Shannon and Fisher Information
Theoretical Neuroscience

Maneesh Sahani

20 February 2017

1. **Stochastic processes and entropy rates**

   (a) Prove that the two definitions of the entropy rate given in class:

   $$\lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots X_n) \qquad \text{and} \qquad \lim_{n\to\infty} H(X_n \mid X_{n-1} \ldots X_1),$$

   are equivalent. [Hint: If $a_n \to a$ as $n \to \infty$, what can be said about the running averages $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$?]

   Now consider a point process $\mathcal{P}_\lambda$ with a constant mean rate constrained to be λ. We are interested in the form of the maximum entropy process consistent with this constraint.

   (b) First, consider the stochastic process defined by taking successive inter-event intervals generated by $\mathcal{P}_\lambda$. How does the constraint on $\mathcal{P}_\lambda$'s rate constrain the ISI process? What is the maximum entropy ISI distribution [recall the discussion of maximum entropy distributions from lecture]? What does this imply about $\mathcal{P}_\lambda$?

   (c) Now consider the stochastic process defined by counting events from $\mathcal{P}_\lambda$ that fall in successive intervals of length Δ. What does the mean rate constraint for the point process mean as a constraint for this discretised counting process? What is the maximum entropy counting process under this constraint? Is this consistent with the form of $\mathcal{P}_\lambda$ you obtained above?

   (d) Suppose we were to expect spike trains in the brain to achieve maximum entropy with constrained spike rate. Which of the two preceding approaches to the obtaining the maximum entropy distribution is likely to be the more relevant to the brain. [Hint: how does the process obtained in the second case depend on Δ? Is there a preferred Δ for the brain?]

2. **Communication through a probabilistic synapse**

   (a) The Blahut-Arimoto algorithm.

   In this part of the question, we derive an algorithm to find an input distribution that achieves the capacity of an arbitrary discrete channel.

   i. Given a channel characterised by the conditional distribution $P(R|S)$, we wish to find a source distribution $P(S)$ that maximises the mutual information $I(R;S)$. Show that

   $$I(R;S) \geq \sum_{s,r} P(s)P(r|s) \log \frac{Q(s|r)}{P(s)}$$

   for any conditional distribution $Q(S|R)$. When is equality achieved?

ii. Use this result to derive (in closed form) an iterative algorithm to find the optimal $P(S)$.* This is called the Blahut-Arimoto algorithm. Prove that the algorithm converges to a unique maximal value of $I(R;S)$.

* Hint: by analogy to EM, alternate maximisations of the bound on the right hand side with respect to Q and to $P(S)$.

(b) Synaptic failure.

Many synapses in the brain appear to be unreliable; that is, they release neurotransmitter stochastically in response to incoming spikes. Here, we will build an extremely crude model of communication under these conditions.

Assume that the input to the synapse is represented by the number of spikes arriving in a 10 ms interval, while the output is the number of times a vesicle is released in the same period. Let the minimum inter-spike interval be 1 ms (taking into account both the length of the spike and the refractory period), and assume that at most 1 vesicle is released per spike. Thus, both input and output symbols on this channel are integers between 0 and 10 inclusive.

Let the probability of vesicle release be independent for each spike in the input symbol, and be given by α^n where α is a measure of synaptic depression and n is the number of spikes in the symbol. (We are neglecting order-dependent effects within each 10ms symbol, and any interactions between successive symbols. This is a terrible model of synaptic behaviour).

i. Generate (in MATLAB) the conditional distribution of output given input for this synapse. Take $\alpha = 0.9$. Use Blahut-Arimoto to derive the capacity-achieving input distribution and plot it.

ii. Try to interpret your result intuitively. Might this have anything to do with the short "bursts" of action potentials found in many spike trains?

iii. `OPTIONAL`: Improve on the model of synaptic transmission. Consider 5 ms input and output symbols, each being a 5-bit binary number where a 1 indicates a spike or a vesicle release. The probability of transmission for each spike in the symbol is again α^n but n is now the number of vesicles released *so far* for this symbol. Construct a new conditional distribution table and repeat the optimisation. Do you get a qualitatively similar result?

3. **Estimation Theory**

(a) We derived the Fisher information $J(\theta)$ as the expected value of the second derivate (curvature) of the log-likeihood in the lecture.

i. Repeat the derivation for a *vector* parameter (or stimulus in our setting) $\boldsymbol{\theta}$, showing that the Fisher information in this case is given by a matrix.

As mentioned in the lecture, there is an alternate definition in terms of the first derivative. For vector parameters this is:

$$J(\boldsymbol{\theta}_0) = \mathrm{Cov}_{\boldsymbol{\theta}_0}\left(\nabla \log p(n|\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0}\right).$$

where $\mathrm{Cov}_{\boldsymbol{\theta}_0}$ means the covariance evaluated under $p(n|\boldsymbol{\theta}_0)$.

ii. Demonstrate that these two definitions are the same (or more precisely, give conditions under which these two definitions are the same).

(b) Consider an LNP model:

$$p(n|\mathbf{x}) = \mathsf{Poiss}(g(\mathbf{w}\cdot\mathbf{x}))$$

i. What is $J(\mathbf{x})$ (the Fisher Information about the stimulus value available to the rest of the brain)? How does it depend on $\mathbf{w}$?

ii. What is $J(\mathbf{w})$ (the Fisher Information about the weight vector available to an experimenter)? How does it depend on $\mathbf{x}$?

4. **Fisher information and refractory firing**

Consider a hypothetical cell, which responds to the presentation of a stimulus with a continuous feature s by firing at a homogeneous rate $f(s)$ in a (fixed) interval $[0, T]$. Assume that the firing rate is 0 outside this interval. We will be interested in the contributions made to the Fisher information by spike-timing, with and without a refractory period.

First, assume that the firing is Poisson.

(a) What is the probability of observing spikes at times $\{t_1 \dots t_n\} \subset [0, T]$?

(b) What is the Fisher information $J_{t,\text{Poiss}}(s)$ associated with this probability density function, assuming that the relevant interval $[0, T]$ is known? How does it compare to the Fisher information $J_{n,\text{Poiss}}$ associated with the distribution of spike counts $P(n|s)$?

Now consider refractory firing. Recall that one way to model a refractory period is to use a gamma-interval renewal process in place of a Poisson process. Thus, now assume that the cell's firing follows a gamma-interval process with the same mean rate $f(s)$ and with integral gamma order γ.

(c) What is the probability of observing spikes at times $\{t_1 \dots t_n\} \subset [0, T]$ from this process?

(d) What is the Fisher information $J_{t,\text{Gamma}}(s)$ associated with the new probability density function? You may assume that T is long enough to neglect contributions due to the first spike, and due to the silence after the last spike.

Finally, we wish to see how much of this information gain is available in the spike count.

(e) Which signal (count or spike-timing) do you expect to carry more information for this process? Why?

(f) Find an expression for the the distribution of spike counts $P(n|s)$ under the gamma-interval model.

(g) Write down the expression for the corresponding Fisher information $J_{n,\text{Gamma}}$, and thus for $J_{n,\text{Gamma}} - J_{t,\text{Gamma}}$. You need not necessarily evaluate the expectation. Identify the term(s) responsible for the difference between $J_{n,\text{Gamma}}$ and $J_{t,\text{Gamma}}$.

5. **Population Coding** (Courtesy of Peter Dayan)

Shadlen and collaborators have claimed that if the activities of neurons in population codes are corrupted by *correlated* noise, then there is a sharp limit to the useful number of neurons in the population. *Prima facie* this is wrong – the stronger the correlations, the lower the entropy of the noise, and therefore the stronger the signal.

Resolve this issue for the case of additive and multiplicative noise by considering the following three models for the noisy activities r_1 and r_2 of two neurons which form a population code for a real-valued quantity x:

$$\text{a)} \quad \begin{cases} r_1^a = x + \epsilon_1 \\ r_2^a = x + \epsilon_2 \end{cases} \tag{1}$$

$$\text{b)} \quad \begin{cases} r_1^b = x(1-\delta) + \epsilon_1 \\ r_2^b = x(1+\delta) + \epsilon_2 \end{cases} \tag{2}$$

$$\text{c)} \quad \begin{cases} r_1^c = x(1-\delta)(1+\eta_1) \\ r_2^c = x(1+\delta)(1+\eta_2) \end{cases} \tag{3}$$

where $\delta \neq 0$ is known, and, ϵ and η are Gaussian, with mean 0 and covariance matrices:

$$\Sigma = \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix}$$

(a) What is the maximum likelihood estimator (MLE) for x on the basis of r_1 and r_2 in each case?

(b) What is the appropriate measure of the expected accuracy of the MLE, and why?

(c) How does the expected accuracy in each case depend on the degree of correlation c?

(d) What conclusions would you draw about the clash between Shadlen and common sense?