Problem Set 7(**Graded Homework - To be Submitted on Dec 23**)

For the Exercise Sessions on Dec 02, Dec 09 and Dec 16

Last name	First name	SCIPER Nr	Points

Problem 1: Exponential Families and Maximum Entropy 1

Let $Y = X_1 + X_2$. Find the maximum entropy of Y under the constraint $\mathbb{E}[X_1^2] = P_1$, $\mathbb{E}[X_2^2] = P_2$:

(a) If X_1 and X_2 are independent.

(b) If X_1 and X_2 are allowed to be dependent.

Solution 1. *(a)* If X_1 and X_2 are independent,

$$\mathrm{Var}[Y] = \mathrm{Var}[X_1 + X_2] = \mathrm{Var}[X_1] + \mathrm{Var}[X_2] \le \mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] = P_1 + P_2 \tag{1}$$

where equality holds when $\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$. Thus we have

$$\max_{f(y)} h(Y) \quad \le \quad \frac{1}{2}\log(2\pi e(P_1 + P_2)) \tag{2}$$

where equality holds when Y is Gaussian with zero mean, which requires X_1 and X_2 to be independent and Gaussian with zeros mean.

(b) For dependent X_1 and X_2, we have

$$\mathrm{Var}(Y) \le \mathbb{E}[Y^2] = \mathbb{E}[(X_1 + X_2)^2] = \mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] + 2\mathbb{E}[X_1X_2] \le (\sqrt{P_1} + \sqrt{P_2})^2 \tag{3}$$

where the first equality holds when $\mathbb{E}[Y] = \mathbb{E}[X_1] + \mathbb{E}[X_2] = 0$, and the send equality holds when $X_2 = \sqrt{\frac{P_2}{P_1}}X_1$. Hence, $\max_{f(y)} h(Y) \le \frac{1}{2}\log(2\pi e(\sqrt{P_1} + \sqrt{P_2})^2)$, where equality holds when Y is Gaussian with zero mean, which requires X_1 and X_2 to be Gaussian with zero mean and $X_2 = \sqrt{\frac{P_2}{P_1}}X_1$.

Problem 2: Exponential Families and Maximum Entropy 2

Find the maximum entropy density f, defined for $x \ge 0$, satisfying $\mathbb{E}[X] = \alpha_1$, $\mathbb{E}[\ln X] = \alpha_2$. That is, maximize $-\int f \ln f$ subject to $\int xf(x)dx = \alpha_1$, $\int(\ln x)f(x)dx = \alpha_2$, where the integral is over $0 \le x < \infty$. What family of densities is this?

Solution 2. The maximum entropy distribution subject to constraints

$$\int xf(x)dx = \alpha_1 \tag{4}$$

and

$$\int (\ln x) f(x) dx = \alpha_2 \tag{5}$$

is of the form

$$f(x) = e^{\lambda_0 + \lambda_1 x + \lambda_2 \ln x} = c x^{\lambda_2} e^{\lambda_1 x} \tag{6}$$

which is of the form of a Gamma distribution. The constants should be chosen so as to satisfy the constraints. We need to solve the following equations

$$\int_0^\infty f(x) dx = \int_0^\infty c x^{\lambda_2} e^{\lambda_1 x} dx = 1 \tag{7}$$

$$\int_0^\infty x f(x) dx = \int_0^\infty c x^{\lambda_2 + 1} e^{\lambda_1 x} dx = \alpha_1 \tag{8}$$

$$\int_0^\infty (\ln x) f(x) dx = \int_0^\infty c x^{\lambda_2} e^{\lambda_1 x} \ln x dx = \alpha_2 \tag{9}$$

Thus, the Gamma distributions $f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$ with

$$\mathbb{E}[X] = k\theta = \alpha_1 \qquad \mathbb{E}[\ln X] = \psi(k) + \ln(\theta) = \alpha_2 \tag{10}$$

is the exponential family we want.

Problem 3: Exponential Families and Maximum Entropy 3

For $t > 0$, consider a family of distributions supported on $[t, +\infty]$ such that $\mathbb{E}[\ln X] = \frac{1}{\alpha} + \ln t$, $\alpha > 0$.

1. What is the parametric form of a maximum entropy distribution satisfying the constraint on the support and the mean?
2. Find the exact form of the distribution.

Solution 3. (i) The maximum entropy distribution has the parametric form $e^{\theta \ln x - A(\theta)} = x^\theta e^{-A(\theta)}$.

(ii) Let us first find the value of $A(\theta)$ from the density constraint $\int_t^\infty x^\theta e^{-A(\theta)} dx = 1$. This gives $e^{-A(\theta)} = -\frac{\theta+1}{t^{\theta+1}}$.

Next we find θ from the mean constraint $\int_t^\infty x^\theta e^{-A(\theta)} \ln x \, dx = \frac{1}{\alpha} + \ln t$. This gives $\frac{t^{\theta+1}((\theta+1)\ln t - 1)}{t^{\theta+1}(\theta+1)} = \ln t - \frac{1}{\theta+1} = \frac{1}{\alpha} + \ln t$ and therefore $\theta = -(\alpha+1)$. The resulting form of the distribution is

$$p(x) = \frac{\alpha t^\alpha}{x^{\alpha+1}}$$

.

Problem 4: Exponential Families and Maximum Entropy 4: I-projections

Let P denote the zero-mean and unit-variance Gaussian distribution. Assume that you are given N iid samples distributed according to P and let $\hat{P}_N$ be the empirical distribution.

Let Π denote the set of distributions with second moment $\mathbb{E}[X^2] = 2$. We are interested in

$$\lim_{N \to \infty} \frac{1}{N} \log \Pr\{\hat{P_N} \in \Pi\} = -\inf_{Q \in \Pi} D(Q \| P).$$

(a) Determine $-\mathrm{arginf}_{Q \in \Pi} D(Q \| P)$, i.e., determine the element Q for which the infinum is taken on.

(b) Determine $-\inf_{Q \in \Pi} D(Q \| P)$.

Solution 4. We are looking for the I-projection of P onto Π, call the result Q. Since Π is a linear family with a single constraint on the expected value of x^2 we know that the density of the minimizing distribution has the form

$$q(x) = p(x)e^{\theta x^2 - A(\theta)}.$$

If we insert $p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ this gives us

$$q(x) = e^{-\frac{x^2}{2} + \theta x^2 - \tilde{A}(\theta)}.$$

We recognize the right-hand side to be the density of a zero-mean Gaussian distribution and by assumption this distribution has second moment 2. Hence, the solution is a zero-mean Gaussian distribution with variance 2, i.e., $q(x) = \frac{1}{\sqrt{4\pi}} e^{-\frac{x^2}{4}}$. The asymptotic exponent is given by the KL distance between these two distributions. We have

$$\begin{aligned} D(q\|p) &= \int \frac{1}{\sqrt{4\pi}} e^{-\frac{x^2}{4}} \log \frac{\frac{1}{\sqrt{4\pi}} e^{-\frac{x^2}{4}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}} dx \\ &= \frac{1}{2} \log \frac{1}{2} + \int \frac{1}{\sqrt{4\pi}} e^{-\frac{x^2}{4}} [-\frac{x^2}{4} + \frac{x^2}{2}] dx \\ &= \frac{1}{2}(\log \frac{1}{2} + 1) = \frac{1}{2}(-\log 2 + 1) \sim 0.153426. \end{aligned}$$

To summarize

1. $-\text{arginf}_{Q\in\Pi} D(Q\|P)$ is given by $q(x) = \frac{1}{\sqrt{4\pi}} e^{-\frac{x^2}{4}}$.
2. $-\inf_{Q\in\Pi} D(Q\|P) = -0.153426$.

Problem 5: Choose the Shortest Description

Suppose $\mathcal{C}_0 : \mathcal{U} \to \{0,1\}^*$ and $\mathcal{C}_1 : \mathcal{U} \to \{0,1\}^*$ are two prefix-free codes for the alphabet $\mathcal{U}$. Consider the code $\mathcal{C} : \mathcal{U} \to \{0,1\}^*$ defined by

$$\mathcal{C}(u) = \begin{cases} [0, \mathcal{C}_0(u)] & \text{if length}\mathcal{C}_0(u) \le \text{length}\mathcal{C}_1(u) \\ [1, \mathcal{C}_1(u)] & \text{else.} \end{cases}$$

Observe that $\text{length}(\mathcal{C}(u)) = 1 + \min\{\text{length}(\mathcal{C}_0(u)), \text{length}(\mathcal{C}_1(u))\}$.

(a) Is $\mathcal{C}$ a prefix-free code? Explain.

(b) Suppose $\mathcal{C}_0, \dots, \mathcal{C}_{K-1}$ are K prefix-free codes for the alphabet $\mathcal{U}$. Show that there is a prefix-free code $\mathcal{C}$ with

$$\text{length}(\mathcal{C}(u)) = \lceil \log_2 K \rceil + \min_{0\le k<K-1} \text{length}(\mathcal{C}_k(u)).$$

(c) Suppose we are told that U is a random variable taking values in $\mathcal{U}$, and we are also told that the distribution p of U is one of K distributions $p_0, \dots, p_{K-1}$, but we do not know which. Using (b) describe how to construct a prefix-free code $\mathcal{C}$ such that

$$\mathbb{E}[\text{length}(\mathcal{C}(U))] \le \lceil \log_2 K \rceil + 1 + H(U).$$

[Hint: From class we know that for each k there is a prefix-free code $\mathcal{C}_k$ that descibes each letter u with at most $\lceil -\log_2 p_k(u) \rceil$ bits.]

Solution 5. (a) Yes, $\mathcal{C}$ is a prefix-free code. We can prove it by contradiction. Suppose there exist $u, v \in \mathcal{U}$ such that $\mathcal{C}(u)$ is a prefix of $\mathcal{C}(v)$. Then they must start with the same bit. Without loss of generality, let us assume they start with 0, then we have $\mathcal{C}(u) = 0\mathcal{C}_0(u)$ is a prefix of $\mathcal{C}(v) = 0\mathcal{C}_0(v)$. This requires $\mathcal{C}_0(u)$ is a prefix of $\mathcal{C}_0(v)$ which contradicts to $\mathcal{C}_0$ is prefix free code.

(b) Generalizing the given construction, we can construct the code $\mathcal{C}(u)$ for any $u \in \mathcal{U}$ as follows.

$$\mathcal{C}(u) = \text{Bin}(i^*)\mathcal{C}_{i^*}(u) \tag{11}$$

where $i^* = \arg\min_{0 \le k \le K-1\}} \text{length}\mathcal{C}_i(u)$ and $\text{Bin}(i^*)$ is the binary representation of number i^*. The length of such code is exactly the given expression and by the same reason in (a), we can show that it is prefix-free.

(c) As the hint suggests, we can use prefix free code $\mathcal{C}_k$ such that $\text{length}(\mathcal{C}_k) \le \lceil -\log_2 p_k(u) \rceil$ and construct the prefix-free code $\mathcal{C}$ as in [b]. Then we have

$$\begin{aligned}
\text{length}(\mathcal{C}(u)) &= \lceil \log_2 K \rceil + \min_{0 \le k < K-1} \text{length}(\mathcal{C}_k(u)) && (12)\\
&\le \lceil \log_2 K \rceil + 1 - \min_{0 \le k < K-1} \log_2 p_k(u) && (13)\\
&\le \lceil \log_2 K \rceil + 1 - \log_2 p(u) && (14)
\end{aligned}$$

Taking expectation at both sides, we get that

$$\mathbb{E}[\text{length}(\mathcal{C}(U))] \le \lceil \log_2 K \rceil + 1 + H(U). \tag{15}$$

Problem 6: Prediction and coding

After observing a binary sequence $u_1, \ldots, u_i$, that contains $n_0(u^i)$ zeros and $n_1(u^i)$ ones, we are asked to estimate the probability that the next observation, u_{i+1} will be 0. One class of estimators are of the form

$$\hat{P}_{U_{i+1}|U^i}(0|u^i) = \frac{n_0(u^i) + \alpha}{n_0(u^i) + n_1(u^i) + 2\alpha} \quad \hat{P}_{U_{i+1}|U^i}(1|u^i) = \frac{n_1(u^i) + \alpha}{n_0(u^i) + n_1(u^i) + 2\alpha}.$$

We will consider the case $\alpha = 1/2$, this is known as the Krichevsky–Trofimov estimator. Note that for $i = 0$ we get $\hat{P}_{U_1}(0) = \hat{P}_{U_1}(1) = 1/2$.

Consider now the joint distribution $\hat{P}(u^n)$ on $\{0,1\}^n$ induced by this estimator,

$$\hat{P}(u^n) = \prod_{i=1}^{n} \hat{P}_{U_i|U^{i-1}}(u_i|u^{i-1}).$$

(a) Show, by induction on n that, for any n and any $u^n \in \{0,1\}^n$,

$$\hat{P}(u_1, \ldots, u_n) \ge \frac{1}{2\sqrt{n}} \left(\frac{n_0}{n}\right)^{n_0} \left(\frac{n_1}{n}\right)^{n_1},$$

where $n_0 = n_0(u^n)$ and $n_1 = n_1(u^n)$.

[Hint: if $0 \le m \le n$, then $(1 + 1/n)^{n+1/2} \ge \frac{m+1}{m+1/2}(1 + 1/m)^m$]

(b) Conclude that there is a prefix-free code $\mathcal{C} : \mathcal{U} \to \{0,1\}^*$ such that

$$\text{length}\,\mathcal{C}(u_1, \ldots, u_n) \le n h_2\left(\frac{n_0(u^n)}{n}\right) + \frac{1}{2}\log n + 2,$$

with $h_2(x) = -x \log x - (1-x)\log(1-x)$.

(c) Show that if $U_1, \ldots, U_n$ are i.i.d. Bernoulli, then

$$\frac{1}{n}\mathbb{E}[\text{length}\, \mathcal{C}(U_1, \ldots, U_n)] \leq H(U_1) + \frac{1}{2n}\log n + \frac{2}{n}$$

Solution 6. (a) For $n = 1$, we have $\hat{P}(u_1) = \hat{P}_{U_1}(u_i) = \frac{1}{2}$. If $u_1 = 0$, $n_0(u_1) = 1$ and $n_1(u_1) = 0$. Hence, $\hat{P}(u_1) = \frac{1}{2} = \frac{1}{2\sqrt{n}}(\frac{n_0}{n})^{n_0}(\frac{n_1}{n})^{n_1}$. It is easy to show that for $u_1 = 1$, the inequality still holds with equality.

For $n = k \geq 1$, let's assume that $\hat{P}(u_1, \ldots, u_k) \geq \frac{1}{2\sqrt{k}}\left(\frac{n_0}{k}\right)^{n_0}\left(\frac{n_1}{k}\right)^{n_1}$. For $n = k+1$, it is sufficient to check $u_{k+1} = 0$, as the case $u_{i+1} = 1$ is the same if we also exchange the roles of n_0 and n_1. In this case, $n_0(u^{k+1}) = n_0(u^k) + 1$ and $n_1(u^{k+1}) = n_1(u^k)$.

$$\begin{aligned}
\hat{P}(u_1, \ldots, u_k, 0) &= \hat{P}_{U_{k+1}|U^k}(0|u^k)\hat{P}_{U^k}(u^k) \\
&\geq \frac{n_0(u^k) + \frac{1}{2}}{n_0(u^k) + n_1(u^k) + 1}\frac{1}{2\sqrt{k}}\Big(\frac{n_0(u^k)}{k}\Big)^{n_0(u^k)}\Big(\frac{n_1(u^k)}{k}\Big)^{n_1(u^k)} \\
&= \underbrace{\frac{(k+1)^{k+1/2}}{k^{k+1/2}}\frac{(n_0(u^k)+\frac{1}{2})n_0(u^k)^{n_0(u^k)}}{(n_0(u^k)+1)^{n_0(u^k)+1}}}_{f(u^k)}\frac{1}{2\sqrt{k+1}}\left(\frac{n_0(u^{k+1})}{k+1}\right)^{n_0(u^{k+1})}\left(\frac{n_1(u^{k+1})}{k+1}\right)^{n_1(u^{k+1})}
\end{aligned}$$

We need to show that $f(u^k) \geq 1$ for any $u^k \in \{0,1\}^k$, but this follows from the hint. Therefore, we proved that our induction hypothesis is true for any $n = k+1$, given the condition that $n = k$ cases is satisfied. By induction, we have for any integer $n \geq 1$

$$\hat{P}(u_1, \ldots, u_n) \geq \frac{1}{2\sqrt{n}}\Big(\frac{n_0}{n}\Big)^{n_0}\Big(\frac{n_1}{n}\Big)^{n_1},$$

Proof the hint: We need to show that:

$$\left(1 + \frac{1}{k}\right)^{k+1/2} \geq \underbrace{\frac{n_0(u^k)+1}{n_0(u^k)+\frac{1}{2}}\left(1 + \frac{1}{n_0(u^k)}\right)^{n_0(u^k)}}_{g(n_0(u^k)) = g(n_0)}.$$

Now, consider the function $g(x) = \frac{x+1}{x+\frac{1}{2}}(1+\frac{1}{x})^x$ for $x \geq 1$. Since we have that $n_0(u^k) \leq k$, if $g(x)$ is an increasing function then we would have:

$$\begin{aligned}
g(n_0(u^k)) \leq g(k) = \frac{k+1}{k+\frac{1}{2}}(1+\frac{1}{k})^k &= \frac{k+1}{(k+\frac{1}{2})\sqrt{1+\frac{1}{k}}}(1+\frac{1}{k})^{k+1/2} \\
&= \frac{\sqrt{k(k+1)}}{k+\frac{1}{2}}(1+\frac{1}{k})^{k+1/2} \\
&< \left(1+\frac{1}{k}\right)^{k+1/2},
\end{aligned}$$

and the result would follow (the last inequality is due to $\sqrt{k(k+1)} < \sqrt{k(k+1)+1/4} = k+1/2$). Hence, we just need to show that $g(x)$ is an increasing function, *i.e.* that $\frac{d}{dx}g(x) \geq 0$. A simple way of doing this is by showing that $\ln g(x)$ is an increasing function, which would then imply the result for $g(x)$. If we compute the differentiation of $\ln g(x)$, we get

$$\frac{d}{dx}\ln g(x) = \frac{1}{x+1} - \frac{1}{x+\frac{1}{2}} + \ln\left(1+\frac{1}{x}\right) - \frac{1}{x+1} = \ln(x+1) - \ln x - \frac{1}{x+\frac{1}{2}}$$

Now observe:

$$\ln(x+1) - \ln x = \int_x^{x+1} \frac{1}{u} du = \mathbb{E}\left[\frac{1}{U}\right],$$

where U is a unifom random variable between x and $x+1$. Also,

$$\frac{1}{x+1/2} = \frac{1}{\mathbb{E}[U]}.$$

Thus:

$$\frac{d}{dx}\ln g(x) = \mathbb{E}\left[\frac{1}{U}\right] - \frac{1}{\mathbb{E}[U]}$$

and the positivity of $\frac{d}{dx}\ln g(x)$ follows from the convexity of the function $u \to 1/u$ (and Jensen's inequality).

(b) Consider the code with length function $L(u^n) = \lceil -\log \hat{P}(u^n) \rceil$. We can check that such code satisfies the Kraft Inequity.

$$\sum_{u^n} 2^{-L(u^n)} = \sum_{u^n} 2^{-\lceil -\log \hat{P}(u^n)\rceil} \leq \sum_{u^n} \hat{P}(u^n) = 1$$

Hence, there exists a prefix-free code with length function $L(u^n)$.

$$\begin{aligned}
\text{length}\, \mathcal{C}(u_1,\ldots,u_n) = \lceil -\log \hat{P}(u^n) \rceil &\leq -\log \hat{P}(u^n) + 1 \\
&\leq -\log\left(\frac{1}{2\sqrt{n}}\left(\frac{n_0}{n}\right)^{n_0}\left(\frac{n_1}{n}\right)^{n_1}\right) + 1 \\
&= 2 + \frac{1}{2}\log n + n\left[-\frac{n_0}{n}\log(\frac{n_0}{n}) - \frac{n_1}{n}\log\frac{n_1}{n}\right] \\
&= 2 + \frac{1}{2}\log n + nh_2(\frac{n_0}{n})
\end{aligned}$$

(c) Let $\Pr(U_i = 0) = \theta$, $\forall i \in \{1,\ldots,n\}$. Since $U_1,\ldots,U_n$ are i.i.d, we have $\mathbb{E}[n_0(u^n)] = \sum_{i=1}^n \mathbb{E}[n_0(u_i)] = n\theta$ and $H(U_i) = h_2(\theta)$ for all i.

$$\begin{aligned}
\mathbb{E}[\text{length}\, \mathcal{C}(U_1,\ldots,U_n)] &\leq \mathbb{E}[nh_2(\frac{n_0(u^n)}{n}) + \frac{1}{2}\log n + 2] \\
&= n\mathbb{E}[h_2(\frac{n_0(u^n)}{n})] + \frac{1}{2}\log n + 2 \\
&\leq nh_2(\frac{\mathbb{E}[n_0(u^n)]}{n}) + \frac{1}{2}\log n + 2 \\
&= nh_2(\theta) + \frac{1}{2}\log n + 2 \\
&= nH(U_1) + \frac{1}{2}\log n + 2
\end{aligned}$$

Therefore,

$$\frac{1}{n}\mathbb{E}[\text{length}\, \mathcal{C}(U_1,\ldots,U_n)] \leq H(U_1) + \frac{1}{2n}\log n + \frac{2}{n}$$

Problem 7: Universal codes

Suppose we have an alphabet $\mathcal{U}$, and let Π denote the set of distributions on $\mathcal{U}$. Suppose we are given a family of S of distributions on $\mathcal{U}$, i.e., $S \subset \Pi$. For now, assume that S is finite.

Define the distribution $Q_S \in \Pi$

$$Q_S(u) = Z^{-1} \max_{P\in S} P(u)$$

where the normalizing constant $Z = Z(S) = \sum_u \max_{P\in S} P(u)$ ensures that Q_S is a distribution.

(a) Show that $D(P\|Q) \le \log Z \le \log|S|$ for every $P \in S$.

(b) For any S, show that there is a prefix-free code $\mathcal{C} : \mathcal{U} \to \{0,1\}^*$ such that for any random variable U with distribution $P \in S$,

$$E[\text{length}\,\mathcal{C}(U)] \le H(U) + \log Z + 1.$$

(Note that $\mathcal{C}$ is designed on the knowledge of S alone, it cannot change on the basis of the choice of P.) [Hint: consider $L(u) = -\log_2 Q_S(u)$ as an 'almost' length function.]

(c) Now suppose that S is not necessarily finite, but there is a finite $S_0 \subset \Pi$ such that for each $u \in \mathcal{U}$, $\sup_{P\in S} P(u) \le \max_{P\in S_0} P(u)$. Show that $Z(S) \le |S_0|$.

Now suppose $\mathcal{U} = \{0,1\}^m$. For $\theta \in [0,1]$ and $(x_1, \dots, x_m) \in \mathcal{U}$, let

$$P_\theta(x_1, \dots, x_n) = \prod_i \theta^{x_i}(1-\theta)^{1-x_i}.$$

(This is a fancy way to say that the random variable $U = (X_1, \dots, X_n)$ has i.i.d. Bernoulli θ components). Let $S = \{P_\theta : \theta \in [0,1]\}$.

(d) Show that for $u = (x_1, \dots, x_m) \in \{0,1\}^m$

$$\max_\theta P_\theta(x_1, \dots, x_m) = P_{k/m}(x_1, \dots, x_m)$$

where $k = \sum_i x_i$.

(e) Show that there is a prefix-free code $\mathcal{C} : \{0,1\}^m \to \{0,1\}^*$ such that whenever $X_1, \dots, X_n$ are i.i.d. Bernoulli,

$$\frac{1}{m}\mathbb{E}[\text{length}\,\mathcal{C}(X_1, \dots, X_m)] \le H(X_1) + \frac{1 + \log_2(1+m)}{m}.$$

Solution 7. (a) From the definition $Q_S(u) = Z^{-1}\max_{P\in S} P(u)$, we have $Q_S(u) \ge P(u)/Z$. Hence, $Z \ge P(u)/Q_S(u)$ and

$$D(P\|Q) = \sum_u P(u)\log\frac{P(u)}{Q(u)} \le \sum_u P(u)\log Z = \log Z$$

From $Z = Z(S) = \sum_u \max_{P\in S} P(u)$, we have $Z \le \sum_u \sum_{P\in S} P(u) = \sum_{P\in S}\sum_u P(u) = |S|$. So $\log Z \le \log|S|$.

(b) For any S, we can find a binary code with length function $L(u) = \lceil -\log_2 Q_S(u)\rceil$ for the codeword $\mathcal{C}(u)$. Since the length function of this binary code satisfies the Kraft Inequality,

$$\sum_u 2^{-L(u)} = \sum_u 2^{-\lceil -\log_2 Q_S(u)\rceil} \le \sum_u 2^{\log_2 Q_S(u)} \le \sum_u Q_S(u) = 1$$

there exists a prefix-free code $\mathcal{C}$ with length function $L(u)$. And the expected length of such code can be computed as

$$\begin{aligned}
\mathbb{E}[\text{length}\,\mathcal{C}(U)] = \mathbb{E}[L(U)] &= \mathbb{E}[\lceil -\log_2 Q_S(u)\rceil] \\
&\le \mathbb{E}[1 - \log_2 Q_S(u)] \\
&= 1 + \mathbb{E}[\log_2 \frac{P(u)}{Q_S(u)} + \log_2 \frac{1}{P(u)}] \\
&= 1 + D(P\|Q) + H(U) \\
&\le 1 + \log Z + H(U)
\end{aligned}$$

(c) Similar as we showed in (a),

$$Z(S) = \sum_u \max_{P \in S} P(u) \leq \sum_u \sup_{P \in S} P(u) \leq \sum_u \max_{P \in S_0} P(u) \leq \sum_u \sum_{P \in S_0} P(u) = |S_0|$$

(d) Rewrite the definition of P_θ:

$$P_\theta(x_1, \ldots, x_m) = \prod_i \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_i x_i}(1-\theta)^{\sum_i (1-x_i)} = \theta^k (1-\theta)^{m-k}$$

Thus, $\log P_\theta = k \log \theta + (m-k)\log(1-\theta)$.

Compute the differentiation of $\log P_\theta$ w.r.t θ:

$$\frac{d}{d\theta} \log P_\theta = \frac{k}{\theta} - \frac{m-k}{1-\theta}$$

Set $\frac{d}{d\theta} \log P_\theta = 0$, we get $\hat{\theta} = k/m$. As logarithm is an increasing function, P_θ is maximized when $\log P_\theta$ is maximized.

(e) From (b) we know that there exists a prefix-free code such that

$$\mathbb{E}[\text{length}\, \mathcal{C}(X_1, \ldots, X_m)] \leq H(X_1, \ldots, X_m) + \log Z + 1$$

where $H(X_1, \ldots, X_m) = mH(X_1)$, since they are i.i.d. From (d), we know that $S_0 = \{P_{k/m} : k = \sum_i^m x_i\}$ has the property in (c). Since each x_i is binary, k is an integer between 0 and m. So $|S_0| = m+1$, we have $Z(S) \leq |S_0| = m+1$. Therefore we have

$$\frac{1}{m}\mathbb{E}[\text{length}\, \mathcal{C}(X_1, \ldots, X_m)] \leq H(X_1) + \frac{\log(1+m)+1}{m}$$