Handout 16 Information Theory and Coding
Solutions to Midterm exam Oct. 29, 2024

Problem 1. (16 points)

Suppose $f : [0, \infty) \to \mathbb{R} \cup \{\pm\infty\}$ is a decreasing, convex function, and p and q are probability distributions on an alphabet $\mathcal{U}$ (i.e., $p(u) \geq 0$ and $\sum_{u \in \mathcal{U}} p(u) = 1$, similarly for q). Define

$$K_f(p, q) = \sum_{u : p(u) > 0} p(u)\, f\left(\frac{q(u)}{p(u)}\right).$$

(a) (2 pts) Show that $K_f(p, q) \geq f(1)$, and equality happens if $q = p$.
Hint: Make sure to use convexity.

Solution: Using the convexity of f, Jensen's inequality gives

$$K_f(p, q) = \sum_{u : p(u) > 0} p(u)\, f\left(\frac{q(u)}{p(u)}\right) \geq f\left(\sum_{u : p(u) > 0} p(u)\, \frac{q(u)}{p(u)}\right) = f\left(\sum_{u : p(u) > 0} q(u)\right).$$

Moreover, $\sum_{u : p(u) > 0} q(u) \leq 1$ and f is decreasing so that $f\left(\sum_{u : p(u) > 0} q(u)\right) \geq f(1)$.

When $q = p$, we get

$$K_f(p, p) = \sum_{u : p(u) > 0} p(u)\, f\left(\frac{p(u)}{p(u)}\right) = \sum_{u : p(u) > 0} p(u)\, f(1) = f(1).$$

Suppose U is a random variable with distribution p. A "prediction" about U is a probability distribution q on $\mathcal{U}$ — basically saying "I believe we will see the value u with probability $q(u)$". A prediction q is assigned a score via $\operatorname{score}(q) = \mathbb{E}\left[\frac{1}{q(U)}\right] = \sum_u \frac{p(u)}{q(u)}$.

(b) (3 pts) Let $p_{1/2}(u) = \frac{p(u)^{1/2}}{A}$, where $A = \sum_u p(u)^{1/2}$ to ensure that $p_{1/2}$ is a probability distribution. Show that for any probability distribution q, $\operatorname{score}(q) \geq A^2$, with equality if $q = p_{1/2}$.
Hint: First show that with $f(x) = 1/x$, $\operatorname{score}(q) = A^2 K_f(p_{1/2}, q)$.

Solution: Following the hint, we consider $f(x) = 1/x$ (which is indeed convex and decreasing) and first show that $\operatorname{score}(q) = A^2 K_f(p_{1/2}, q)$. We have

$$\begin{aligned}
A^2 K_f(p_{1/2}, q) &= A^2 \sum_{u : p(u) > 0} \frac{\left(p_{1/2}(u)\right)^2}{q(u)} \\
&= A^2 \sum_{u : p(u) > 0} \frac{\left(p^{1/2}(u)/A\right)^2}{q(u)} \\
&= \sum_{u : p(u) > 0} \frac{p(u)}{q(u)} \\
&= \sum_u \frac{p(u)}{q(u)} \\
&= \operatorname{score}(q).
\end{aligned}$$

By the derivation in part (a), $K_f(p_{1/2}, q) \geq f(1)$ with equality when $q = p_{1/2}$, from which we conclude that $\text{score}(q) \geq A^2 f(1) = A^2$ with equality when $q = p_{1/2}$.

(c) (3 pts) Suppose $c : \mathcal{U} \to \{0,1\}^*$ is a uniquely decodable code. Show that $\mathbb{E}[2^{\text{length}(c(U))}] \geq A^2$.

Solution: We would like to make a particular choice for a distribution q and use the result in part (b). The way score is defined suggests the choice $q(u) = 2^{-\text{length}(c(u))}$, however we need to q to be a probability distribution. By normalizing appropriately, our guess becomes $q(u) = 2^{-\text{length}(c(u))}/(\sum_v 2^{-\text{length}(c(v))})$. This choice gives

$$
\begin{aligned}
\text{score}(q) &= \sum_u p(u) 2^{\text{length}(c(u))} \left(\sum_v 2^{-\text{length}(c(v))} \right) \\
&\leq \sum_u p(u) 2^{\text{length}(c(u))} \\
&= \mathbb{E}[2^{\text{length}(c(U))}],
\end{aligned}
$$

where the inequality follows from Kraft's inequality since c is uniquely decodable. The result then follows from part (b) since $\text{score}(q) \geq A^2$.

Fix $\alpha > 0$.

(d) (3 pts) Replace the score function above with $\text{score}_\alpha(q) = \mathbb{E}[q(U)^{-\alpha}] = \sum_u \frac{p(u)}{q(u)^\alpha}$. Show that for any q, $\text{score}_\alpha(q) \geq (A_{1/1+\alpha})^{1+\alpha}$, with equality if $q = p_{1/1+\alpha}$, where we define $p_s(u) = p(u)^s / A_s$ where $A_s = \sum_u p(u)^s$.

Hint: Choose f appropriately and express $\text{score}_\alpha(q)$ in terms of $K_f(p_s, q)$ for some s.

Solution: The proof strategy is similar to the one in part (b). Let us consider $f(x) = 1/x^\alpha$ (which is indeed convex and decreasing, as required) and show that $\text{score}_\alpha(q) = (A_{1/1+\alpha})^{1+\alpha} K_f(p_{1/1+\alpha}, q)$. We have

$$
\begin{aligned}
(A_{1/1+\alpha})^{1+\alpha} K_f(p_{1/1+\alpha}, q) &= (A_{1/1+\alpha})^{1+\alpha} \sum_{u:p(u)>0} \frac{\left(p_{1/1+\alpha}(u)\right)^{1+\alpha}}{q(u)^\alpha} \\
&= (A_{1/1+\alpha})^{1+\alpha} \sum_{u:p(u)>0} \frac{\left(p^{1/(1+\alpha)}(u)/A_{1/1+\alpha}\right)^{1+\alpha}}{q(u)^\alpha} \\
&= \sum_{u:p(u)>0} \frac{p(u)}{q(u)^\alpha} \\
&= \sum_u \frac{p(u)}{q(u)^\alpha} \\
&= \text{score}_\alpha(q).
\end{aligned}
$$

By the derivation in part (a), $K_f(p_{1/1+\alpha}, q) \geq f(1)$ with equality when $q = p_{1/1+\alpha}$, from which we conclude that $\text{score}_\alpha(q) \geq (A_{1/1+\alpha})^{1+\alpha} f(1) = (A_{1/1+\alpha})^{1+\alpha}$ with equality when $q = p_{1/1+\alpha}$.

(e) (2 pts) Show that for any uniquely decodable code $c : \mathcal{U} \to \{0,1\}^*$,

$$
\mathbb{E}[2^{\alpha\,\text{length}(c(U))}] \geq (A_{1/1+\alpha})^{1+\alpha}.
$$

Solution: As in part (c), we select $q(u) = 2^{-\operatorname{length}(c(u))}/(\sum_v 2^{-\operatorname{length}(c(v))})$. This choice of distribution gives

$$\begin{aligned}
\operatorname{score}_\alpha(q) &= \sum_u p(u) 2^{\alpha \operatorname{length}(c(u))} \left(\sum_v 2^{-\operatorname{length}(c(v))} \right)^\alpha \\
&\leq \sum_u p(u) 2^{\alpha \operatorname{length}(c(u))} \\
&= \mathbb{E}[2^{\alpha \operatorname{length}(c(U))}],
\end{aligned}$$

where the inequality follows from Kraft's inequality since c is uniquely decodable. The result then follows from part (d) since $\operatorname{score}_\alpha(q) \geq (A_{1/1+\alpha})^{1+\alpha}$.

(f) (3 pts) Show that there exists a prefix-free code $c : \mathcal{U} \to \{0,1\}^*$ such that

$$\mathbb{E}[2^{\alpha \operatorname{length}(c(U))}] \leq 2^\alpha (A_{1/1+\alpha})^{1+\alpha}.$$

Solution: Notice that in order to prove an upper-bound on $\mathbb{E}[2^{\alpha \operatorname{length}(c(U))}]$, we somehow have to make a choice in which $q = p_{1/1+\alpha}$ for otherwise we know from previous parts that we obtain a lower-bound on $\mathbb{E}[2^{\alpha \operatorname{length}(c(U))}]$.

Our strategy is the following: we construct a prefix-free code by choosing the length of codewords according to

$$\operatorname{length}(c(u)) = \left\lceil -\log(p_{1/1+\alpha}(u)) \right\rceil, \quad u \in \mathcal{U}.$$

This choice of lengths satisfies Kraft's inequality, and hence, there exists a prefix-free code with these lengths (such as a Shannon code, see Problem 3 in Homework 2). Thus, we get

$$\begin{aligned}
\mathbb{E}[2^{\alpha \operatorname{length}(c(U))}] &= \sum_u p(u) 2^{\alpha \left\lceil -\log(p_{1/1+\alpha}(u)) \right\rceil} \\
&\leq \sum_u p(u) 2^{\alpha \left(-\log(p_{1/1+\alpha}(u)) + 1 \right)} \\
&= 2^\alpha \sum_u p(u) 2^{\log(p_{1/1+\alpha}(u)^{-\alpha})} \\
&= 2^\alpha \sum_u \frac{p(u)}{p_{1/1+\alpha}(u)^\alpha} \\
&= 2^\alpha \operatorname{score}_\alpha(p_{1/1+\alpha}).
\end{aligned}$$

Finally, part (d) tells us that the α-score of $p_{1/1+\alpha}$ is precisely equal to $(A_{1/1+\alpha})^{1+\alpha}$, so that indeed

$$\mathbb{E}[2^{\alpha \operatorname{length}(c(U))}] \leq 2^\alpha (A_{1/1+\alpha})^{1+\alpha}.$$

Remarks: In the lectures, we saw that a possible choice of the score is $\operatorname{score}(q) = \mathbb{E}\left[\log \frac{1}{q(U)}\right]$. The problem of minimizing this score is equivalent to the problem of minimizing the expected codeword length (with the identification $q(u) \propto 2^{-\operatorname{length}(c(u))}$), and the minimizer is $q = p$, i.e., $\operatorname{length}(c(u)) = -\log p(u)$ rounded up. In this problem, we see how different choices of the score such as $\mathbb{E}[q(U)^{-\alpha}]$ can lead to surprising observations, such as the "best prediction" not even being $q = p$, but rather $q = p_{1/1+\alpha}$. That is, if the objective is to minimize the expected value of $2^{\alpha \operatorname{length} c(u))}$ rather than simply $\operatorname{length}(c(u))$, the optimal choice is $\operatorname{length}(c(u)) = -\log p_{1/1+\alpha}$ rounded up. The quantity $\mathbb{E}[2^{\alpha \operatorname{length}(c(U))}]$ as a function of α is the moment generating function of $\operatorname{length}(c(U))$, which is useful in obtaining tail probability bounds such as $\Pr\{\operatorname{length}(c(U)) \geq l\} = \Pr\left\{2^{\alpha \operatorname{length}(c(U))} \geq 2^{\alpha l}\right\} \leq 2^{-\alpha l}\mathbb{E}[2^{\alpha \operatorname{length}(c(U))}]$, using the Markov inequality. The quantity $K_f(p, q)$ is called an f-divergence, usually denoted by $D_f(q \,\|\, p)$, which can be defined for any convex f. A special example is the KL divergence, obtained by taking $f(x) = x \log x$ or $f(x) = -\log x$. If $f(1) = 0$, we get $D_f(p \,\|\, p) = 0$. Other well-known examples include the squared Hellinger distance, total variation, chi-squared divergence, and so on.

Problem 2. (12 points)

For this problem, we define the following notation different from that used in the lectures. Fix a natural number n. Let $(X_1, \ldots, X_n)$ be a vector of binary random variables, with each X_i taking values in $\{0, 1\}$. For $i, j = 1, \ldots, n$, let $X_i^j = (X_i, \ldots, X_j)$ if $i \leq j$ and empty if $i > j$. Let $X_{\neq i}$ denote the vector X_1^n without the i^{th} element, i.e., $X_{\neq i} = (X_1^{i-1}, X_{i+1}^n)$. Also let $X_{(\bar{i})}$ denote the vector X_1^n with its i^{th} element flipped, i.e., $X_{(\bar{i})} = (X_1^{i-1}, 1 - X_i, X_{i+1}^n)$.

(a) (3 pts) Show that $\sum_{i=1}^n H(X_i \mid X_{\neq i}) \leq H(X_1^n)$.

Solution: Using the fact that conditioning reduces entropy, we have

$$\begin{aligned} \sum_{i=1}^n H(X_i \mid X_{\neq i}) &= \sum_{i=1}^n H(X_i \mid X_1^{i-1}, X_{i+1}^n) \\ &\leq \sum_{i=1}^n H(X_i \mid X_1^{i-1}) = H(X_1^n), \end{aligned}$$

with the last equality immediate from the chain rule of entropy.

Let A be a subset of $\{0, 1\}^n$, i.e., A consists of binary vectors of length n. Denote by $E(A)$ the set of pairs of vectors in A that differ at *exactly* one position, i.e.,

$$\begin{aligned} E(A) &= \{(x_1^n, \tilde{x}_1^n) \in A \times A \text{ such that } \tilde{x}_i \neq x_i \text{ for exactly one } i\} \\ &= \{(x_1^n, \tilde{x}_1^n) \in A \times A \text{ such that } \tilde{x}_1^n = x_{(\bar{i})} \text{ for some } i\}. \end{aligned}$$

Let $(X_1, \ldots, X_n)$ be randomly and uniformly chosen from A.

(b) (3 pts) Fix $x_1^n \in A$. Compute $H(X_i \mid X_{\neq i} = x_{\neq i})$.
Hint: Consider two cases: $x_{(\bar{i})} \in A$ and $x_{(\bar{i})} \notin A$.

Solution: As suggested by the hint, first suppose $x_{(\bar{i})} \in A$. Then, as both x_1^n and $x_{(\bar{i})}$ (which, by definition is the vector x_1^n with the i^{th} element flipped) are in A, given that $X_{\neq i} = x_{\neq i}$, X_i could either be 0 or 1 with equal probability, since X_1^n is picked uniformly from A. Hence, $H(X_i \mid X_{\neq i} = x_{\neq i}) = 1$ when $x_{(\bar{i})} \in A$. On the other hand, if $x_{(\bar{i})}$ is not in A, then given $X_{\neq i} = x_{\neq i}$, X_i must be equal to the i^{th} element of x_1^n, and hence $H(X_i \mid X_{\neq i} = x_{\neq i}) = 0$ when $x_{(\bar{i})} \notin A$. Putting the two together, we have $H(X_i \mid X_{\neq i} = x_{\neq i}) = \mathbb{1}\{x_{(\bar{i})} \in A\}$ for any $x_1^n \in A$.

(c) (3 pts) Show that $H(X_i \mid X_{\neq i}) = \frac{1}{|A|}\sum_{x_1^n \in A} \mathbb{1}\{x_{(\bar{i})} \in A\}$.

Solution: Let p denote the distribution induced by the uniformly drawn X_1^n. By definition of conditional entropy, $H(X_i \mid X_{\neq i}) = \sum_{x_{\neq i}} p(x_{\neq i}) H(X_i \mid X_{\neq i} = x_{\neq i})$. We can write this as

$$\begin{aligned} H(X_i \mid X_{\neq i}) &= \sum_{x_{\neq i} \in \{0,1\}^{n-1}} p(x_{\neq i}) H(X_i \mid X_{\neq i} = x_{\neq i}) \\ &= \sum_{x_{\neq i} \in \{0,1\}^{n-1}} p(x_{\neq i}) \left(\sum_{x_i \in \{0,1\}} p(x_i) \right) H(X_i \mid X_{\neq i} = x_{\neq i}) \\ &= \sum_{x_1^n \in \{0,1\}^n} p(x_1^n) H(X_i \mid X_{\neq i} = x_{\neq i}) \\ &= \frac{1}{|A|} \sum_{x_1^n \in A} \mathbb{1}\{x_{(\bar{i})} \in A\}, \end{aligned}$$

since for a given $x_1^n \in A$, $H(X_i \mid X_{\neq i} = x_{\neq i}) = \mathbb{1}\{x_{(\bar{i})} \in A\}$. Note that we could not have written this equality without first fixing an x_1^n.

(d) (3 pts) Show that $\sum_{i=1}^n H(X_i \mid X_{\neq i}) = \frac{|E(A)|}{|A|}$ and conclude that $|E(A)| \leq |A| \log |A|$.
Hint: Use (a).

Solution: Using the result in (c), we directly have

$$\begin{aligned} \sum_{i=1}^n H(X_i \mid X_{\neq i}) &= \sum_{i=1}^n \frac{1}{|A|} \sum_{x_1^n \in A} \mathbb{1}\{x_{(\bar{i})} \in A\} \\ &= \frac{1}{|A|} \sum_{x_1^n \in A} \sum_{i=1}^n \mathbb{1}\{x_{(\bar{i})} \in A\}. \end{aligned}$$

Observe that for $x_1^n \in A$, $\sum_{i=1}^n \mathbb{1}\{x_{(\bar{i})} \in A\}$ is equal to 1 if either one of $(x_1^n, x_{(\bar{i})})$ or $(x_{(\bar{i})}, x_1^n)$ is in A. Hence, summing this over all x_1^n in A, we get $|E(A)|$, and we have that $\sum_{i=1}^n H(X_i \mid X_{\neq i}) = \frac{|E(A)|}{|A|}$. From part (a), we know that $\sum_{i=1}^n H(X_i \mid X_{\neq i}) \leq H(X_1^n)$, which is in turn less than $\log |A|$ since X_1^n is distributed on A, and we are done.

Remarks: The set $\{0,1\}^n$ is the binary hypercube, equipped with a graph structure by defining the edge relation as in the definition of $E(A)$, i.e., two points in $\{0,1\}^n$ are connected by an edge if they differ at exactly one position. The subset A is then a subgraph of $\{0,1\}^n$ and $E(A)$ is then the set of *directed* edges induced by A. The result in (d) shows that the density of directed edges induced by a subgraph A (i.e., $\frac{|E(A)|}{|A|}$) is at most $\log |A|$ (if we considered unordered pairs in the definition, we would get (undirected) edges and a density of $\frac{1}{2}\log |A|$). Equality occurs if and only if A is a lower-dimensional hypercube in $\{0,1\}^n$ (i.e., X_i are all independent when X_1^n is uniformly distributed on A), so the inequality is tight. See Fig. 1 for an illustration. The inequality derived in part (a) is sometimes called Han's inequality.

PROBLEM 3. (15 points)

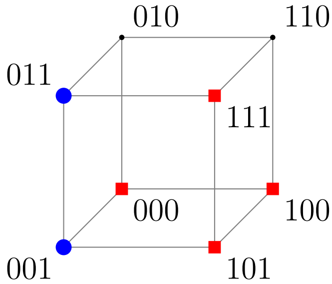

Figure 1: Example of binary hypercube with $n = 3$ with the set A_1 marked by red squares and A_2 marked by large blue circles. $|A_1| = 4$, $|E(A)| = 6 < 8 = 4 \cdot 2 = |A_1| \log |A_1|$. $|A_2| = 2$, $|E(A)| = 2 = 2 = 2 \cdot 1 = |A_2| \log |A_2|$.

(a) (2 pts) Suppose p is a probability distribution on $\mathcal{U}$. Show that for any probability distribution q on $\mathcal{U}$, $\max_{u\in\mathcal{U}} \log \frac{p(u)}{q(u)} \geq 0$. Additionally, show that $\min_q \max_{u\in\mathcal{U}} \log \frac{p(u)}{q(u)} = 0$, where the minimization is over all probability distributions q on $\mathcal{U}$.

Solution: Recall that the KL divergence $D(p\|q)$ between two probability distributions p and q is always non-negative and hence

$$\begin{aligned}
0 \leq D(p\|q) &= \sum_u p(u) \log \frac{p(u)}{q(u)} \\
&\leq \sum_u p(u) \max_v \log \frac{p(v)}{q(v)} \\
&= \left(\max_v \log \frac{p(v)}{q(v)} \right) \sum_u p(u) \\
&= \max_v \log \frac{p(v)}{q(v)}.
\end{aligned}$$

Finally, notice that for $q = p$, we have $\max_v \log \frac{p(v)}{q(v)} = \log(1) = 0$, so that the minimum over distributions q of $\max_v \log \frac{p(v)}{q(v)}$ is in indeed equal to zero.

(b) (2 pts) Show that $\min_q \max_{u\in\mathcal{U}} \log \frac{f(u)}{q(u)} = \log K$, where $K = \sum_u f(u)$ for a nonnegative function f.

Hint: Use (a).

Solution: In order to leverage part (a), we need to deal with probability distributions. Following the hint, we can render $f(u)$ a probability distribution by appropriately normalizing it, that is by considering $p(u) = f(u)/K$. With this, we have

$$\begin{aligned}
\min_q \max_{u\in\mathcal{U}} \log \frac{f(u)}{q(u)} &= \min_q \max_{u\in\mathcal{U}} \log \frac{Kf(u)}{Kq(u)} \\
&= \min_q \max_{u\in\mathcal{U}} \left\{ \log K + \log \frac{f(u)}{Kq(u)} \right\} \\
&= \min_q \max_{u\in\mathcal{U}} \log K + \underbrace{\min_q \max_{u\in\mathcal{U}} \log \frac{f(u)}{Kq(u)}}_{=0 \text{ from part (a)}} \\
&= \log K.
\end{aligned}$$

Suppose from now on that for every θ in some parameter set Θ, we have a probability distribution p_θ on $\mathcal{U}$.

(c) (2 pts) Show that $\min_q \max_{u\in\mathcal{U},\theta\in\Theta} \log \frac{p_\theta(u)}{q(u)} = S$, where $S = \log \sum_{u\in\mathcal{U}} \max_{\theta\in\Theta} p_\theta(u)$.

Hint: Use (b).

Solution: The only difference here compared to previous parts is that we have an additional maximum over the parameters θ. By remembering that the logarithm is an increasing function, we can swap a maximum with a log. Formally, this means

$$\min_q \max_{u\in\mathcal{U},\theta\in\Theta} \log \frac{p_\theta(u)}{q(u)} = \min_q \max_{u\in\mathcal{U}} \log \frac{\max_{\theta\in\Theta} p_\theta(u)}{q(u)}.$$

It remains to use our result from part (b) with the choice $f(u) = \max_{\theta\in\Theta} p_\theta(u)$ (which is indeed nonnegative), so that $K = \sum_{u\in\mathcal{U}} \max_{\theta\in\Theta} p_\theta(u)$. Combining everything yields $\min_q \max_{u\in\mathcal{U},\theta\in\Theta} \log \frac{p_\theta(u)}{q(u)} = \log \sum_{u\in\mathcal{U}} \max_{\theta\in\Theta} p_\theta(u)$ as expected.

Let us also note that part (a) informs us that the value S is attained for the distribution $q = (\max_{\theta\in\Theta} p_\theta(u))/K$.

(d) (3 pts) Suppose we do not know the probability distribution of a random variable U, except that the distribution is one of the p_θ above. Show that there is a prefix-free code $c : \mathcal{U} \to \{0,1\}^*$ such that, for every $\theta \in \Theta$ and every $u \in \mathcal{U}$, length $c(u) \le \log \frac{1}{p_\theta(u)} + S + 1$, where S is as in part (c) above.

Solution: We use Shannon coding with codewords lengths given by the probability distribution encountered in part (e), $\max_{\phi\in\Theta} p_\phi(u)/K$, where $K = \sum_{u\in\mathcal{U}} \max_{\phi\in\Theta} p_\phi(u)$. This choice of lengths satisfies Kraft's inequality, hence there exists a prefix-free code with these codeword lengths. For any $u \in \mathcal{U}$, we have

$$\begin{aligned}
\text{length}(c(u)) &= \left\lceil -\log\left(\frac{\max_{\phi\in\Theta} p_\phi(u)}{K}\right)\right\rceil \\
&\le \log\left(\frac{1}{(\max_{\phi\in\Theta} p_\phi(u))/K}\right) + 1 \\
&= \log\left(\frac{p_\theta(u)}{(\max_{\phi\in\Theta} p_\phi(u))/K}\right) + \log\left(\frac{1}{p_\theta(u)}\right) + 1 \\
&\le \max_{u\in\mathcal{U},\theta\in\Theta} \log\left(\frac{p_\theta(u)}{(\max_{\theta\in\Theta} p_\phi(u))/K}\right) + \log\left(\frac{1}{p_\theta(u)}\right) + 1.
\end{aligned}$$

From part (c), we know $\min_q \max_{u\in\mathcal{U},\theta\in\Theta} \log \frac{p_\theta(u)}{q(u)} = S$, with S attained when $q = (\max_{\theta\in\Theta} p_\theta(u))/K$. Hence $\max_{u\in\mathcal{U},\theta\in\Theta} \log\left(\frac{p_\theta(u)}{(\max_{\theta\in\Theta} p_\phi(u))/K}\right) = S$ and we conclude that

$$\text{length}(c(u)) \le S + \log\left(\frac{1}{p_\theta(u)}\right) + 1.$$

Suppose we know that $U_1, U_2, \ldots,$ are i.i.d. Bernoulli(θ) random variables, but we do not know the value of $\theta \in [0,1]$. For $u^n \in \{0,1\}^n$, define $p_\theta(u^n) = \theta^{k(u^n)}(1-\theta)^{n-k(u^n)}$, where $k(u^n)$ is the number of 1's in the sequence $(u_1, \ldots, u_n)$. With this definition, $\Pr(U^n = u^n) = p_\theta(u^n)$.

(e) (3 pts) Show that for any u^n, we have $\max_{\theta\in[0,1]} p_\theta(u^n) = \left(\frac{k}{n}\right)^k \left(1-\frac{k}{n}\right)^{n-k}$ with $k = k(u^n)$, and conclude that

$$\sum_{u^n\in\{0,1\}^n} \max_{\theta\in[0,1]} p_\theta(u^n) = \sum_{i=0}^{n} \binom{n}{i} \left(\frac{i}{n}\right)^i \left(1-\frac{i}{n}\right)^{n-i}.$$

Hint: Differentiate $\log p_\theta(u^n)$ with respect to θ.

Solution: First, notice that the probability of a sequence u^n solely depends on the number of 1's appearing in it. As such, sequences with the same number of 1's are assigned the same probability. Since there are $\binom{n}{k}$ sequences with k ones, we can rewrite the sum over all sequences as follows:

$$\sum_{u^n\in\{0,1\}^n} \max_{\theta\in[0,1]} \theta^{k(u^n)}(1-\theta)^{n-k(u^n)} = \sum_{k=0}^{n} \binom{n}{k} \max_{\theta\in[0,1]} \theta^k(1-\theta)^{n-k} \tag{1}$$

Next, we evaluate $\max_{\theta\in[0,1]} \theta^k(1-\theta)^{n-k}$. Since the logarithm is an increasing function, the parameter θ maximizing $p_\theta(u^n)$ is the same as the parameter maximizing $\log p_\theta(u^n)$. To find the optimal parameter, compute the derivative of the function $g(\theta) = \log\left(\theta^k(1-\theta)^{n-k}\right) = k\log\theta + (n-k)\log(1-\theta)$ and set it to 0. Doing so will give the optimal parameter $\theta^\star = \frac{k}{n}$. It remains to show that this is a maximum by inspecting the sign of the second derivative of $g(\theta)$. We find that $g''(\theta) = -\frac{k}{\theta^2} - \frac{n-k}{(1-\theta)^2} \leq 0$, so that g is concave and hence $\theta^\star$ is indeed a maximum.

Overall, we just proved that

$$\max_{\theta\in[0,1]} \theta^{k(u^n)}(1-\theta)^{n-k(u^n)} = \left(\frac{k}{n}\right)^k \left(1-\frac{k}{n}\right)^{n-k},$$

and plugging this back in Eq. (1) gives the desired result.

(f) (3 pts) Show that for each n, there is a prefix-free code $c_n : \{0,1\}^n \to \{0,1\}^*$ such that, for every $\theta \in [0,1]$ and every $u^n \in \{0,1\}^n$,

$$\text{length}\, c_n(u^n) \leq \log\frac{1}{p_\theta(u^n)} + \log(1+n) + 1.$$

Hint: Use (d) and (e).

Solution: We know from part (d) that given a family of distribution $(p_\theta)_{\theta\in\Theta}$, designing a Shannon code with the probability distribution given by $\max_{\phi\in\Theta} p_\phi(u)/K$, where $K = \sum_{u\in\mathcal{U}} \max_{\phi\in\Theta} p_\phi(u)$ will give codewords lengths such that

$$\begin{aligned}\text{length}(c(u)) &\leq \log\sum_{u\in\mathcal{U}} \max_{\theta\in\Theta} p_\theta(u) + \log\left(\frac{1}{p_\theta(u)}\right) + 1\\ &= \log\left(\sum_{u\in\mathcal{U}} \max_{\theta\in\Theta} p_\theta(u)\right) + \log\left(\frac{1}{p_\theta(u)}\right) + 1.\end{aligned}$$

In this part, the family of distributions is given by $(p_\theta)_{\theta\in[0,1]}$ where $p_\theta(u^n) = \theta^{k(u^n)}(1-\theta)^{n-k(u^n)}$ and $k(u^n)$ is the number of ones in u^n. For this family of distributions, the

Shannon coding mentioned above is such that

$$\operatorname{length}(c(u^n)) \leq \log\left(\sum_{u^n\in\{0,1\}^n} \max_{\theta\in[0,1]} \theta^{k(u^n)}(1-\theta)^{n-k(u^n)}\right) + \log\left(\frac{1}{p_\theta(u^n)}\right) + 1$$

and with our result from part (e) we can write

$$\operatorname{length}(c(u^n)) \leq \log\left(\sum_{k=0}^{n} \binom{n}{k} \left(\frac{k}{n}\right)^k \left(1-\frac{k}{n}\right)^{n-k}\right) + \log\left(\frac{1}{p_\theta(u^n)}\right) + 1. \quad (2)$$

We see from here that we can reach the desired result if we can show

$$\sum_{k=0}^{n} \binom{n}{k} \left(\frac{k}{n}\right)^k \left(1-\frac{k}{n}\right)^{n-k} \leq n+1.$$

To do so, we upper bound each term of that last sum. Since for any $0 \leq k \leq n$, the term $\binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k}$ can be seen as the probability of a binomial random variable with parameter k/n being equal to k (or the probability of observing k heads out of n i.i.d. fair coin tosses), it is upper bounded by 1. Hence,

$$\sum_{u^n\in\{0,1\}^n} \max_{\theta\in[0,1]} \theta^{k(u^n)}(1-\theta)^{n-k(u^n)} \leq n+1,$$

and using this back in Eq. (2) and dividing by n gives the desired result.

Remarks: Normalizing the result in part (f) by dividing by n, we have $\frac{1}{n}\operatorname{length} c_n(u^n) \leq \frac{1}{n}\log\frac{1}{p_\theta(u^n)} + \frac{1}{n}\log(1+n) + \frac{1}{n}$. As $n\to\infty$, the last two terms go to zero, and we obtain that the lengths of the codewords are nearly $\log\frac{1}{p_\theta(u^n)}$, which is what we would have chosen had we known the parameter θ. Thus, this result shows universal compression in a "point-wise" sense — not only can we make the *average lengths* equal to the optimal average without knowing the distribution $\left(\text{as } \mathbb{E}\left[\frac{1}{n}\log\frac{1}{p_\theta(U^n)}\right] = H(U)\right)$, but also for *each sequence* u^n. By a tighter analysis, the $\log(1+n)$ term can be improved to $\log\left(1+\sqrt{\frac{\pi}{2}n}\right) \approx \frac{1}{2}\log(1+n)$.