

Answer: maximum entropy distributions are „most random”
- For discrete RV, uniform distribution has maximum entropy (see earlier)
- For continuous RV, need to consider additional constraints on the distributions.
- Different constraints lead to different maximum entropy distributions.
Important results:

• for a **fixed range** of the RV, i.e. \( \text{min} < x < \text{max} \), the uniform distribution has maximum entropy

• for a **fixed variance**, the Gaussian distribution has the highest entropy (another reason why the Gaussian is so special)

• for a **fixed mean** and \( p(x) = 0 \) if \( x < 0 \), the exponential distribution has the highest entropy. Do some neurons in the brain have approximately exponential firing rate distributions because this allows them to be most “informative” given a fixed average firing rate, which corresponds to a certain level of average energy? consumption? Proof: see blackboard 4.2
Cortical Neurons

- neurons in visual cortex of cat and monkey have close to exponential firing rate distributions; may maximize entropy for fixed energy consumption [Baddeley et al., 1997]
**Making a neuron’s response uniform**

**Question:** how much should neuron respond to stimulus of strength $s$ with distribution $p(s)$ such that the pdf of its response $r$ is uniform in interval $[0, R]$?

We seek a monotonic function $f$ such that the distribution of $r=f(s)$ is uniform.

**Solution:**

$$f(s) = R \int_{s_{\text{min}}}^{s} p(s') ds' = RF_S(s)$$

**cumulative distribution function**
Proof:
Probability of stimulus $s$ falling in $[s, s+\Delta s]$ is $p(s)\Delta s$. This produces responses falling in $[f(s), f(s+\Delta s)]$.
If distribution of output responses is flat ($p(r) = 1/R$), then the probability of the response falling in this interval should be $|f(s+\Delta s)-f(s)|/R$.

Thus we want: $|f(s+\Delta s)-f(s)|/R = p(s)\Delta s$

Now assuming that $f$ is monotonically increasing and considering the limit $\Delta s \to 0$, we have:

$$\frac{df(s)}{ds} = Rp(s)$$

The solution to this is the result from the previous slide.
Uniform distribution of responses in large monopolar cell of fly

contrast response of neuron compared to integral of natural contrast probability distribution
Histogram equalization

Passing a random variable through its own cumulative distribution function creates a random variable that is uniformly distributed over [0,1].

This is frequently used in signal processing, e.g. image processing.
Entropy maximization for population of neurons

Just making each neuron informative individually does not imply that the population codes effectively: neurons could be redundant.

Ideal situation is when neurons are each maximizing entropy, i.e. they should have the same marginal distribution (probability equalization) and they should be independent (factorial code). For a factorial code the joint entropy becomes the sum of the individual entropies.

Independence is usually difficult to achieve. Decorrelation is usually easier to obtain.
Generalization of Mutual Information to multiple RVs: Total correlation, multivariate constraint, multiinformation

\[ I(X;Y) \equiv D(P(X,Y) \| P(X)P(Y)) = \sum_x \sum_y P(x,y) \log \left( \frac{P(x,y)}{P(x)P(y)} \right) \]

\[ I(X_1, X_2, ..., X_n) = D(p(X_1, X_2, ..., X_n) \| p(X_1)p(X_2)...p(X_n)) \]

\[ I(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} H(X_i) - H(X_1, X_2, ..., X_n), \text{ where} \]

\[ H(X_1, X_2, ..., X_n) \equiv -E[\log p(X_1, X_2, ..., X_n)] \]
Recall: independence implies decorrelation but not vice versa.

If decorrelation is the goal we want:

$$\text{Cov}(r_i, r_j) = \sigma^2 \delta_{ij}$$

This also makes all variances identical. This problem is generally more tractable (see book for details).

Nice examples: prediction of receptive field properties in the Retina and LGN based on decorrelation idea.
4.3 Entropy and Information for Spike Trains
Motivation

• it’s obviously useful to try to quantify the entropy in a spike train or the information about a stimulus in a spike train

• for long spike train recordings, entropy will typically grow linearly with the length of the spike train

• to get a measure that’s independent of recording duration we can calculate *entropy rates*:
  • entropy per time [bits/s]
  • entropy per spike [bits/spike]

• basic problem: need to estimate probabilities of various temporal patterns of action potentials appearing
Approach

• divide spike train of duration $T$ into independent subunits of much smaller duration $T_s$

• now divide each subunit $T_s$ into small intervals $\Delta t$, such that at most one spike falls into each interval

• we can now represent a part of the spike train as a binary vector of length $T_s / \Delta t$ : 0=no spike, 1=spike

• We now have many such binary vectors $B$ and can estimate their probabilities of occurring $P[B]$

• This gives a spike train entropy rate estimate:

$$\dot{H} = -\frac{1}{T_s} \sum_B P[B] \log_2 P[B]$$
\[ \dot{H} = -\frac{1}{T_s} \sum_B P[B] \log_2 P[B] \]

- this estimate will only be accurate if subsequent subunits of duration \( T_s \) are independent, so there must not be a correlation between bins which are \( T_s \) apart \( \rightarrow \) want big \( T_s \)

- if \( T_s \) very big, however, we can't estimate the \( P[B] \) accurately anymore \( \rightarrow \) want small \( T_s \)

- recipe: increase \( T_s \) until estimation of probabilities becomes problematic and extrapolate
Example

Strong et al. 1998: H1 visual neuron in the fly
• to estimate mutual information we have to subtract the noise entropy rate from the total entropy rate: for this consider entropy of patterns $B$ generated **by the same stimulus** and average this over stimuli.

• result in example above:
  • entropy rate: 157 bits/s
  • noise entropy rate: 79 bits/s
  • information rate: 78 bits/s (corresponds to 1.8 bits/spike)

• Note: binning size $\Delta t$ can influence results. More information can be extracted from more accurately measured spike trains.
4.4 Summary
Key Concepts

• entropy
• mutual information
• conditional entropy
• Kullback-Leibler Divergence
• maximum entropy distributions
• histogram equalization
• factorial code
• decorrelation
• entropy rate, information rate