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 $f \equiv g$.

Solution:

a. As in the proof that mutual information is nonnegative, we note that $-\log(\cdot)$ is strictly convex, so we can apply Jensen's inequality with $Z = \frac{f(X)}{g(X)}$:

$$-D(f \parallel g) = \int f(x) \log \frac{g(x)}{f(x)}\, dx \leq \log \int f(x) \frac{g(x)}{f(x)}\, dx = \log \int g(x)\, dx = 0.$$

Furthermore, we have equality if and only if $\frac{g(x)}{f(x)} = c$ for all x. But, as both are probability densities, we must have $c = 1$, so $f \equiv g$ holds whenever $D(f \parallel g) = 0$.

b. As differential entropy is translation-invariant, assume without loss of generality that f and g are zero-mean.

$$0 \leq D(f \parallel g) = \int f(x) \log \frac{f(x)}{g(x)}\, dx = -h(f) - \int f(x) \log g(x)\, dx.$$

We compute the second term to be

$$\int f(x) \log g(x)\, dx = \int f(x) \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/(2\sigma^2)}\right) dx$$

$$
\begin{aligned}
&= \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\int f(x)\ dx - \log(e)\int f(x)\frac{x^2}{2\sigma^2}\ dx \\
&= -\frac{1}{2}\log(2\pi\sigma^2) - \log(e)\frac{\sigma^2}{2\sigma^2} \\
&= -\frac{1}{2}\log(2\pi e\sigma^2) = -h(g).
\end{aligned}
$$

Therefore $h(g) - h(f) \geq 0$, with equality if and only if $D(f \parallel g) = 0$, i.e. $f \equiv g$.

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.

   Solution:

   a. By the Markov property, $X = X_{n-1}$ and $Y = X_{n+1}$ are conditionally independent given $Z = X_n$. That is, $p(X \mid Z) \cdot p(Y \mid Z) = p(X, Y \mid Z)$. Then

$$I(X;Y \mid Z) = \mathbb{E}\left(\log \frac{p(X,Y \mid Z)}{p(X \mid Z)\, p(Y \mid Z)}\right) = \mathbb{E}(\log 1) = 0.$$

   b. Given the current state of a Markov chain, no information can be gained about the past by observing the future, and vice versa.

   c. As we have seen, by the linearity of expectation,

$$\begin{aligned} I(X;Y \mid Z) &= \mathbb{E}\left(\log \frac{p(X,Y \mid Z)}{p(X \mid Z)\, p(Y \mid Z)}\right) \\ &= \mathbb{E}(-\log p(X \mid Z)) + \mathbb{E}(\log p(X \mid Y, Z)) \\ &= H(X \mid Z) - H(X \mid Y, Z). \end{aligned}$$

   Now, by part a, because X and $\hat{X} = g(Y)$ are conditionally independent given Y, we have $I(X; \hat{X} \mid Y) = 0$, or $H(X \mid Y) = H(X \mid \hat{X}, Y)$, which also equals $H(X \mid \hat{X}) - I(X; Y \mid \hat{X})$. Conditional mutual information is nonnegative by Jensen's inequality, and therefore $H(X \mid Y) \leq H(X \mid \hat{X})$.

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 $D(\pi \parallel \mu P)$.

Solution:

a. As in the proof that mutual information is nonnegative, we note that $-\log(\cdot)$ is strictly convex, so we can apply Jensen's inequality with $Z = \frac{p(X)}{q(X)}$:

$$-D(p \parallel q) = \sum_x p(x)\log\frac{q(x)}{p(x)} \leq \log\sum_x p(x)\frac{q(x)}{p(x)} = \log\sum_x q(x) = \log 1 = 0.$$

Furthermore, we have equality if and only if $\frac{q(x)}{p(x)} = c$ for all x. But, as both are probability distributions, we must have $c = 1$, so $p \equiv q$ holds whenever $D(p \parallel q) = 0$.

b. Let P be the transition matrix of the Markov chain. Then

$$\begin{aligned}
D(\pi \parallel \mu P) &= \sum_y \pi(y)\log\left(\frac{\pi(y)}{(\mu P)(y)}\right) \\
&= -\sum_y \log\left(\sum_x \frac{\mu(x)P(x,y)}{\pi(y)}\right)\pi(y) \\
&= \sum_y\left[-\log\left(\sum_x \frac{\mu(x)}{\pi(x)}\frac{\pi(x)P(x,y)}{\pi(y)}\right)\right]\pi(y)
\end{aligned}$$

Now, we observe that $\nu(x) = \frac{\pi(x)P(x,y)}{\pi(y)}$ is a probability distribution, so the inner term is $-\log\mathbb{E}_{x\sim\nu}(\frac{\mu(x)}{\pi(x)})$. We can now apply Jensen's inequality:

$$\begin{aligned}
&\leq \sum_y\left[\sum_x -\log\left(\frac{\mu(x)}{\pi(x)}\right)\frac{\pi(x)P(x,y)}{\pi(y)}\right]\pi(y) \\
&= \sum_x -\log\left(\frac{\mu(x)}{\pi(x)}\right)\left[\sum_y \pi(x)P(x,y)\right] = D(\pi \parallel \mu).
\end{aligned}$$

Intuitively, this says that applying P can only bring the distribution closer to stationarity, at least in terms of relative entropy. Furthermore, we have equality if and only if $\frac{\mu(x)}{\pi(x)}$ is constant. By the same reasoning as in part (a), we must have $\mu \equiv \pi$.

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 $\mathbb{P}(X_0 = 0 \mid X_2 = 2)$?

Solution:

a. The Markov chain is one recurrent, aperiodic class.

b. By solving $\pi P = \pi$, we have

$$\pi = \frac{1}{11} \begin{bmatrix} 4 & 4 & 3 \end{bmatrix}.$$

Thus $\pi(1) = 4/11$.

c. By the definition of conditional probability,

$$\mathbb{P}(X_0 = 0 \mid X_2 = 2) = \frac{\mathbb{P}(X_0 = 0, X_2 = 2)}{\mathbb{P}(X_2 = 2)} = \frac{\mathbb{P}(X_0 = 0, X_1 = 2, X_2 = 2)}{\mathbb{P}(X_2 = 2)}.$$

Note that we used the fact that the only possible two-step path from $X_0 = 0$ to $X_2 = 2$ in this chain is $0 \to 2 \to 2$. Now, $\mathbb{P}(X_2 = 2) = \mathbb{P}(X_0 = 2)$ because X_0 is chosen according to the stationary distribution π, so

$$\frac{\mathbb{P}(X_0 = 0, X_1 = 2, X_2 = 2)}{\mathbb{P}(X_2 = 2)} = \frac{\pi(0) \cdot (1/2) \cdot (1/3)}{\pi(2)} = \frac{2}{9}.$$

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

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

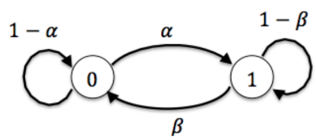

Figure 1: Markov chain for this Problem

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 $n \to \infty$?

Solution:

a. The probability transition matrix is

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}.$$

b. Since $(P - \lambda_i I)x = 0$ has non-zero solution u_i, we have $\det(P - \lambda_i I) = 0$, i.e., λ_1 and λ_2 are solutions to

$$\det \begin{bmatrix} 1-\alpha-\lambda & \alpha \\ \beta & 1-\beta-\lambda \end{bmatrix} = \lambda^2 - (2-\alpha-\beta)\lambda + 1 - \alpha - \beta.$$

Then we get $\lambda_1 = 1$, and $\lambda_2 = 1 - \alpha - \beta$. Then we can get u_1 and u_2: $u_1 = \begin{bmatrix} 1 & 1 \end{bmatrix}^\mathsf{T}$ and $u_2 = \begin{bmatrix} \alpha & -\beta \end{bmatrix}^\mathsf{T}$. Further, we can see that if we let

$$U = \begin{bmatrix} u_1 & u_2 \end{bmatrix} = \begin{bmatrix} 1 & \alpha \\ 1 & -\beta \end{bmatrix},$$

and

$$\Lambda = \begin{bmatrix} 1 & 0 \\ 0 & 1-\alpha-\beta \end{bmatrix},$$

we have $PU = U\Lambda$, which is equivalent to $P = U\Lambda U^{-1}$.

c. We have

$$P^n = U\Lambda U^{-1} \cdots U\Lambda U^{-1} = U\Lambda^n U^{-1}.$$

d. Let $\pi(n) = \begin{bmatrix}\Pr(X_n = 0) & \Pr(X_n = 1)\end{bmatrix}$ be the PMF of X_n. Then we have

$$\pi(n) = \pi(0)P^n = \pi(0)U\Lambda^n U^{-1}.$$

Since we have $\pi(0) = \begin{bmatrix}1 & 0\end{bmatrix}$, by some computation, we get

$$\pi(n) = \frac{1}{\alpha+\beta}\begin{bmatrix}\beta + \alpha(1-\alpha-\beta)^n & \alpha - \alpha(1-\alpha-\beta)^n\end{bmatrix}.$$

e. By the Big Theorem, the fraction of time spent in state 0 converges to the stationary distribution at state 0, $\pi(0)$. The stationary distribution is

$$\pi = \frac{1}{\alpha+\beta}\begin{bmatrix}\beta & \alpha\end{bmatrix},$$

so $\pi(0) = \beta/(\alpha+\beta)$.

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.

Solution:

a. Suppose that detailed balance holds. Then for all $y \in \mathcal{X}$,

$$(\pi P)(y) = \sum_{x\in\mathcal{X}} \pi(x)P(x, y) = \sum_{x\in\mathcal{X}} \pi(y)P(y, x) = \pi(y) \sum_{x\in\mathcal{X}} P(y, x) = \pi(y).$$

b. $P(x,y)$ is the probability that we propose y with $g(x,\cdot)$, then accept y:

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

c. We check that detailed balance holds for any pair of states (x,y). Observe that if

$$\frac{f(y)}{f(x)}\frac{g(y,x)}{g(x,y)} \le 1,$$

then $A(x,y)$ is equal to this ratio, and its reciprocal is at least 1, which makes $A(y,x) = 1$. Thus, assume without loss of generality that $A(y,x) = 1$, swapping x and y if this were not true. Then $P(y,x) = g(y,x)$, and

$$\begin{aligned}
p(x)P(x,y) &= p(x)g(x,y)A(x,y) \\
&= p(x)g(x,y)\frac{f(y)g(y,x)}{f(x)g(x,y)} \\
&= p(x)\frac{f(y)}{f(x)}g(y,x) \\
&= p(y)g(y,x) \\
&= p(y)P(y,x).
\end{aligned}$$

Note that $p(x)\frac{f(y)}{f(x)} = p(y)$ follows from the fact that f is directly proportional to p.

d. The lazy chain is aperiodic as it has self-loops. Now, suppose $\pi = \pi P$ is a stationary distribution of the original chain. The transition probability matrix P' of the lazy chain is $\frac{1}{2}P + \frac{1}{2}I$, where I is the identity matrix, so

$$\pi P' = \frac{1}{2}\pi P + \frac{1}{2}\pi I = \frac{1}{2}\pi + \frac{1}{2}\pi = \pi.$$

In other words, π is also a stationary distribution for the lazy chain.