UC Berkeley
Department of Electrical Engineering and Computer Sciences

EECS 126: Probability and Random Processes

Homework 07
Fall 2023

1. **Entropy Maximization by Gaussians**

   For a continuous random variable X with density f, we define its *differential entropy* as

   $$h(f) := -\mathbb{E}(\log f(X)) = -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx.$$

   Note that differential entropy is translation-invariant. For a Gaussian with variance σ^2, we have $h(f) = \frac{1}{2}\log(2\pi e\sigma^2)$. Then the *relative entropy*, or Kullback–Leibler divergence, between two continuous distributions f and g is

   $$D(f \parallel g) = \mathbb{E}_{X\sim f}\left(\log \frac{f(X)}{g(X)}\right) = \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)}\, dx.$$

   a. Show that $D(f \parallel g) \geq 0$, with equality iff $f \equiv g$, i.e. $f(x) = g(x)$ for all x. *Hint*: if φ is strictly concave, Jensen's inequality states that $\varphi(\mathbb{E}(Z)) \geq \mathbb{E}(\varphi(Z))$, with equality iff Z is constant.

   Remark: by this result, it is often useful to think about $D(\cdot \parallel \cdot)$ as a sort of distance function, though it is asymmetric. A genuine information-theoretic metric is the variation of information $VI(X;Y) = H(X,Y) - I(X;Y)$.

   b. Let g be a Gaussian PDF with variance σ^2, and let f be an arbitrary PDF with the same variance. Show that differential entropy is maximized by taking $f \equiv g$.

2. **Mutual Information for Markov Chain**

In the discussion, we stated without proof the fact that $H(X \mid Y) \leq H(X \mid \hat{X})$, where $\hat{X} = g(Y)$. Here, we will explore why this inequality is true. We define the *conditional mutual information* between random variables X and Y given Z to be

$$I(X;Y \mid Z) \coloneqq \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y \mid z)}{p(x \mid z)\, p(y \mid z)}.$$

a. Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain. Show that $I(X_{n-1}; X_{n+1} \mid X_n) = 0$ for any $n \geq 1$.

b. Give an interpretation of part a.

c. Show that $I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z)$. Returning to the setting of Homework 06 Q4, conclude that $H(X \mid Y) \leq H(X \mid \hat{X})$.

Hint: Show that $I(X; \hat{X} \mid Y) = 0$ using part a.

3. **Relative Entropy and Stationary Distributions**

The *relative entropy*, or Kullback–Leibler divergence, between two distributions p and q is defined as the following. Note that this definition is not symmetric.

$$D(p \parallel q) = \mathbb{E}_{X \sim p}\left(\log \frac{p(X)}{q(X)}\right) = \sum_x p(x) \log \frac{p(x)}{q(x)}.$$

a. Show that $D(p \parallel q) \geq 0$, with equality if and only if $p(x) = q(x)$ for all x. *Hint*: if φ is strictly concave, Jensen's inequality states that $\varphi(\mathbb{E}(Z)) \geq \mathbb{E}(\varphi(Z))$, with equality if and only if Z is constant.

Remark: by this result, it is often useful to think about $D(\cdot \parallel \cdot)$ as a sort of distance function, though it does not satisfy symmetry or the triangle inequality. Instead, $D(\cdot \parallel \cdot)$ is a type of *divergence* function. A genuine information-theoretic metric is the variation of information $VI(X;Y) = H(X,Y) - I(X;Y)$.

b. Show that for any irreducible Markov chain with stationary distribution π, any other stationary distribution μ must be equal to π. *Hint*: consider $D(\pi \parallel \mu P)$.

4. **Markov Chain Practice**

Consider a Markov chain with three states 0, 1, 2, and suppose its transition probabilities are $P(0,1) = P(0,2) = \frac{1}{2}$, $P(1,0) = P(1,1) = \frac{1}{2}$, $P(2,0) = \frac{2}{3}$, and $P(2,2) = \frac{1}{3}$.

a. Classify the states in the chain. Is this chain periodic or aperiodic?

b. In the long run, what fraction of time does the chain spend in state 1?

c. Suppose that X_0 is chosen according to the steady-state or stationary distribution. What is $\mathbb{P}(X_0 = 0 \mid X_2 = 2)$?

5. **Two-State Chain with Linear Algebra**

Consider the Markov chain $(X_n,\ n \in \mathbb{N})$, shown in Figure 1, where $\alpha, \beta \in (0,1)$.

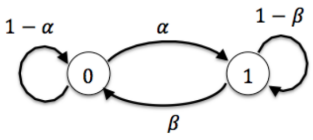

Figure 1: Markov chain for this Problem

a. Find the probability transition matrix P.

b. Find two real numbers λ_1 and λ_2 such that there exists two non-zero vectors u_1 and u_2 such that $Pu_i = \lambda_i u_i$ for $i = 1, 2$. Further, show that P can be written as $P = U\Lambda U^{-1}$, where U and Λ are 2×2 matrices and Λ is a diagonal matrix.

Hint: This is called the eigendecomposition of a matrix.

c. Find P^n in terms of U and Λ for each $n \in \mathbb{N}$.

d. Assume that $X_0 = 0$. Use the result in part (c) to compute the PMF of X_n for all $n \in \mathbb{N}$.

e. What does the fraction of time spent in state 0, $n^{-1}\sum_{i=1}^{n} \mathbb{1}\{X_i = 0\}$, converge to (almost surely) as $n \to \infty$?

6. **Metropolis–Hastings**

We will prove some properties of the *Metropolis–Hastings* algorithm, an example of Markov Chain Monte Carlo (MCMC) sampling that you will see more of in lab. The goal of MH is to draw samples from a distribution $p(x)$; the algorithm assumes that

- We can compute $p(x)$ up to a normalizing constant C via $f(x)$, and
- We have a proposal distribution $g(x, \cdot)$.

The steps in making a transition are:

i. Propose the next state y according to the distribution $g(x, \cdot)$.

ii. Accept the proposal with probability

$$A(x, y) = \min\left\{1, \frac{f(y)}{f(x)}\frac{g(y, x)}{g(x, y)}\right\}.$$

iii. If the proposal is accepted, move the chain to y; otherwise, stay at x.

Remark. The normalizing factor $C = 1/\sum_{x\in\mathcal{X}} f(x)$ is sometimes called the *partition function*, and it can be difficult to compute for large sets $\mathcal{X}$, even if $f(x)$ is efficient to compute.

In the following, we will verify that the Metropolis–Hastings chain has stationary distribution p, and in fact approaches stationarity after running for some time, at which point we can draw samples from p by sampling from the chain.

a. The key to why Metropolis–Hastings works is the **detailed balance equations**. Suppose we have a finite irreducible Markov chain on a state space $\mathcal{X}$ with transition probability matrix P. Show that if there exists a distribution π on $\mathcal{X}$ satisfying detailed balance,

$$\pi(x)P(x, y) = \pi(y)P(y, x) \quad \text{for all } x, y \in \mathcal{X},$$

then $\pi P = \pi$ is a stationary distribution of the chain.

b. Returning to the Metropolis–Hastings chain, find $P(x, y)$. For simplicity, assume $x \neq y$.

c. Show that the target distribution $p(x)$ satisfies the detailed balance equations for $P(x, y)$, and conclude that $p(x)$ is the stationary distribution of the chain.

d. If the chain is aperiodic, then it will converge to the stationary distribution. If not, we can force the chain to be aperiodic by considering the **lazy chain**: on each transition, the chain decides not to move with probability $\frac{1}{2}$, independently of the propose-accept step. Explain why the lazy chain is aperiodic, and explain why the stationary distribution is the same as before.