Assignment 6
Theoretical Neuroscience

Liam Paninski, Maneesh Sahani

15 November 2004

1. **Decision theory**

   (a) Prove that the optimal decision function $q(data)$ given in class exists and is unique.

   (b) Suppose we observe two cells whose firing rates are given by (λ_1, λ_2) when a stimulus is present (i.e., $\theta = 1$) and (λ_0, λ_0) when it is not. Assume that the firing can be modelled by a homogeneous Poisson process. Compute the optimal discrimination threshold T for squared loss, as a function of the prior probability $p(\theta = 1) = a$.

   (c) The Bayes-optimal estimator under squared loss is the conditional mean; what is the optimal estimator under absolute (L1) loss $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$? Is this estimator always unique?

   (d) We introduced the idea of sufficient statistics for a single distribution in class. Generalize this to the case of decision theory, where there are two distributions to consider. What are the sufficient statistics for distinguishing two multidimensional Gaussians with identical covariance? With different covariances?

   (e) (Optional: estimation theory). Consider data X generated from a parameterized distribution $P(X; \theta)$. Let $\hat{\theta}_0(X)$ be an estimator for the parameter θ, and $T(X)$ be a sufficient statistic for θ. A Rao-Blackwell estimator can be derived from $\hat{\theta}_0(X)$ according to the following definition:

   $$\hat{\theta}_{RB}(X) = E[\hat{\theta}_0(X) \mid T(X)].$$

   Recall that since $\theta \to T \to X \to \hat{\theta}_0$, $\hat{\theta}(X) \perp \theta \mid T(X)$. This means that the expectation above can be obtained without knowing θ, and the Rao-Blackwell estimator is observable.
   Use Jensen's inequality to prove the Rao-Blackwell theorem:

   $$E[L(\theta, \hat{\theta}_{RB}(X))] \leq E[L(\theta, \hat{\theta}(X))]$$

   where L is any convex loss function.

2. **Asymptotics**

   (a) Derive the distribution of the minimizer of the random function

   $$f(\vec{x}) = \vec{x}^t A \vec{x} + \vec{b} \cdot \vec{x} + c,$$

   where A and c are a constant matrix and scalar, respectively, and $\vec{b}$ is Gaussian with mean zero and covariance C. What does this have to do with the asymptotic distribution of the MLE? (Hint:

what if A is taken to be the Fisher information matrix? What if $A = C$?) What does it have to do with the asymptotic distribution of the minimizer of the random function

$$M_N(\theta) = \sum_{i=1}^{N} f(x_i, \theta),$$

with $f(s,t) = (s-t)^2$? Can you think of an estimation problem this loss function $M(\theta)$ might correspond to? (Hint: what if $f(x_i, \theta)$ is the loglikelihood of x_i?)

(b) We derived the Fisher information $J(\theta)$ as a geometric (curvature) quantity in the lecture. There is an alternate definition (which we actually used implicitly in computing the asymptotic covariance of the MLE):

$$J(\theta_0) = \mathrm{Cov}_{\theta_0}\left(\nabla \log p(n|\theta)\Big|_{\theta_0} \right).$$

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

(c) i. Compute the Fisher information in the parameter p for the multinomial distribution. Specifically, we are given N data points $\{x_i\}_{1 \le i \le N}$, drawn i.i.d. from some discrete distribution $\{p(j)\}_{1 \le j \le m}$ on m points. What is the MLE for p here? What is the covariance distribution of the MLE, for any fixed N? How does this finite-N covariance compare to the Fisher information $J(p)$?

ii. Discuss the shape of this function $J(p)$ as p varies in the binomial case, $m = 2$. What happens as $p \to 0$ or $p \to 1$? What are the implications?

iii. (Optional) Compute the Fisher information $J(\vec{k})$ for the LN model $p(n|\vec{x}) = \mathit{Poiss}(f(\vec{k} \cdot \vec{x}))$.

3. **More fly data**

Go back to the best model you fit to the fly data in the file c1p8.mat. Compute (by Monte Carlo, if necessary) the mean-square Bayes optimal estimator of the velocity $x(t)$ given the spike train. Compute an estimator which minimizes the ϵ-insensitive loss function

$$E_\epsilon(s,t) = \begin{cases} 0 & |s-t| < \epsilon, \\ |s-t| - \epsilon & |s-t| \ge \epsilon. \end{cases}$$

Is this estimator unique? Compare the estimated velocity to the true velocity by plotting the two signals against each other and as functions of time. Do the two estimators behave significantly differently (i.e., lead to significantly different estimates)? Can you give a reasonable explanation of any observed differences?

Compare these estimators to the optimal linear estimator (OLE), with or without regularization. (It might be easier to fit the OLE in the Fourier domain, due to the temporal stationarity of $x(t)$.) Which estimator achieves a lower mean square error? Discuss the qualitative differences (if any) between the Bayesian MMSE and linear estimators.

4. **Information theory**

(a) Use Jensen's inequality to prove that the KL divergence $\mathsf{KL}[p\|q]$ is convex in the pair (p, q); i.e., if (p_1, q_1) and (p_2, q_2) are two pairs of distributions and a is a scalar, then

$$\mathsf{KL}[ap_1 + (1-a)p_2 \| aq_1 + (1-a)q_2] \le a\mathsf{KL}[p_1\|q_1] + (1-a)\mathsf{KL}[p_2\|q_2].$$

(b) Differential entropy. What is the differental entropy of the continuous distribution with density $u_a(x)$ that is uniform on the interval $[-a/2, a/2]$ and 0 elsewhere? When is this entropy equal to 0? Compare to the case of the entropy of a discrete distribution being 0. What happens when $a = 0.5$? What happens as $a \to \infty$? Do these results correspond to your intuitive notion of uncertainty in a distribution?

(c) Is the KL divergence defined on continuous distributions "well-behaved"? Calculate $KL[u_1 \| u_2]$ as well as $KL[u_2 \| u_1]$ (u_a as defined above). Interpret these results in terms of the "coding penalty" discussed in class.

5. **Stochastic processes and entropy rates**

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

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

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

(b) 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 the constraint.

i. 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 process? What does this imply about $\mathcal{P}_\lambda$?

ii. Now consider the stochastic process defined by counting events from $\mathcal{P}_\lambda$ that fall in successive intervals of length Δ. How is the mean rate constraint reflected in this counting process? What is the maximum entropy counting process under this constraint? What does this imply about $\mathcal{P}_\lambda$?

iii. 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. [Hint: how does the process obtained in the second case depend on Δ?]