APMA 1740/2610 2025: Homework 2

1. **Strange outcome, fair die.** Let $X_{1:n}$ be the outcomes of iid rolls of a fair six-sided die with faces labeled $\{1, \ldots, 6\}$. Define the empirical distribution $\widehat{p} \triangleq \left(\widehat{p}_1(X_{1:n}), \ldots, \widehat{p}_6(X_{1:n})\right)$ by

$$\widehat{p}_x \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i = x}$$

When n is large the average outcome will typically be about 3.5 because of the LLN, i.e,

$$\bar{X} \triangleq \frac{1}{n} \sum_{i=1}^{n} X_i = \sum_{x=1}^{6} x \widehat{p}_x \approx \sum_{x=1}^{6} x \frac{1}{6} = 3.5$$

Suppose, however, that you observe the rare event

$$3.0 < \bar{X} < 3.2 \tag{1}$$

(a) Use large deviations theory to guess the value of $\widehat{p}$, i.e. guess the empirical distribution conditioned on observing a rare sample mean in the sense of equation (1). Let's call your guess p^*. Assume that n is very large. Your answer should be a table of 6 numbers. Hint: (i) There are two constraints here (defined by the two inequalities in (1)), but only one can be active. (ii) Once you decide on an active constraint, say $\mathbb{E}[\mathcal{E}(X)] = \theta$, you have to determine the parameter λ in the Gibbs representation of p^*. Do this by gradient descent on the log-partition function (actually, on $\log Z_\lambda - \lambda\theta$) as described in class. Check your λ to make sure that the constraints are (approximately) satisfied.

Let $B = \{q : \mathbb{E}_q[X] \in (3.0, 3.2)\}$

By the LDP, we are interested in

$$p^* = \arg\min_{q \in B} D(q \parallel \hat{p})$$

subject to the constraints

- $\mathbb{E}[\mathcal{E}(X)] < 3.2$
- $\mathbb{E}[\mathcal{E}(X)] > 3.0$

Let us begin with the active constraint $\mathbb{E}[\mathcal{E}(X)] = 3.2$ since $\bar{X} = 3.5 > 3.2$ and Sanov implies the p^* will occur on the boundary.

From previous work, we know the solution to the minimization problem is given by

$$p_x^* = \frac{1}{Z_\lambda} \hat{p}_x e^{\lambda x}$$

where $Z_\lambda = \sum_{x=1}^{6} \hat{p}_x e^{\lambda x}$.

We need now to solve for λ. Luckily, we know that

$$\frac{\partial}{\partial \lambda} = \log Z_\lambda - 3.2\lambda = \mathbb{E}_\lambda[X] - 3.2 = -3.2 + \sum_{x=1}^{6} \frac{x \cdot p_x^{\lambda x}}{Z_\lambda}$$

is convex in λ and minimized when $\mathbb{E}_\lambda[\mathcal{E}(X)] = 3.2$, i.e. at exactly the λ we need to find.

Let $\lambda^{(0)} = \frac{1}{6}$ and $\varepsilon = 10^{-6}$. We will use gradient descent to find the optimal λ by:

$$\lambda^{(t+1)} = \lambda^{(t)} - (\mathbb{E}_\lambda[X] - \theta)$$

until $\left|\lambda^{(t+1)} - \lambda^{(t)}\right| < \varepsilon$.

Running the algorithm, we find

$$\lambda \approx -0.1031$$

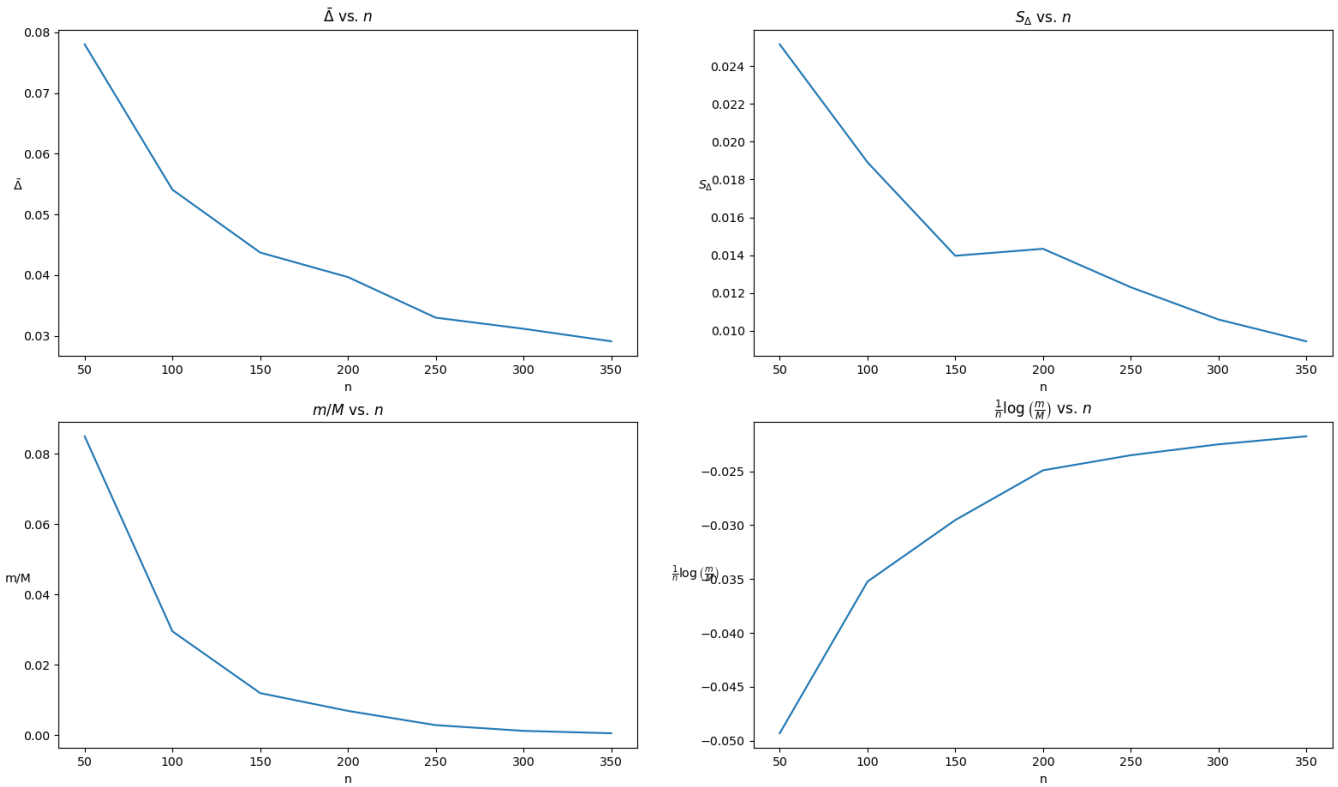
and

$$p_1^* \approx 0.217$$
$$p_2^* \approx 0.192$$
$$p_3^* \approx 0.173$$
$$p_4^* \approx 0.156$$
$$p_5^* \approx 0.140$$
$$p_6^* \approx 0.127$$

(b) (Do not submit) Perform a series of Monte Carlo computer experiments: For each of $n = 50, 100, 150, \ldots, 400$, repeatedly sample n rolls of a fair die.[1] For each sample of size n, check for the rare event from equation (1). Accumulate $m = 100$ examples of this event for each value of n. Let M be the number of samples needed to see m events,[2] so that m/M is an estimate of the probability of the rare event. Let $\widehat{p}^{(1)}, \ldots, \widehat{p}^{(m)}$ be the observed empirical distributions for each of the m rare events and let $\Delta_1, \ldots, \Delta_m$ be the corresponding distances from the empirical distributions to your prediction p^* above, where the distance is defined as

$$\Delta_k \triangleq \max_{1 \le x \le 6} \left| p_x^* - \widehat{p}_x^{(k)} \right| \qquad k = 1, \ldots, m$$

Let $\bar{\Delta}$ and S_Δ be the empirical mean and standard deviation of the Δ_k's, respectively.
Plot m/M, $\frac{1}{n}\log(m/M)$, $\bar{\Delta}$, and S_Δ versus n. Interpret the four plots.



(c) For each of the four plots, as $n \to \infty$, what is the limiting value of each quantity? Why?
All four plots go to 0 as $n \to \infty$. As n increases, LLN says we should see the rare event less and less frequently, so m/M should go to 0. The LLN also says that the empirical distribution should converge to the true distribution, so $\bar{\Delta}$ and S_Δ should go to 0 as well. We expect $\log(m/M) \to -\infty$ but $1/n \to 0$ more quickly and we take the convention $0 \log 0 = 0$.

[1] An easy way to do this in Matlab is with the command `randi(6,n,1)`.

[2] You can get a rough idea of how big M will need to be from the CLT, which says that $\bar{X}$ is approximately Normal with mean 3.5 and variance σ^2/n, where σ^2 is the variance of a single X_i.

2. (Do not submit) **Strange outcome, weighted die.** Repeat problem 1 using an *un*-fair die with $h = (h_1, \ldots, h_6) = (.1, .1, .2, .1, .2, .3)$ and with the rare event defined by *both*

$$3.6 < \bar{X} < 3.8$$

and

$$U \triangleq \frac{1}{n} \sum_{i=1}^{n} X_i^2 < 17$$

Note that both of these are unusual for large n, since we expect $\bar{X} \approx 4.1$ and $U \approx 19.7$ from the LLN. For part (a) you will need to determine which constraints are active. Try each individually and see if the other is satisfied. If neither works individually, you will need to use both simultaneously. For part (b) only use $n = 50, 100, 150, \ldots, 300$.[3]

[3] An easy way to simulate this particular unfair die in Matlab is to first define the table `lookup = [1 2 3 3 4 5 5 6 6 6]` and then get n rolls with `lookup(randi(10,n,1))`.

3. **Exponential families.** The exponential families first came up as solutions to the maximum entropy problem under a set of c constraints

$$\mathbb{E}_p[\mathcal{E}_k(X)] = \theta_k \qquad k = 1, 2, \ldots, c \tag{2}$$

We called these constraints *linear*, since they are linear in the components of p^*, as can be seen by writing out, explicitly, the expectation in (2):

$$\sum_{x=1}^{s} p^*(x)\mathcal{E}_k(x) = \theta_k \qquad k = 1, 2, \ldots, c$$

Solving the optimization problem led us to exponential families of the form

$$p^*(x) = \frac{1}{Z_{\lambda_1,\ldots,\lambda_c}} e^{\sum_{k=1}^{c} \lambda_k \mathcal{E}_k(x)} \qquad \forall x \in \{1, \ldots, s\}$$

The more general LDP problem (minimize $D(q||p)$ given a pmf p on $\{1, \ldots, s\}$), under these very same constraints, led us instead to a more general collection of exponential families, namely all distributions on $\{1, \ldots, s\}$ of the form

$$p^*(x) = \frac{p(x)}{Z_{\lambda_1,\ldots,\lambda_c}} e^{\sum_{k=1}^{c} \lambda_k \mathcal{E}_k(x)} \qquad \forall x \in \{1, \ldots, s\}$$

These derivations were in the context of discrete ($X \in \{1, \ldots, s\}$) and univariate ($X \in \mathbb{R}^1$) random variables. But the approach is more general, in two important ways: (i) we could just as well have started with continuous random variables that had density functions instead of probability mass functions (see, for example, problem 5 on this HW), and (ii) we could also have started with multivariate random variables, also known as random vectors, in d dimensions ($X \in \mathbb{R}^d$).

Let X be a continuous or discrete random variable with a pdf or pmf $f(x) = f(x; \vec{\lambda})$, where $\vec{\lambda} = \{\lambda_1, \ldots, \lambda_c\}$ is a collection of parameters that define a family of distributions. This is an "exponential family" with "canonical" or "natural" parametrization if f can be written in the form

$$f(x; \vec{\lambda}) = \frac{1}{Z(\vec{\lambda})} p(x) e^{\vec{\lambda} \cdot \vec{\mathcal{E}}(x)} \tag{3}$$

where $p(x)$ is any function that is non-negative on the range of X, and $\vec{\mathcal{E}}(x) = (\mathcal{E}_1(x), \ldots, = \mathcal{E}_c(x))$ are the "sufficient statistics".[4] (As we shall see in class, there is indeed something canonical about the canonical parametrization. Namely, that the normalizing constant, i.e. the "partition function" $Z(\vec{\lambda})$, becomes the key to computing the values of the parameters, whether we are given samples and want to do maximum-likelihood estimation or we want to compute an LDP distribution from the θ parameters.) Nothing about this representation is unique—the same family can be written with different parameters and different sufficient statistics, and p and Z can be multiplied by a common arbitrary constant, since this will not change the ratio. As for the generalization to multivariate X, instead of mapping $\mathbb{R}^1$ to $\mathbb{R}^1$, f, p, and the sufficient statistics $\mathcal{E}_1, \ldots, \mathcal{E}_c$ map $\mathbb{R}^n$ to $\mathbb{R}^1$.

The problem is to determine which of the following standard statistical families is an exponential family. For any that are, identify the sufficient statistics.

(a) Binomial$(100, p)$ with parameter $p \in [0, 1]$. (Hint: start by writing f as

$$f(x) = \binom{100}{x} p^x (1-p)^{100-x} \mathbb{1}_{x \in \{0,1,\ldots,100\}}$$

and note that this one-parameter family will have only one sufficient statistic, and the associated λ will be a function of p: $\lambda = \lambda(p)$.)

$$\begin{aligned} f(x) &= \binom{100}{x} p^x (1-p)^{100-x} \mathbb{1}_{x \in \{0,1,\ldots,100\}} \\ &= \binom{100}{x} \left(\frac{p}{1-p}\right)^x (1-p)^{100} \mathbb{1}_{x \in \{0,1,\ldots,100\}} \\ &= \binom{100}{x} e^{x \log \frac{p}{1-p} + 100 \log(1-p)} \mathbb{1}_{x \in \{0,1,\ldots,100\}} \end{aligned}$$

[4] In statistics, $T_k(x)$, $k = 1, \ldots, c$ is traditional. We used $\mathcal{E}_k$, since these arose as energies in the Gibbs thought experiment.

Let $h(x) = \binom{100}{x} \mathbb{1}_{x \in \{0,1,\ldots,100\}}$, $\mathcal{E}(x) = x$, and $\lambda(p) = \log \frac{p}{1-p}$. Then

$$f(x) = \frac{1}{Z(\lambda)} p(x) e^{\lambda \mathcal{E}(x)}$$

where

$$\frac{1}{Z(\lambda)} = e^{100 \log(1-p)} = e^{100 \log \frac{1}{1+e^{\lambda}}} = (1+e^{\lambda})^{-100} \implies Z(\lambda) = (1+e^{\lambda})^{100}$$

(b) Binomial(n, p) for parameter (n, p)

Analogously,

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \mathbb{1}_{x \in \{0,1,\ldots,n\}}$$

so with

$$\begin{aligned} h(x) &= \binom{n}{x} \mathbb{1}_{x \in \{0,1,\ldots,n\}} \\ \mathcal{E}(x) &= x \\ \lambda(p) &= \log \frac{p}{1-p} \\ Z(\lambda) &= (1+e^{\lambda})^n \end{aligned}$$

we have

$$f(x) = \frac{1}{Z(\lambda)} p(x) e^{\lambda \mathcal{E}(x)}$$

(c) (Do not submit) Normal(μ, σ^2) for parameter (μ, σ^2)

(d) Uniform$(0, m)$ for parameter m

$$f(x) = \frac{1}{m} \mathbb{1}_{x \in [0,m]}$$

but this has no parameter, so it is not an exponential family.

(e) (Do not submit) Gamma(α, β) for parameter (α, β) (i.e. $f(x; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \mathbb{1}_{x>0} x^{\alpha-1} e^{-\beta x}$)

4. **Samples from an exponential family.** Let $X_{1:n}$ be iid from an exponential family with sufficient statistics $\vec{\mathcal{E}} = \mathcal{E}_{1:c}$ over $\mathbb{R}$. Show that the joint distribution of $X_{1:n}$ is an exponential family over $\mathbb{R}^n$ and identify the sufficient statistics.

Since $X_{1:n}$ are iid, the joint distribution is

$$\begin{aligned} f(x_1, \ldots, x_n; \vec{\lambda}) &= \prod_{k=1}^{n} f(x_k; \vec{\lambda}) \\ &= \prod_{k=1}^{n} \frac{1}{Z(\vec{\lambda})} p(x_k) e^{\vec{\lambda} \cdot \mathcal{E}(x_k)} \\ &= \frac{1}{Z(\vec{\lambda})^n} \left(\prod_{k=1}^{n} p(x_k) \right) e^{\vec{\lambda} \cdot \sum_{k=1}^{n} \mathcal{E}(x_k)} \end{aligned}$$

If we define $\tilde{Z}(\vec{\lambda}) = Z(\vec{\lambda})^n$, $\tilde{p}(x_1, \ldots, x_n) = \prod_{k=1}^{n} p(x_k)$, and $\tilde{\mathcal{E}}(x_1, \ldots, x_n) = \sum_{k=1}^{n} \mathcal{E}(x_k)$, then we have

$$f(\vec{x}; \vec{\lambda}) = \frac{1}{\tilde{Z}(\vec{\lambda})} \tilde{h}(\vec{x}) e^{\vec{\lambda} \cdot \tilde{\mathcal{E}}(\vec{x})}$$

where each $\tilde{\mathcal{E}}_{1:c} = \sum_{k=1}^{n} \mathcal{E}_{1:c}(x_k) = n\bar{\mathcal{E}}$.

5. **The partition function has a few more tricks up its sleeve.** A common theme in statistical mechanics is the deep relationship between the properties of a physical system and the partition function of its equilibrium distribution. A similar theme emerges in statistics, as illustrated by our derivation, in the lectures, of a collection of properties relating derivatives of the partition function to the computation of parameters in exponential families. In this problem we will exploit this connection further, in order to get closed form solutions to some expectations that would be otherwise very difficult to compute.

Let X be a continuous random variable with pdf $f(x) = f(x; \vec{\lambda})$, where $\lambda = \{\lambda_1, \ldots, \lambda_c\}$ is a collection of parameters that define a family of distributions. This is an "exponential family with "canonical" or "natural" parametrization if f can be written in the form

$$f(x; \lambda) = \frac{1}{Z(\lambda)} p(x) e^{\vec{\lambda} \cdot T(x)} \tag{4}$$

where $p(x)$ is any function that is everywhere non-negative and $T(x) = (T_1(x), \ldots, T_c(x))$ are the "sufficient statistics".[5] This problem is about the particular family of pdfs defined by

$$f(x; \beta) = \frac{\mathbb{1}\{x > 0\}}{\sqrt{2\pi x^3}} \exp\left(-\frac{(x-\beta)^2}{2x\beta^2}\right)$$

for $x \in \mathbb{R}$ and parameter $\beta > 0$. The purpose of this problem is to use properties of exponential families to compute the mean and variance of X in terms of β.

But before jumping into it, let's pause to consider the direct approach, which means finding closed form expression for two integrals:

$$\mu \doteq \mathbb{E}[X] = \int_0^\infty \frac{x}{\sqrt{2\pi x^3}} \exp\left(-\frac{(x-\beta)^2}{2x\beta^2}\right) dx \quad \sigma^2 \doteq \text{Var}[X] = \int_0^\infty \frac{(x-\mu)^2}{\sqrt{2\pi x^3}} \exp\left(-\frac{(x-\beta)^2}{2x\beta^2}\right) dx$$

Not an easy task. So let's come at it from a different direction:

(a) Show that f can be expressed as an exponential family of the form given in equation (4), where $\lambda = \lambda(\beta)$ is a scalar function of β. Explicitly identify the function $\lambda(\beta)$, as well as $p(x)$, $T(x)$, and $Z(\lambda)$. (For the latter, be sure to write Z in terms of λ and not β.)

$$\begin{aligned}
f(x; b) &= \frac{\mathbb{1}\{x > 0\}}{\sqrt{2\pi x^3}} \exp\left(-\frac{(x-\beta)^2}{2x\beta^2}\right) \\
&= \frac{\mathbb{1}\{x > 0\}}{\sqrt{2\pi x^3}} \exp\left(-\frac{x^2 - 2x\beta + \beta^2}{2x\beta^2}\right) \\
&= \frac{\mathbb{1}\{x > 0\}}{\sqrt{2\pi x^3}} \exp\left(-\frac{x}{2\beta^2} + \frac{1}{\beta} - \frac{1}{2x}\right) \\
&= \frac{\mathbb{1}\{x > 0\}}{\sqrt{2\pi x^3}} e^{1/\beta} e^{-1/2x} e^{-x/2\beta^2}
\end{aligned}$$

Hence with

$$\begin{aligned}
p(x) &= \frac{\mathbb{1}\{x > 0\}}{\sqrt{2\pi x^3}} e^{-\frac{1}{2x}} \\
T(x) &= -x \\
\lambda(\beta) &= \frac{1}{2\beta^2} \\
Z(\lambda) &= e^{1/\beta} = e^{\sqrt{-2\lambda}}
\end{aligned}$$

we see that f is an exponential family.

(b) Use the properties of (natural) exponential families that we proved in class to compute the mean and variance of $T(X)$ in terms of λ. For full credit, you must use properties of exponential families. In other words, you should be solving this problem with differentiation, not integration.

In class, we showed that

$$\frac{\partial}{\partial \lambda} \log Z_\lambda = \mathbb{E}_\lambda[T_k(X)]$$

[5] In statistics, T (instead of $\mathcal{E}$) is traditional.

Hence,

$$\mathbb{E}_\lambda[T(X)] = \frac{\partial}{\partial\lambda}\log Z_\lambda = \frac{\partial}{\partial\lambda}\sqrt{2\lambda} = -\frac{1}{\sqrt{2\lambda}}$$

Further,

$$\frac{\partial^2}{\partial\lambda^2}\log Z_\lambda = \mathrm{Var}_\lambda(T(X))$$

so

$$\mathrm{Var}_\lambda[T(X)] = \frac{\partial^2}{\partial\lambda^2}\log Z_\lambda = -\frac{\partial}{\partial\lambda_1}\frac{1}{\sqrt{2\lambda}} = -(2\lambda)^{-3/2}$$

(c) Finally, use the function $\lambda(\beta)$ derived in (a) to express the mean and variance of X in terms of β.

We have $\lambda(\beta) = \frac{1}{2\beta^2}$, so

$$\mathbb{E}[X] = -\mathbb{E}[T(X)] = -\frac{1}{\sqrt{2\lambda}} = -\frac{1}{\sqrt{\frac{1}{\beta^2}}} = -\frac{1}{\frac{1}{\beta}} = \boxed{-\beta}$$

$$\mathrm{Var}[X] = \mathrm{Var}[T(X)] = \left(-2\cdot\frac{1}{2\beta^2}\right)^{-3/2} = \boxed{-\beta^3}$$

This is almost magical. We sidestepped two daunting integrations by simply identifying Z and then taking derivatives. But it wasn't quite a "free lunch"—after all, we were given the fact that f is normalized, which is not so obvious. In any case, you can perhaps begin to see why a closed-form expression for the partition function is so highly valued when modeling physical systems with Gibbs and related distributions.

6. **Prefix codes and uniquely decodable codes.** Recall that a binary code $C : \{1, \ldots, s\} \to \{0,1\}^*$ is a **prefix code** (also known as "instantaneous code") if $C(x)$ is not the prefix of any other code word $C(y)$ (e.g. 001 *is a prefix* of 0010). On the other hand, C is **uniquely decodable** if any finite string of codewords has a unique decoding. In other words, if $C(x_1) \cdots C(x_i) = C(y_1) \cdots C(y_j)$ then $i = j$ and $x_k = y_k$ for each $k = 1 : i$.[6]

Clearly, prefix codes are uniquely decodable (just read off the source symbols left-to-right), but uniquely decodable codes are not necessarily prefix codes. A theorem (proved in problem 11) shows that for every uniquely decodable code, C, there is a prefix code $\tilde{C}$ with codewords of the same length, i.e. $|\tilde{C}(x)| = |C(x)|$ for all x.

Consider the following three codes for the source $\{1, \ldots, 4\}$:

x	$C_1(x)$	$C_2(x)$	$C_3(x)$
1	00	0	0
2	01	1	01
3	10	01	011
4	11	11	111

(a) For each of the three codes, list all possible decodings of 001111.

C_1	C_2	C_3
144	112222	124
	11224	
	11242	
	11422	
	1144	
	13222	
	1324	
	1342	

(b) Which of these codes is uniquely decodable? Why?

Only C_1 and C_3 are uniquely decodable since C_1 is prefix and $\sum 2^{-|C_3(x)|} = 1 \leq 1$ so C_3 is uniquely decodable by Kraft-McMillan.

(c) One of these codes is uniquely decodable but not prefix. Find a prefix code with the same code lengths.

x	$\tilde{C}_3(x)$
1	0
2	10
3	110
4	111

is prefix and has the same code lengths as C_3.

(d) If $p = p_{1:4} = (.7, .1, .1, .1)$ (a probability on the source), then $\mathbb{E}_p|C(X)| \doteq \sum_{x=1:4} |C(x)|p_x$ is the average number of bits per symbol. Compute $\mathbb{E}_p|C(X)|$ for each of these three codes, and compare these averages to the entropy $H(p)$ (remember to use $\log_2$). Comment on which code might be the most useful.

$$\begin{aligned}
\mathbb{E}_p\,|C_1(X)| &= 0.7 \cdot |C_1(1)| + 0.1 \cdot |C_1(2)| + 0.1 \cdot |C_1(3)| + 0.1 \cdot |C_1(4)| \\
&= 0.7 \cdot 2 + 0.1 \cdot 2 + 0.1 \cdot 2 + 0.1 \cdot 2 = \boxed{2} \\
\mathbb{E}_p\,|C_2(X)| &= 0.7 \cdot |C_2(1)| + 0.1 \cdot |C_2(2)| + 0.1 \cdot |C_2(3)| + 0.1 \cdot |C_2(4)| \\
&= 0.7 \cdot 1 + 0.1 \cdot 1 + 0.1 \cdot 2 + 0.1 \cdot 2 = \boxed{1.2} \\
\mathbb{E}_p\,|C_3(X)| &= 0.7 \cdot |C_3(1)| + 0.1 \cdot |C_3(2)| + 0.1 \cdot |C_3(3)| + 0.1 \cdot |C_3(4)| \\
&= 0.7 \cdot 1 + 0.1 \cdot 2 + 0.1 \cdot 3 + 0.1 \cdot 3 = \boxed{1.5}
\end{aligned}$$

but

$$H(p) = -\sum_{x=1}^{4} p_x \log_2 p_x = -(0.7)(\log_2 0.7) - (3)(0.1)(\log_2 0.1) \approx 1.3568$$

[6] Here, $x_k \in \{1, \ldots, s\}$, $k = 1 : i$, are i source symbols and $C(x_1) \cdots C(x_i)$ is the concatenation of the corresponding codes; the right-hand side is interpreted analogously.

By a Theorem from class, the optimal code C^* satisfies

$$H(p) \leq \mathbb{E}_p |C^*| \leq H(p) + 1$$

so we know an optimal code must satisfy

$$1.3568 \leq \mathbb{E}_p |C^*| \leq 2.3568$$

Hence both C_1 and C_3 are candidates and are uniquely decodable. Since C_3 has the shorter expected code length, however, it is likely to be more useful.

(e) Repeat part (d) for $p = (.1, .1, .1, .7)$.

$$\begin{aligned}
\mathbb{E}_p |C_1(X)| &= 0.1 \cdot 2 + 0.1 \cdot 2 + 0.1 \cdot 2 + 0.7 \cdot 2 = \boxed{2} \\
\mathbb{E}_p |C_2(X)| &= 0.1 \cdot 1 + 0.1 \cdot 1 + 0.1 \cdot 2 + 0.7 \cdot 2 = \boxed{1.8} \\
\mathbb{E}_p |C_3(X)| &= 0.1 \cdot 1 + 0.1 \cdot 2 + 0.1 \cdot 3 + 0.7 \cdot 3 = \boxed{2.7}
\end{aligned}$$

and

$$H(p) = -(3)(0.1)(\log_2 0.1) - (0.7)(\log_2 0.7) \approx 1.3568$$

In this case, C_3 is no longer in the optimal range and C_2 is not uniquely decodable. Hence C_1 is the best code.

7. **Block coding.** Let $X_1, X_2, \ldots$ be iid with pmf $p = p_{1:2} = (0.3, 0.7)$. Find a binary and uniquely decodable code C on the source $\{1, 2\}$ that has expected code length per source symbol within 0.025 bits of the smallest possible value. (You might need to encode blocks.)

We know that for $C_n^* = \arg\min_{C_n} \mathbb{E}\,|C(X_{1:n})|$,

$$H(p) \leq \frac{1}{n}\mathbb{E}\,|C_n^*(X_{1:n})| \leq H(p) + \frac{1}{n}$$

Since we want to be within 0.025 bits of the optimal code, we need $n = \frac{1}{0.025} = 40$.

We can calculate

$$H(p) = -0.3\log_2 0.3 - 0.7\log_2 0.7 \approx 0.881$$

So we need

$$0.881 \leq \frac{1}{40}\mathbb{E}\,|C(X_{1:n})| \leq 0.906$$

We also know that for each symbol, we will need roughly $\log\frac{1}{0.3} \approx 1.7$ and $\log\frac{1}{0.7} \approx 0.5$ bits so it is reasonable to use blocks of length 2.

Now,

x	$p(x)$
11	0.09
12	0.21
21	0.21
22	0.49

We want the highest probability symbols to have the shortest code lengths, so consider the code:

x	$C(x)$
11	111
12	110
21	10
22	0

which is prefix and has

$$\mathbb{E}\,|C(x)| = 0.09(3) + 0.21(3) + 0.21(2) + 0.49(1) = 1.81 \implies 0.905 \text{ bits/symbol}$$

which is within the desired range.

8. **Shannon codes.** Let $X_{1:n}$ be iid Bernoulli(p), with $p = 0.2$. For each positive integer n, each $q \in (0,1)$ and each $x_{1:n} \in \{0,1\}^n$, define

$$f_{n,q}(x_{1:n}) \triangleq \prod_{i=1}^{n} q^{x_i}(1-q)^{1-x_i}$$

which is the probability mass function of n iid Bernoulli(q) random variables.

(a) Use the Kraft-McMillan theorem to show that, for each positive integer n and each $q \in (0,1)$, there exists a (variable length, binary) prefix code, say $C_{n,q}$, over $\{0,1\}^n$ (a source with $s = 2^n$ symbols) with

$$|C_{n,q}(x_{1:n})| = \lceil -\log_2 f_{n,q}(x_{1:n}) \rceil$$

for all $x_{1:n} \in \{0,1\}^n$, where $\lceil a \rceil$ is the ceiling function, i.e., the smallest integer greater than or equal to a. (This "Shannon code" was the key construction in our proof of Corollary 1 to the Kraft-McMillan Inequality.)

By the Kraft-McMillan Inequality, for any code lengths $\ell_1, \ldots, \ell_{2^n}$ with

$$\sum_{x=1}^{2^n} 2^{-\ell_x} \leq 1$$

there exists a prefix code with $|C(x)| = \ell_x$.

In particular, this means that for any n and q, it suffices to show that

$$\sum_{x=1}^{2^n} 2^{-\ell_x} = \sum_{x=1}^{2^n} 2^{-\lceil -\log_2 f_{n,q}(x_{1:n}) \rceil} \leq 1$$

And in fact,

$$\begin{aligned}
\sum_{x=1}^{2^n} 2^{-\lceil -\log_2 f_{n,q}(x_{1:n}) \rceil} &\leq \sum_{x=1}^{2^n} 2^{\log_2 f_{n,q}(x_{1:n}+1)} \\
&= \sum_{x=1}^{2^n} f_{n,q}(x_{1:n}) \\
&= \sum_{x=1}^{2^n} \prod_{i=1}^{n} q^{x_i}(1-q)^{1-x_i} \\
&= (q + [1-q])^n \\
&= 1^n = 1
\end{aligned}$$

(b) (Do not submit) For all (q,n) combinations of $q = .01, .1, .2, .5, .8$ and $n = 1, 2, 3, 4, 5, 10, 20, 100, 500$, approximate $\frac{1}{n}\mathbb{E}\,|C_{n,q}(X_{1:n})|$ by generating 10^5 independent realizations of $X_{1:n}$ and averaging the corresponding code lengths for each realization (and then divide by n).[7] As n varies, this gives a graph for each of the five values of q. Plot these 5 graphs on the same plot.[8]

(c) (Do not submit) For each of the five values of q, compute the limiting value as $n \to \infty$, i.e.,

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}\,|C_{n,q}(X_{1:n})|$$

(d) (Do not submit) Interpret the graphs. Which code is best? Is it possible to do much better than this code? Why or why not?

---

[7] One way to generate $X_{1:n}$ in Matlab is `(rand(1,n)<p)`. Also, `ceil` is the ceiling function in Matlab.

[8] The matlab command `hold on` will let you put multiple graphs on the same plot. Use `hold off` to stop this behavior and revert to the default where each plotting command clears the current plot.

For 2610 or for extra credit:

9. **Large deviations and the "rate function."** Some of you will have been introduced to large deviations through Cramèr's Theorem and the rate function, $I(\theta)$. Here's one version, in the discrete setting:

   Let p be a pmf on $\{1, \ldots, s\}$, let $\mathcal{E}(x)$ be a function on $\{1, \ldots, s\}$, and let $X_{1:n} \sim \ iid\, p$. Then for any $\theta > \mathbb{E}_h[\mathcal{E}(X)]$:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{E}(X_i) \geq \theta\right) = -I(\theta) \tag{5}$$

   where

$$I(\theta) = \sup_{\lambda\in\mathbb{R}} [\theta\lambda - \phi(\lambda)]$$

   is the "rate function" and $\phi(\lambda)$ is the "cumulant generating function" of $\mathcal{E}(X)$

$$\phi(\lambda) = \log \mathbb{E}_h\left[e^{\lambda\mathcal{E}(X)}\right]$$

   The point of this problem is to connect the rate function $\phi(\lambda)$ (from large-deviation theory) to Sanov's theorem.

   To this end, for some given function $\mathcal{E} : \{1, \ldots, s\} \to \mathbb{R}$ (not identically zero), consider the exponential family

$$p_x = p_x(\lambda) = \frac{1}{Z_\lambda} h_x e^{\lambda\mathcal{E}(x)}$$

   and let θ be some fixed constant. In class we outlined a proof that the function $\log Z_\lambda - \theta\lambda$ is strictly convex in λ, and its unique minimum

$$\lambda^* = \arg\min_{\lambda} (\log Z_\lambda - \theta\lambda)$$

   satisfies the equation $\mathbb{E}_{p(\lambda^*)}[\mathcal{E}(X)] = \theta$. (The detailed proof is in the notes.)

   Show that

$$D(p(\lambda^* \| p)) = I(\theta)$$

10. **An elegant proof of the Kraft inequality.** The first part of the Kraft-McMillan theorem is known as the Kraft inequality. It says that every binary prefix code C over a discrete alphabet Ω satisfies $\sum_{w \in \Omega} 2^{-|C(w)|} \leq 1$. Here is a simple method of proof (notable perhaps because it uses probabilities to prove something that is purely combinatorial): Let $\dagger$ be a symbol that is not in Ω. Define the function $f : \{0,1\}^\infty \to \Omega \cup \{\dagger\}$ as follows: for any infinite binary sequence $z_{1:\infty}$, if $C(w) = z_{1:\ell}$ for some $w \in \Omega$ and some ℓ, then define $f(z_{1:\infty}) = w$; otherwise, define $f(z_{1:\infty}) = \dagger$. Now, let $Z_1, Z_2, \ldots$ be iid Bernoulli$(1/2)$, and consider $\mathbb{P}(f(Z_{1:\infty}) = w)$ for each $w \in \Omega$. Use this to prove the Kraft inequality. Somewhere in your proof you should use the fact that the code is a prefix code![9]

[9] The same method of proof can be used for d-ary codes; i.e. codes that use the symbols $\{0, 1, \ldots, d-1\}$, instead of just $\{0, 1\}$.

11. **Uniquely decodable codes are no shorter than prefix codes.** Recall the statement of the Kraft-McMillan theorem for (binary) prefix codes: Given a set of source symbols $\{1, 2, \ldots, s\}$, every prefix code C satisfies

$$\sum_{x=1}^{s} \frac{1}{2^{|C(x)|}} \leq 1$$

and for every set of lengths $\ell_1, \ell_2 \cdots \ell_s$ satisfying

$$\sum_{x=1}^{s} \frac{1}{2^{\ell_x}} \leq 1$$

there exists a prefix code with $|C(x)| = \ell_x$.

As noted in class, the proof makes use of a simple but powerful relationship between prefix codes and binary trees. It was also noted that both parts of the theorem still hold with "prefix" replaced by "uniquely decodable." The point of this assignment is to prove the extension of the theorem from prefix codes to uniquely decodable codes.

Notice that there is really nothing to do for the second part of the theorem, since prefix codes are obviously uniquely decodable. Hence, what needs to be shown is that every uniquely decodable code C satisfies the Kraft inequality:

$$\sum_{x=1}^{s} \frac{1}{2^{|C(x)|}} \leq 1 \tag{6}$$

Use the following steps to prove 6:

(a) Show that for any integer $m > 0$

$$\Big(\sum_{x=1}^{s} \frac{1}{2^{|C(x)|}}\Big)^m = \sum_{\{x_{1:m} \in \{1,\ldots,s\}^m\}} \frac{1}{2^{|C(x_1)\cdots C(x_m)|}} = \sum_{l=1}^{\infty} q_l^{(m)} \frac{1}{2^l} \tag{7}$$

where $q_l^{(m)} = \#\{x_{1:m} \in \{1, \ldots, s\}^m : |C(x_1) \cdots C(x_m)| = l\}$.[10]

(b) Let $n = \min_{x \in \{1,\ldots,s\}} |C(x)|$ and $N = \max_{x \in \{1,\ldots,s\}} |C(x)|$ and show that

i. $q_l^{(m)} = 0$ unless $mn \leq l \leq mN$, and

ii. $q_l^{(m)} \leq 2^l$

(c) Conclude that for all m

$$\sum_{l=1}^{\infty} q_l^{(m)} \frac{1}{2^l} \leq m(N - n) + 1$$

and, putting this together with (7), conclude that uniquely decodable codes satisfy the Kraft inequality.

[10] $\#A$ is the number of elements in a set A.