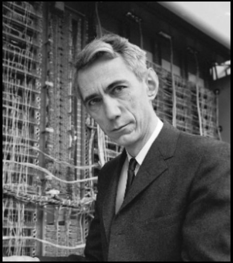


4. Information Theory

Sungyoon Kim

04. 14. 2023

Introduction

- Until now, we have talked about how our world (the ”stimulus”) interacts with our brain (the ”response”).
- Encoding is how our brain responds to the world (with ”neural spikes”), and decoding is how we can interpret the response of our brain to understand the stimulus.
- Can we quantify how ”good” an encoding scheme is, or how ”close” the stimulus and the response are related?
- Shannon’s information theory!

Table of Contents

- Entropy and mutual information
  - Motivation
  - Definition of entropy and mutual information
  - Entropy and mutual information for continuous variables
- Information and entropy maximization
  - Single neuron case
  - Population of neurons
  - Application to retinal ganglion cell receptive fields
- Entropy and information for spike trains

Entropy: Motivation

- Consider an encoding scheme that maps each stimulus to a response (in a probabilistic manner). How can we determine ”how informative” the scheme is?
- Basic idea 1: When a high-probability event happens, it does not give us a lot of information (because it is likely to happen). On the other hand, when a low-probability event happens, it gives us a lot of information.
- *Information* $\approx$ *Surprise*

Designing Surprise

- For a probability distribution P, we may define how ”surprising” or ”informative” a specific event r_s is by defining a surprise function $h(x) : \mathbb{R}^+ \rightarrow \mathbb{R}$ that maps a probability to a ”surprise”.
- Property 1: From basic idea 1, we know that the surprise function should be a decreasing function of x.
- Property 2: When two independent events happen, we want the surprise to be simply added. For independent events r_1 and r_2,

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

must hold, and $h(xy) = h(x) + h(y)$ should hold in general.

Entropy

- $h(x) = -\log(x)$ is a good candidate for illustrating surprise. We use base 2 for log as a convention from communications theory (related to ”bits”).
- The entropy, or the quantitive measure of surprise for the probability distribution is simply the expectation of surprise.

Definition

For a discrete probability distribution P, we define its entropy to be the expectation of surprise

$$H = -\sum_r P[r] log(P[r])$$

Examples

- When the response is fixed for each stimulus, we can see that each $P[r]log(P[r]) = 0$ for every response. Thus, the entropy for this encoding scheme is zero (i. e. it is not informative at all)
- When the response is fixed in a twofold manner, with r_+ and r_-, the entropy becomes

$$H = -(1 - P[r_+]) \log(1 - P[r_+]) - P[r_+] \log(P[r_+])$$

  Jensen's inequality and the fact that $-x \log(x)$ is concave leads to the fact that the entropy is maximized when $P[r_+] = P[r_-] = \frac{1}{2}$

Mutual Information: Motive

- We have defined the ”concept of entropy” to quantify the total information that the probability distribution has. Is it enough to understand how informative the response is?
- **No.**
- Suppose the response is fixed in a twofold manner, and for ”any” stimulus the neuron responds as either r_+ or r_- randomly. Then, the entropy is maximized, but the response is giving ”zero” information about the stimulus!
- Entropy is the capacity of the response (how informative it can be). We need another concept to understand how informative it is regarding to ”a particular stimulus”.

Mutual Information: Motive

- There could be two reasons that the response is holding information:
    1) Different stimuli lead to different responses. The variety of the response due to the variety of stimuli contributed to the total information.
    2) The response may change ”even if” the stimulus is the same (noisy channel, stochasticity, ...). The variety due to the noise in the response may contribute to the total information
- We want to eliminate the effect of 2) and only see the effect of 1) to understand how the response and stimulus is related.
- 1): Mutual Information, 2): Noise Entropy
- **Mutual Information + Noise Entropy = Total Entropy**

Noise Entropy

- The information of the response for each stimulus is

$$H_s = -\sum_r P[r|s]\log(P[r|s])$$

- Remark: When each response is deterministic for each stimulus, $H_s = 0$.
- The noise entropy is simply the expectation of H_s.

Definition

For a discrete probability distribution P, stimuli s and response r, we define the noise entropy as the expectation of H_s,

$$H_{noise} = \sum_s P[s]H_s = -\sum_{s,r} P[s]P[r|s]\log(P[r|s])$$

Mutual Information

- Subtract the noise entropy from the full response entropy to obtain mutual information.

Definition

For a discrete probability distribution P, stimuli s and response r, we define the mutual information of s and r to be

$$I_m = H - H_{noise} = -\sum_r P[r] \log(P[r]) + \sum_{s,r} P[s]P[r|s] \log(P[r|s])$$

- Using $P[r] = \sum_s P[s]P[r|s]$ and simplifying leads to

$$I_m = \sum_{s,r} P[s]P[r|s] \log(\frac{P[r|s]}{P[r]}) = \sum_{s,r} P[r,s] \log(\frac{P[r,s]}{P[r]P[s]})$$

Mutual Information: Intuitions and remarks

- We can see that

  $$I_m = \sum_{s,r} P[r,s] \log(\frac{P[r,s]}{P[r]P[s]}) = \sum_{s,r} P[r,s](-\log(P[r]P[s]) + \log(P[r,s]))$$

  The term $-\sum_{s,r} P[r,s]\log(P[r]P[s])$ gives the amount of information ”when the response and stimulus are independent”, and the term $-\sum_{s,r} P[r,s]\log(P[r,s])$ gives the amount of information of the current joint distribution. The discrepancy between the two is happening as r and s are related, and subtracting the two leads to mutual information.
- We can also understand I_m as the KL divergence

  $$D_{KL}(P,Q) = \sum_{r} P[r] \log(\frac{P[r]}{Q[r]})$$

  between the distribution $P[r,s]$ and $P[r]P[s]$.

Mutual Information: Intuitions and remarks

- $I_m \leq -\sum_s P[s] \log(P[s]), I_m \leq -\sum_r P[r] \log(P[r])$, and $0 \leq I_m$
- I_m is symmetric with respect to stimulus and response.
- When the response and stimulus is totally independent, $I_m = 0$. On the other hand, when the response is deterministic, $I_m = -\sum_s P[s] \log(P[s])$, becoming maximal.

Entropy and mutual information for continuous variables

- Consider the probability density function of the response, $p[r]$.
- The difference between the continuous case and the discrete case is that when the variables become continuous, we cannot consider the probability of ”each event”. Rather, we should consider the probability of ”each interval”.
- When the resolution of the response is Δr, we may write the entropy as

$$H = -\sum_r p[r]\Delta r \log(p[r]\Delta r) = -\sum_r p[r]\log(p[r])\Delta r - \log(\Delta r)$$

- As $\Delta r \to 0$, the entropy diverges!

Entropy and mutual information for continuous variables

- For continuous case, we can only obtain the entropy up to an additive constant. We write
$$\lim_{\Delta r \to 0} H + \log(\Delta r) = -\int dr p[r] \log(p[r])$$
where $\log \Delta r$ is best thought of as a limit on the resolution.
- However, when the two entropies are subtracted, we can obtain the exact value.
- As mutual information is the subtraction between the full entropy and the noise entropy, it is exactly determined as the following integral,
$$I_m = \int ds \int dr p[s] p[r|s] \log(\frac{p[r|s]}{p[r]})$$

Information and entropy maximization: motives and objectives

- ”Are neural responses maximizing mutual information?”
- To answer the question, we need both experimental results and theoretical calculations that gives us the optimal response.
- Maximizing the mutual information involves two steps, maximizing the total entropy and minimizing the noise entropy.
- In this section we study how to maximize the total entropy ”within given constraints”, and later on discuss the effect of noise entropy.
- Possible constraints: maximal firing rate, average firing rate, variance of firing rate, ...

Entropy maximization: Single Neuron, maximal firing rate

- Constraint: Maximal firing rate r_{max}, $\int_0^{r_{max}} dr p[r] = 1$

Problem

Maximize

$$-\int_0^{r_{max}} dr p[r] \log(p[r])$$

subject to

$$\int_0^{r_{max}} dr p[r] = 1$$

- Solution: Lagrange multipliers

Entropy maximization: Single Neuron, maximal firing rate

The Lagrangian for the optimization problem becomes

$$\int_0^{r_{max}} drp[r]\log(p[r]) + \lambda \int_0^{r_{max}} drp[r] = \int_0^{r_{max}} dr(\lambda p[r] + p[r]\log p[r])$$

The critical point of $x \log x + \lambda x$ is $x = 2^{-\lambda-1}$ by direct calculation. Thus, for each λ, $p[r] = 2^{-\lambda-1}$ becomes the stationary point of the Lagrangian, and the integral, if it has its critical value, has one when $p[r]$ is constant.

Now, let's show that $p[r] = \frac{1}{r_{max}}$ becomes the probability distribution that maximizes the entropy. As the KL- divergence is always positive, we know

$$-\int_0^{r_{max}} drp[r]\log(r_{max}p[r]) \leq 0, \quad -\int_0^{r_{max}} drp[r]\log(p[r]) \leq \log(r_{max}),$$

finishing the proof.

Histogram Equalization

- Suppose each response r for stimulus s is given as $r = f(s)$, and assume f is monotone.
- The probability that the stimulus is in $[s, s + \Delta s]$ is $p[s]\Delta s$, where the probability that the response is in $[f(s), f(s + \Delta s)]$ is $p[f(s)](f(s + \Delta s) - f(s))$.
- For optimal $p[r] = \frac{1}{r_{max}}$, we see that

  $$p[s]\Delta s = \frac{f(s + \Delta s) - f(s)}{r_{max}}$$

  and as $\Delta s \to 0$ we obtain

  $$\frac{df}{ds} = r_{max} p[s], \quad f(s) = r_{max} \int_{s_{min}}^{s} p[u] du$$

- Intuition: The formula can be understood as a change of variables between two probability density functions, $p[r] = \frac{1}{r_{max}}$ and $p[s]$.

Entropy Maximization in LMC

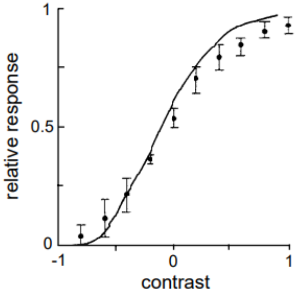


Figure 4.2: Contrast response of the fly LMC (data points) compared to the integral of the natural contrast probability distribution (solid curve). The relative response is the amplitude of the membrane potential fluctuation produced by the onset of a light or dark image with a given level of contrast divided by the maximum response. Contrast is defined relative to the background level of illumination, I_b, as $(I - I_b)/I_b$. (Adapted from Laughlin, 1981.)

Entropy maximization: Single Neuron, average firing rate

- Constraint: Average firing rate $\int_0^\infty dr\ rp[r] = r_{avg}$, $\int_0^\infty dr\ p[r] = 1$

Problem

Maximize

$$-\int_0^{r_{max}} drp[r]\log(p[r])$$

subject to

$$\int_0^\infty dr\ rp[r] = r_{avg}, \quad \int_0^\infty drp[r] = 1$$

The Lagrangian for the optimization problem becomes

$$\int_0^\infty drp[r]\log(p[r]) + \lambda_1 \int_0^\infty drp[r] + \lambda_2 \int_0^\infty dr\ rp[r] = \int_0^\infty dr(\lambda_1 p[r] + \lambda_2 rp[r] + p[r]\log p[r])$$

Entropy maximization: Single Neuron, average firing rate, and more

For given λ_1 and λ_2, the critical point of $x \log x + \lambda_1 x + \lambda_2 r x$ becomes

$$x = C * 2^{-\lambda_1 - \lambda 2r}$$

where C = $2^{\frac{1}{\ln 2}}$. Thus, in this case, the optimal $p[r]$ should be an exponential function.

• When there is an additional constraint on the second-order moment too, $p[r]$ should be an exponential function of a second-order polynomial. Thus, in this case, the optimal $p[r]$ should be a Gaussian.

Entropy maximization: Population of neurons

- Optimizing each neuron does not necessarily mean that the total population of neurons will be optimized: The relation between the neurons may decrease the overall entropy.
- Suppose there are N neurons, and each response vector $\mathbf{r} \in \mathbb{R}^N$, the probability distribution of $\mathbf{r}$ is given as $p[\mathbf{r}]$. Then, the overall entropy is given as

$$H = -\int d\mathbf{r} p[\mathbf{r}] \log(p[\mathbf{r}]) - N \log(\Delta r)$$

- The entropy of a single neuron a is given as

$$H_a = -\int d\mathbf{r} p[\mathbf{r}] \log(p[r_a]) - \log(\Delta r)$$

Entropy maximization: Population of neurons

- We can intuitively see that when the neurons are independent, the overall population will have the maximized entropy.
- Indeed,

  $$H \leq \sum_a H_a$$

  holds, as

  $$\sum_a H_a - H = \int d\mathbf{r} p[\mathbf{r}] \log(\frac{p[\mathbf{r}]}{\Pi_a p_a[r_a]}) \geq 0$$

  as it is the KL divergence between $p[\mathbf{r}]$ and $\Pi_a p_a[r_a]$.
- Equality holds iff $p[\mathbf{r}] = \Pi_a p_a[r_a]$, thus when all neurons are independent.
- Remark: $\sum_a H_a - H$ is the mutual information between N neurons. Thus, decreasing the mutual information between neurons are increasing the overall entropy!

Factorial codes

- For a population of neurons to have maximal entropy, we can see that they need two conditions.

   1) They should be independent.
   2) Each neuron should be optimal

  The encoding scheme that satisfies 1) and 2) are ”factorial codes”, as the probability $p[\mathbf{r}]$ can be factorized as a multiple of $p[r_1]$, $p[r_2]$, ..., $p[r_N]$.

- In general, finding the exact factorial code is hard. We use substitute constraints instead, such as fixing average firing rate and second moment for all neurons, or enforcing

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

  a procedure similar to whitening in signal processing.

Application to RGC receptive fields: Introduction

- For space-time receptive field $D(\mathbf{x}, t)$, the linear estimate of the response of a neuron is given as

$$L(t) = \int_0^\infty d\tau \int d\mathbf{x} D(\mathbf{x}, t) s(\mathbf{x}, t - \tau).$$

We wish to find the optimal $D(\mathbf{x}, t)$ that maximizes the overall entropy for a population of neurons. Then, we will compare the results with experiments.

- Preliminaries

1) The space-time receptive field is seperable as $D_s(\mathbf{x})$ and $D_t(\tau)$.
2) The stimulus $s(\mathbf{x}, t)$ is also separable as $s_s(\mathbf{x})$ and $s_t(t)$.

Thus, the linear estimate is separable with L_s and L_t.

Application to RGC receptive fields: Introduction

- Preliminaries (Continued...)
  3) All receptive fields within the patch we are considering are equivalent, and receptive fields with different centers can be expressed by mere translation, i.e. the cell whose receptive field is centered at **a** has the linear spatial response

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

The Whitening Filter

- It is intractable to find the optimal population of neurons. Rather, we use the approximate approach to enforce the correlation of different neurons to be 0 and the correlation of identical neurons to be σ_L^2.
- Thus, we want D_s that satisfies

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

  Remind that the stimulus is zero-averaged for trials.
- The homogeneity of stimulus implies that

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

The Whitening Filter

- Now, denote $\tilde{D_s}(\kappa)$ and $\tilde{Q_{ss}}(\kappa)$ as Fourier transforms of D_s and $Q_s s$. When we write

$$H(\mathbf{b}-\mathbf{x}) = \int d\mathbf{y} D_s(\mathbf{y}-\mathbf{b}) Q_{ss}(\mathbf{x}-\mathbf{y}) = D_s * Q_{ss},$$

  we can see that

$$\int d\mathbf{x} D_s(\mathbf{x}-\mathbf{a}) H(\mathbf{b}-\mathbf{x}) = D_s * H = Q_{LL}(\mathbf{a}-\mathbf{b}) = \sigma_L^2 \delta(\mathbf{a}-\mathbf{b}).$$

- As Fourier transform of convolution = Multiplication of Fourier transform,

$$D_s\tilde{(\kappa)}\mathcal{F}(H) = \sigma_L^2.$$

- As $\mathcal{F}(h(-t)) = \overline{\mathcal{F}(h(t))}$,

$$\mathcal{F}(H) = \overline{\tilde{D_s}(\kappa)\tilde{Q_{ss}}(\kappa)}$$

The Whitening Filter

- At last, we know that Q_{ss} is actually a function of $|\mathbf{x}-\mathbf{y}|$, thus $\tilde{Q_{ss}}$ is real. Using the fact, we obtain the final equality

  $$|\tilde{D_s}(\kappa)|^2\tilde{Q_{ss}}(\kappa) = \sigma_L^2.$$

  and

  $$|\tilde{D_s}(\kappa)| = \frac{\sigma_L}{\sqrt{\tilde{Q_{ss}}(\kappa)}}$$

- Remark: The result only tells about the magnitude of Fourier transform. Thus, there may be multiple choices of optimal spatial reception fields.
- Experiments show

  $$\tilde{Q_{ss}}(\kappa) \propto \frac{exp(-\alpha|\kappa|)}{|\kappa|^2+|\kappa_0|^2}$$

Filtering input noise

- As $\tilde{Q_{ss}}(\kappa) \propto \frac{exp(-\alpha|\kappa|)}{|\kappa|^2+|\kappa_0|^2}$ and $|\tilde{D_s}(\kappa)| = \frac{\sigma_L}{\sqrt{\tilde{Q_{ss}}(\kappa)}}$, for large frequencies, the amplitude $|\tilde{D_s}(\kappa)|$ boosts.
- This is due to the exponential attenuation of the eye when large frequency signals are inputted. Though the strategy maximizes entropy, it is doing it by ”amplifying the noise”(high-frequency region) rather than ”obtaining meaningful information” - Not a good strategy!
- To alleviate such a problem, we should have additional filtering that filters noise before the signal is inputted to the receptive field.

Filtering input noise

- Now, the final receptive field is

$$\tilde{D}_s(\kappa) = \tilde{D}_w(\kappa)\tilde{D}_\eta(\kappa),$$

where $\tilde{D}_\eta(\kappa)$ is the Fourier transform of the optimal kernel that filters noise, i.e. when input $s_s(\mathbf{x}) + \eta(\mathbf{x})$ is given, output $s_s(\mathbf{x})$.

- From optimal kernel theory,

$$\tilde{D}_\eta(\kappa) = \frac{\mathcal{F}(\langle (s_s(\mathbf{x}) + \eta(\mathbf{x}))s_s(\mathbf{y})\rangle)}{\mathcal{F}(\langle (s_s(\mathbf{x}) + \eta(\mathbf{x}))(s_s(\mathbf{y}) + \eta(\mathbf{y}))\rangle)} = \frac{\tilde{Q}_{ss}(\kappa)}{\tilde{Q}_{ss}(\kappa) + \tilde{Q}_{\eta\eta}(\kappa)}$$

- Substitute to obtain

$$|\tilde{D}_s(\kappa)| = \frac{\sigma_L\sqrt{\tilde{Q}_{ss}(\kappa)}}{\tilde{Q}_{ss}(\kappa) + \tilde{Q}_{\eta\eta}(\kappa)}$$

Temporal Processing

- The temporal receptive field is almost identical,

$$|\tilde{D}_t(\omega)| = \frac{\sigma_L \sqrt{\tilde{Q}_{ss}(\omega)}}{\tilde{Q}_{ss}(\omega) + \tilde{Q}_{\eta\eta}(\omega)}$$

- Here, experiments show that

$$\tilde{Q}_{ss}(\omega) \propto \frac{1}{\omega^2 + \omega_0^2}$$

and the experimental results match well with theory.

Temporal Processing: Experiments

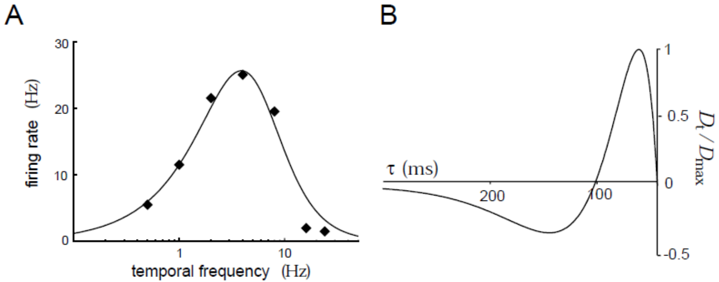
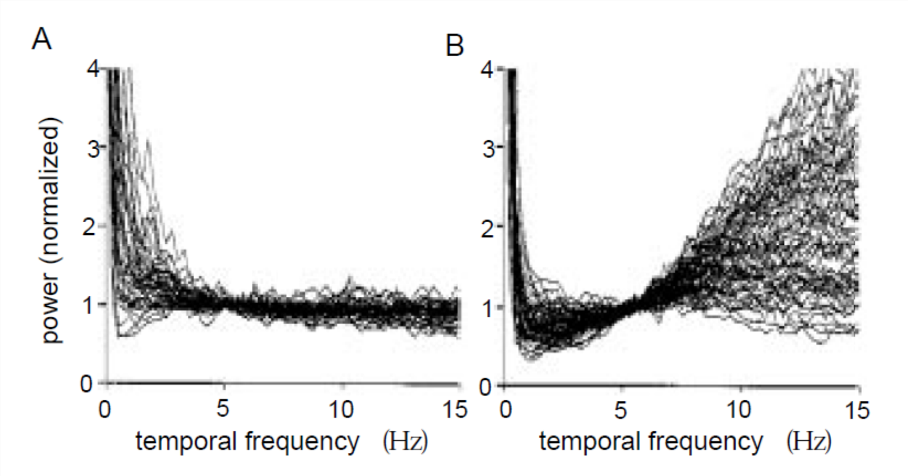


Figure 4.4: A) Predicted (curve) and actual (diamonds) selectivity of an LGN cell as a function of temporal frequency. The predicted curve is based on the optimal linear filter $\tilde{D}_t(\omega)$ with $\omega_0 = 5.5$ Hz. B) Causal, minimum phase, temporal form of the optimal filter. (Adapted from Dong and Atick, 1995; data in A from Saul and Humphrey, 1990.)

Temporal Processing: Experiments



Cortical Coding: Spatial and Temporal information are not independent

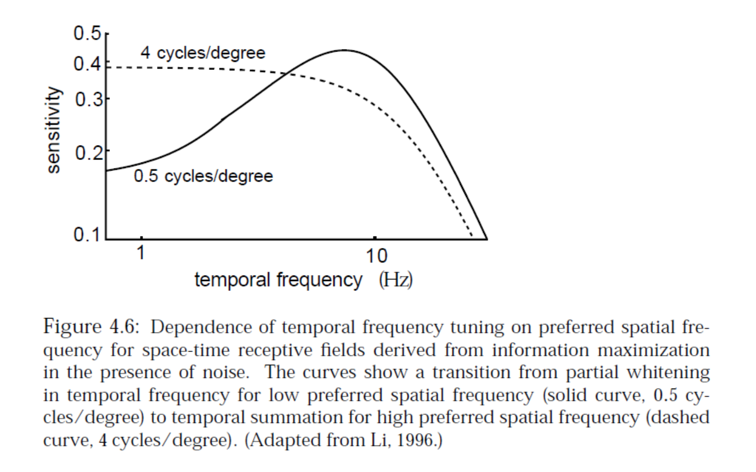


Figure 4.6: Dependence of temporal frequency tuning on preferred spatial frequency for space-time receptive fields derived from information maximization in the presence of noise. The curves show a transition from partial whitening in temporal frequency for low preferred spatial frequency (solid curve, 0.5 cycles/degree) to temporal summation for high preferred spatial frequency (dashed curve, 4 cycles/degree). (Adapted from Li, 1996.)

Entropy and Information for Spike Trains

- For spike trains we should consider the entropy ”rate” instead of the entropy itself because the amount of information increases as time passes. The entropy rate is defined as the total entropy divided by the duration of the spike train noted with $\dot{H}$.
- Consider the interspike interval t as the parameter that carries information, and $p[\tau]$ be the probability distribution function of the interspike interval. For a given time interval T there are $\langle r \rangle T$ spike intervals in expectation. As entropy is maximized when all spike intervals are independent, we can obtain the upper bound of the entropy rate

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

- When the interspike intervals follow the Poisson distribution and are independent, we can obtain the exact entropy rate.

Entropy and Information for Spike Trains: Experimental Calculation

- To obtain experimental measurements of the entropy rate for spike trains, we use a subsequence of spikes that have duration of T_s and assume each subsequences are independent.
- We divide the time interval of length T_s into $\frac{T_s}{\Delta t}$ bins, and encode a binary sequence for each spike pattern - 0 when no spikes, and 1 when spikes.
- Then, for each time interval, the entropy rate becomes

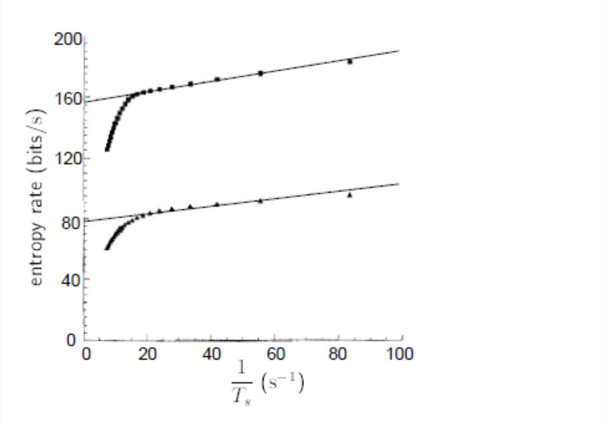
  $$\dot{H} = -\frac{1}{T_s} \sum_B P[B] \log(P[B])$$

  and by experimental results, we can calculate the corresponding value.
- The assumption that each subsequences with length T_s are independent is not true in general. Thus, we need the value of $\dot{H}$ when $T_s \to \infty$ for true entropy rate.

Entropy and Information for Spike Trains: Experimental Calculation

- Fortunately, when $T_s \to \infty$, $\frac{1}{T_s}$ and the true entropy should be proportional for large T_s. Thus, by extrapolating the curve ”Entropy rate v.s. $\frac{1}{T_s}$”, we can find the true entropy rate.
- Empirical results do not show linear dependence between entropy rate and $\frac{1}{T_s}$, as larger T_s needs more data points to calculate the entropy rate. Thus, empirical results show a two-phase relation between the entropy rate and $\frac{1}{T_s}$. Interpolating with data when T_s is not too large is meaningful.
- With the same idea, we can obtain the noise entropy rate and also the (mutual) information rate.

Entropy and Information for Spike Trains: Experiment



Conclusion

- We can define the quantitative amount of information in an encoding scheme for our brain, using Shannon's information theory.
- Our brain tries to maximize the entropy of its response (at least that is what experiments suggest).
- We can obtain the entropy rate of spike trains experimentally.
- Information theory can be used as a great tool to understand brain and how ”informative” each signal is.