Problem Set 5
Distribution Estimation, Property Testing, Exponential Families

For the Exercise Sessions on Nov 5 and 19— **Due: Tue, Nov 25, 10am, on Moodle**

1 Problems for Class

Problem 1: Add-β Estimator

The add-β estimator $q_{+\beta}$ over $[k]$, assigns to symbol i a probability proportional to its number of occurrences plus β, namely,

$$q_i \stackrel{\text{def}}{=} q_i(X^n) \stackrel{\text{def}}{=} q_{+\beta,i}(X^n) \stackrel{\text{def}}{=} \frac{T_i+\beta}{n+k\beta}$$

where $T_i \stackrel{\text{def}}{=} T_i(X^n) \stackrel{\text{def}}{=} \sum_{j=1}^n \mathbf{1}(X_j = i)$. Prove that for all $k \geq 2$ and $n \geq 1$,

$$\min_{\beta \geq 0} r^{l_2^2}_{k,n}(q_{+\beta}) = r^{l_2^2}_{k,n}(q_{+\sqrt{n}/k}) = \frac{1-\frac{1}{k}}{(\sqrt{n}+1)^2}$$

Furthermore, $q_{+\sqrt{n}/k}$ has the same expected loss for every distribution $p \in \Delta_k$.

Solution 1. By definition of variance, $\mathbb{E}(X^2) = V(X) + \mathbb{E}(X)^2$. Hence,

$$\begin{aligned}
\mathbb{E}(p_i - \frac{T_i+\beta}{n+\beta k})^2 &= \frac{1}{(n+k\beta)^2}\mathbb{E}\left(T_i - np_i - \beta(kp_i-1)\right)^2 \\
&= \frac{1}{(n+k\beta)^2}\left(V(T_i) + \beta^2(kp_i-1)^2\right) \\
&= \frac{1}{(n+k\beta)^2}(np_i(1-p_i) + \beta^2(kp_i-1)^2)
\end{aligned}$$

The loss of the add-β estimator for a distribution p is therefore,

$$\mathbb{E}\|p - q_{+\beta}(X^n)\|_2^2 = \sum_{i=1}^k \mathbb{E}\left(p_i - \frac{T_i+\beta}{n+k\beta}\right)^2 = \frac{1}{(n+k\beta)^2}\left(n - \beta^2 k - (n-\beta^2k^2)\sum_{i=1}^k p_i^2\right)$$

The expected l_2^2 loss of an add-β estimator is therefore determined by just the sum of squares $\sum_{i=1}^k p_i^2$ that ranges from $1/k$ to 1. For $\beta \leq \sqrt{n}/k$, the expected loss is maximized when the square sum is $1/k$, and for $\beta \geq \sqrt{n}/k$, when the square sum is 1, yielding

$$r^{l_2^2}_{k,n}(q_{+\beta}) = \max_{p\in\Delta_k} \mathbb{E}\|p - q_{+\beta}(X^n)\|_2^2 = \frac{1}{(n+k\beta)^2}\begin{cases} n(1-\frac{1}{k}) & \text{for } \beta \leq \frac{\sqrt{n}}{k} \\ \beta^2 k(k-1) & \text{for } \beta > \frac{\sqrt{n}}{k} \end{cases}$$

For $\beta \leq \sqrt{n}/k$, the expected loss decreases as β increases, and for $\beta > \sqrt{n}/k$, it increases as β increases, hence the minimum worst-case loss is achieved for $\beta = \sqrt{n}/k$. Furthermore, $q_{+\sqrt{n}/k}$ has the same expected loss for every underlying distribution p.

Problem 2: Uniformity Testing

Let us reconsider the problem of testing against uniformity. In the lecture we saw a particular *test statistics* that required only $O(\sqrt{k}/\epsilon^2)$ samples where ϵ was the ℓ_1 distance.

Let us now derive a test from scratch. To make things simple let us consider the ℓ_2^2 distance. Recall that the alphabet is $\mathcal{X} = \{1, \cdots, k\}$, where k is known. Let U be the uniform distribution on $\mathcal{X}$, i.e., $u_i = 1/k$. Let P be a given distribution with components p_i. Let X^n be a set of n iid samples. A pair of samples (X_i, X_j), $i \neq j$, is said to *collide* if $X_i = X_j$, if they take on the same value.

1. Show that the expected number of collisions is equal to $\binom{n}{2}\|p\|_2^2$.
2. Show that the uniform distribution minimizes this quantity and compute this minimum.
3. Show that $\|p - u\|_2^2 = \|p\|_2^2 - \frac{1}{k}$.

   NOTE: In words, if we want to distinguish between the uniform distribution and distributions P that have an ℓ_2^2 distance from U of at least ϵ, then this implies that for those distributions $\|p\|_2^2 \geq 1/k + \epsilon$. Together with the first point this suggests the following test: compute the number of collisions in a sample and compare it to $\binom{n}{2}(1/k + \epsilon/2)$. If it is below this threshold decide on the uniform one. What remains is to compute the variance of the collision number as a function of the sample size. This will tell us how many samples we need in order for the test to be reliable.
4. Let $a = \sum_i p_i^2$ and $b = \sum_i p_i^3$. Show that the variance of the collision number is equal to

$$\begin{aligned}&\binom{n}{2}a + \binom{n}{2}\left[\binom{n}{2} - \left(1 + \binom{n-2}{2}\right)\right]b + \binom{n}{2}\binom{n-2}{2}a^2 - \binom{n}{2}^2 a^2 \\ =&\binom{n}{2}\left[2b(n-2) + a(1 + a(3-2n))\right]\end{aligned}$$

   by giving an interpretation of each of the terms in the above sum.

   NOTE: If you don't have sufficient time, skip this step and go to the last point.

   For the uniform distribution this is equal to

$$\binom{n}{2}\frac{(k-1)(2n-3)}{k^2} \leq \frac{n^2}{2k}.$$

   NOTE: You don't have to derive this from the previous result. Just assume it.
5. Recall that we are considering the ℓ_2^2 distance which becomes generically small when k is large. Therefore, the proper scale to consider is $\epsilon = \kappa/k$. Use the Chebyshev inequality and conclude that if we have $\Theta(\sqrt{k}/\kappa)$ samples then with high probability the empirical number of collisions will be less than $\binom{n}{2}(1/k + \kappa/(2k))$ assuming that we get samples from a uniform distribution.

NOTE: The second part, namely verifying that the number of collisions is with high probability smaller than $\binom{n}{2}(1/k + \kappa/(2k))$ when we get $\Theta(\sqrt{k}/\kappa)$ samples from a distribution with ℓ_2^2 distance at least κ/k away from a uniform distribution follows in a similar way.

HINT: Note that if p represents a vector with components p_i then $\|p\|_1 = \sum_i |p_i|$ and $\|p\|_2^2 = \sum_i p_i^2$.

Solution 2. 1. There are $\binom{n}{2}$ pairs. For each pair the chance that both values agree is equal to $\sum_i p_i^2 = \|p\|_2^2$.

2. Let u be the vector of length k with all-one entries. Then, by using the Cauchy-Schwartz inequality, $\|p\|_2^2 = \langle p, p\rangle \geq \langle p, u\rangle^2/\langle u, u\rangle = 1/k$.

3. Expanding the expression, we get

$$\|p-u\|_2^2 = \|p\|_2^2 - 2\langle p,u\rangle + \|u\|_2^2 = \|p\|_2^2 - 2/k + 1/k = \|p\|_2^2 - 1/k.$$

4. Recall that in order to count collisions we look at pairs of indices in our samples. Let (i,j), $1 \le i < j \le n$, be one such pair. When computing the variance we are looking at *pairs of pairs*. E.g., (i,j) and (u,v). There are four parts in the expression for the variance. These have the following interpretation. The first part comes from all pairs with *total* overlap, i.e., $(i,j) = (u,v)$. There are $\binom{n}{2}$ such cases. The second part comes from pairs where exactly one index is repeated. The third term comes from pairs with no overlap. And the fourth term is the mean squared so that we convert from the second moment to the variance.

5. By the Chebyshev's inequality, if $C(X^n)$ counts the number of collisions in our sample then, assuming that the sample comes from the uniform distribution,

$$\Pr\{C(X^n) - \binom{n}{2}\frac{1}{k} \ge \binom{n}{2}\frac{\kappa}{2k}\} \le \frac{n^2/(2k)}{\binom{n}{2}^2\frac{\kappa^2}{4k^2}} \le \frac{k}{n^2\kappa^2}.$$

Therefore, as long as n is large compared to $\sqrt{k/\kappa^2}$ the right-hand side goes to zero. In other words, we need $\Theta(\sqrt{k}/\kappa)$ samples.

2 The Homework

Problem 3: l_2 Estimation

Assume that we have two distributions p and q on $\{1,\cdots K\}$. Let $n \in \mathbb{N}$. Let $N_1, N_2 \sim \text{Poi}(n)$ be independent random variables. We are given N iid samples from each, call them $\{X_j\}_{j=1}^{N_1}$ and $\{Y_j\}_{j=1}^{N_2}$, respectively. Let $t_k(X^{N_1})$, $k = 1,\cdots,K$, respectively, $t_k(Y^{N_2})$, denote the empirical counts. E.g.,

$$t_k(x^n) = |\{j \in \{1,\dots n\} : x_j = k\}|.$$

We want to estimate $\|p-q\|_2^2$.

Define $Z = \sum_{k=1}^K (t_k(X^{N_1}) - t_k(Y^{N_2}))^2 - t_k(X^{N_1}) - t_k(Y^{N_2})$. We claim that Z/n^2 is a good estimator for $\|p-q\|_2^2$.

(a) Show that Z is an unbiased estimator of $n^2\|p-q\|_2^2$.

Hint: The expression for Z should look somewhat familiar. The notes are your best friend.

(b) Assuming that $\|p\|_2^2 \le b$ and $\|q\|_2^2 \le b$ show that the variance of Z can be upper bounded in the following way:

$$\begin{aligned}\text{Var}(Z) &\overset{(i)}{=} \sum_{k=1}^K 4n^3(p_k-q_k)^2(p_k+q_k) + 2n^2(p_k+q_k)^2 \\ &\overset{(ii)}{\le} \sum_{k=1}^K 8n^3(p_k-q_k)^2 + 2n^2(p_k^2+q_k^2+2p_kq_k) \\ &\overset{(iii)}{\le} 8n^3\|p-q\|_2^2 + 8n^2b.\end{aligned}$$

Justify each of the three steps.

Hint: Define $R = (U-V)^2 - U - V$, where $U \sim \text{Poi}(\lambda)$ and $V \sim \text{Poi}(\mu)$. A straighforward but tedious calculation shows that $\text{Var}(R) = 4(\lambda-\mu)^2(\lambda+\mu) + 2(\lambda+\mu)^2$.

(c) Show that $\mathbb{P}\{|Z/n^2 - \|p-q\|_2^2| \ge \epsilon\} \le \frac{8n\|p-q\|_2^2+8b}{n^2\epsilon^2}$.

Solution 3.

(a) Define $Z_k = (t_k(X^{N_1}) - t_k(Y^{N_2}))^2 - t_k(X^{N_1}) - t_k(Y^{N_2})$. Note that $t_k(X^{N_1})$ is distributed according to Poi(np_k) and that $t_k(Y^{N_2})$ is distributed according to Poi(nq_k). Hence,

$$\begin{aligned}\mathbb{E}[Z_k] &= \mathbb{E}[(t_k(X^{N_1}) - t_k(Y^{N_2}))^2 - t_k(X^{N_1}) - t_k(Y^{N_2})] \\ &= \mathbb{E}[t_k(X^{N_1})(t_k(X^{N_1}) - 1) + t_k(Y^{N_2})(t_k(Y^{N_2}) - 1) - 2t_k(X^{N_1})t_k(Y^{N_2})] \\ &= n^2p_k^2 + n^2q_k^2 - 2n^2p_kq_k \\ &= n^2(p_k - q_k)^2,\end{aligned}$$

where in the one-before-last line we have used the fact that for a Poisson random variable of parameter λ, $\mathbb{E}[X(X-1)] = \lambda^2$ as well as the independence of $t_k(X^{N_1})$ and $t_k(Y^{N_2})$.

Since $Z = \sum_{k=1}^K Z_k$, it follows that $E[Z] = n^2 \sum_{k=1}^K (p_k - q_k)^2 = n^2\|p - q\|_2^2$, as claimed.

(b)

$$\begin{aligned}\mathrm{Var}(Z) &\overset{(i)}{=} \sum_{k=1}^K 4n^3(p_k - q_k)^2(p_k + q_k) + 2n^2(p_k + q_k)^2 \\ &\overset{(ii)}{\leq} \sum_{k=1}^K 8n^3(p_k - q_k)^2 + 2n^2(p_k^2 + q_k^2 + 2p_kq_k) \\ &\overset{(iii)}{\leq} 8n^3\|p - q\|_2^2 + 8n^2b.\end{aligned}$$

To see step (i) note that we have

$$\mathrm{Var}(Z) = \sum_{k=1}^K \mathrm{Var}(Z_k)$$

since the random variables Z_k are independent. For step (ii) we use $p_k + q_k \leq 2$. For step (iii) use the bounds $\|p\|_2^2 \leq b$ and $\|q\|_2^2 \leq b$ as well as $\sum_{k=1}^K p_kq_k \leq \|p\|_2\|q\|_2 \leq \beta$ (Cauchy Schwartz).

(c) This is just the Chebyshev inequality.

Problem 4: Fisher Goes Exponential

Let $p_\theta(x)$ denote a family of distributions parameterized by θ. Define the Fisher information as

$$I_\theta = \mathbb{E}_\theta[\nabla_\theta \log p_\theta(X)(\nabla_\theta \log p_\theta(X))^T].$$

(a) Let $p_\theta(x) = h(x)e^{\langle\theta,\phi(x)\rangle - A(\theta)}$ be an exponential family. What is the Fisher information in terms of the parameters of the family?

(b) Consider distributions of the form $p_\lambda(x) = \lambda e^{-\lambda x}$, where $\lambda \in \mathbb{R}^+$.

1. Write it in the form of an exponential family.
2. What is $\Theta = \{\theta \in \mathbb{R} : A(\theta) < \infty\}$.
3. Is the family regular?
4. Is it minimal?
5. What is the Fisher information?

(c) Consider distributions of the form $p_p(k) = (1-p)^k p$, where $p \in (0,1)$ and $k \in \mathbb{N}$.

1. Write it in the form of an exponential family.
2. What is $\Theta = \{\theta \in \mathbb{R} : A(\theta) < \infty\}$.
3. Is the family regular?
4. Is it minimal?
5. What is the Fisher information?

Solution 4. (1) We know from the notes that the Fisher information can also be written as $-\mathbb{E}_\theta[\nabla_\theta^2 \log p_\theta(X)]$. This shows that $I_\theta = \nabla_\theta^2 A(\theta)$.

Alternatively, full score also given for showing one of the following equivalent statements: $I_\theta = \mathbb{E}[\phi(x)\phi(x)^\top] - \mathbb{E}[\phi(x)]\mathbb{E}[\phi(x)]^\top$, $I_\theta = \mathrm{Cov}(\phi(x))$, $I_\theta = \mathbb{E}[(\phi(x) - \mathbb{E}[\phi(x)])(\phi(x) - \mathbb{E}[\phi(x)])^\top]$. (Note that rewriting $\mathbb{E}[\phi(x)] = \nabla_\theta A(\theta)$ is also possible)

(2)
1. $p_\Theta(x) = e^{\Theta\phi(x) - \log(1/\Theta)}$ with $h(x) = 1$, $\theta = \lambda$, $\phi(x) = -x$, and $A(\theta) = \log(1/\theta)$,
2. $\Theta = \{\theta > 0\}$
3. The family is regular since the region Θ is open.
4. Yes, the family is minimal.
5. The Fisher information is $\frac{\partial^2 A(\theta)}{\partial \theta^2} = \frac{\partial^2 \log(1/\theta)}{\partial \theta^2} = \frac{1}{\theta^2}$.

(3)
1. $p_\theta(k) = e^{\theta\phi(k) - A(\theta)}$ with $h(k) = 1$, $\theta = \log(1-p)$, $\phi(k) = k$, and $A(p) = \log(1/p)$ so that $p = 1 - e^\theta$ and $A(\theta) = \log(1/(1-e^\theta))$,
2. We have $\Theta = \{\theta < 0\}$.
3. The family is regular, since Θ is not open.
4. Yes, the family is minimal.
5. The Fisher information is $\frac{\partial^2 A(\theta)}{\partial \theta^2} = \frac{\theta^2 \log(1/(1-e^\theta))}{\partial \theta^2} = \frac{e^\theta}{(1-e^\theta)^2} = (1-p)/p^2$.

Problem 5: Exponential Families and Conjugate Priors

Let $p_\theta(x) = h(x)e^{\langle \phi(x), \theta\rangle - A(\theta)}$ denote a generic exponential family with sufficient statistics $\phi(x)$ and parameter θ.

Assume that we receive iid samples from this family, call them $\{x_i\}_{i=1}^n$. From these samples, we want to infer the unknown parameter θ via a maximum a-posteriori (MAP) procedure.

In order to apply a MAP procedure we need to define a prior distribution on the parameter θ. Consider the family of prior distributions $q_{\mu,\lambda}(\theta) = K(\mu,\lambda)e^{\langle \theta, \mu\rangle - \lambda A(\theta)}$, parametrized by (μ, λ). Note that this is also an exponential family. However, we have written it in a slightly non-standard form, where $K(\mu,\lambda)$ denotes the normalization constant which is a function of the parameters (μ,λ).

(i) Write down the posterior distribution $p_{\mu,\lambda}(\theta \mid x_1, \cdots, x_n)$ for a fixed set of parameters (μ,λ).

(ii) If you have not already done so in part (i), write the posterior as explicitly and compactly as you can. Justify why we called $q_{\mu,\lambda}(\theta)$ a conjugate prior.

(iii) Derive the MAP estimator of the parameter θ given the samples $\{x_i\}_{i=1}^n$ starting with the posterior derived in (ii). When will the estimate be unique?

Solution 5.

(i)/(ii) We have (where in the following Z denotes a normalization constant, not necessarily always the

same):

$$\begin{aligned} p_{\mu,\lambda}(\theta \mid x_1, \cdots, x_n) &= \frac{p_{\mu,\lambda}(\theta) p(x_1, \cdots, x_n \mid \theta)}{p(x_1, \cdots, x_n)} \\ &= \frac{1}{Z} K(\mu, \lambda) e^{\langle \theta, \mu \rangle - \lambda A(\theta)} \prod_{i=1}^{n} h(x_i) e^{\langle \phi(x_i), \theta \rangle - A(\theta)} \\ &= \frac{1}{Z} e^{\langle \theta, \mu + \sum_{i=1}^{n} \phi(x_n) \rangle - (\lambda + n) A(\theta)} \\ &= K(\mu + \sum_{i=1}^{n} \phi(x_n), \lambda + n) e^{\langle \theta, \mu + \sum_{i=1}^{n} \phi(x_n) \rangle - (\lambda + n) A(\theta)} \\ &= q_{\mu + \sum_{i=1}^{n} \phi(x_n), \lambda + n}(\theta) \end{aligned}$$

In the first step we used Bayes rule. In the second step we plugged in the various expressions, keeping in mind that $p(x_1, \cdots, x_n)$ only influences the normalization and can therefore be omitted. In the third step we consolidated the expression. In the fourth and firths step we take into account the resulting expression has the same form as the prior but just with different parameters.

The chosen prior is a conjugate prior since the posterior is again a member of the exponential family.

(iii) In order to find the MAP estimate we have to find the θ that maximes $p_{\mu,\lambda}(\theta \mid x_1, \cdots, x_n)$. Let $\tilde{\mu} = \mu + \sum_{i=1}^{n} \phi(x_n)$ and $\tilde{\lambda} = \lambda + n$. Taking the gradient wrt to the parameter θ and setting the result to 0 we arrive at

$$\nabla_\theta e^{\langle \theta, \tilde{\mu} \rangle - \tilde{\lambda} A(\theta)} = e^{\langle \theta, \tilde{\mu} \rangle - \tilde{\lambda} A(\theta)} (\tilde{\mu} - \tilde{\lambda} \nabla_\theta A(\theta)) = 0.$$

The solution is therefore a θ^* so that $\nabla_\theta A(\theta^*) = \mathbb{E}_{X \sim p_{\theta^*}(x)}[\phi(X)] = \frac{\tilde{\mu}}{\tilde{\lambda}} = \frac{\mu + \sum_{i=1}^{n} \phi(x_n)}{\lambda + n}$. If the family $p_\theta(x)$ is minimal, there will be a unique such value θ^*.

3 Additional Problems

Problem 6: Exponential Families and Maximum Entropy: 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 $-\text{arginf}_{Q \in \Pi} D(Q \| P)$, i.e., determine the element Q for which the infinum is taken on.

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

Solution 6. 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 7: Exponential Families and Maximum Entropy

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

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

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

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

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

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

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, $\max_{f(y)} h(Y) \leq \frac{1}{2}\log(2\pi e(\sqrt{P_1} + \sqrt{P_2})^2)$, 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 8: Poisson Sampling

Assume that we have given a distribution p on $\mathcal{X} = \{1, \cdots, k\}$. Let X^n denote a sequence of n iid samples. Let $T_i = T_i(X^n)$ be the number of times symbol i appears in X^n. Then

$$\{T_i = t_i\} = \binom{n}{t_i} p_i^{t_i} (1 - p_i)^{n - t_i}.$$

Note that the random variables T_i are *dependent*, since $\sum_i T_i = n$. This dependence can sometimes be inconvenient.

There is a convenient way of getting around this problem. This is called *Poisson* sampling. Let N be a random variable distributed according to a Poisson distribution with mean n. Let X^N be then an iid sequence of N variables distributed according to p.

Conditioned on $N = n$, what is the induced distribution of the Poisson sampling scheme?

Show that

1. $T_i(X^N)$ is distributed according to a Poisson random variable with mean $p_i n$.
2. The $T_i(X^N)$ are independent.

Solution 8. (1) Recall that the pmf of a $\mathrm{Poi}(n)$ is:

$$(N = N^*) = e^{-n} \frac{n^{N^*}}{N^*!}$$

Using the concept of conditional probability, we have

$$\begin{aligned}
(T_i(X^N) = t_i) &= \sum_{N^* \geq t_i} (N = N^*)(T_i = t_i | N = N^*) \\
&= \sum_{N^* \geq t_i} e^{-n} \frac{n^{N^*}}{N^*!} \binom{N^*}{t_i} p_i^{t_i} (1-p_i)^{N^* - t_i} \\
&= e^{-n} \sum_{N^* \geq t_i} \frac{n^{t_i + N^* - t_i}}{N^*!} \frac{N^*!}{t_i!(N^* - t_i)!} p_i^{t_i} (1-p_i)^{N^* - t_i} \\
&= \frac{e^{-n}}{t_i!} (np_i)^{t_i} \sum_{N^* \geq t_i} \frac{n^{N^* - t_i}}{(N^* - t_i)!} (1-p_i)^{N^* - t_i} \\
&= \frac{e^{-n}}{t_i!} (np_i)^{t_i} \sum_{N^* \geq t_i} \frac{(n - np_i)^{N^* - t_i}}{(N^* - t_i)!} \\
&= \frac{e^{-n}}{t_i!} (np_i)^{t_i} e^{n - np_i} \\
&= e^{np_i} \frac{(np_i)^{t_i}}{t_i!}
\end{aligned}$$

where in the second last line, we use the fact that $e^x = \sum_{i \geq 0} \frac{x^i}{i!}$. The resulting probability is the pmf of $\text{Poi}\,(np_i)$.

(2) Here, to prove that the random variables are mutually independent, we need to show that

$$P\left(T_1(X^N) = t_1, T_2(X^N) = t_2, \ldots, T_k(X^N) = t_k\right) = \prod_{i=1}^{k} P(T_i(X^N) = t_i).$$

We compute

$$\begin{aligned}
P\left(T_1(X^N) = t_1, \ldots, T_k(X^N) = t_k\right) &= \sum_{m \geq 0} P\left(T_1(X^N) = t_1, \ldots, T_k(X^N) = t_k | N = m\right) P(N = m) \\
&= P\left(T_1(X^N) = t_1, \ldots, T_k(X^N) = t_k \Big| N = \sum_{i=1}^{k} t_i\right) P\left(N = \sum_{i=1}^{k} t_i\right)
\end{aligned}$$

since the only N for which the probability is non-zero with this choice of $t_1, \ldots, t_k$ is $N = \sum_{i=1}^{k} t_i$. Moreover, we know that $P\left(T_1(X^N) = t_1, \ldots, T_k(X^N) = t_k | N = \sum_{i=1}^{k} t_i\right)$ follows a multinomial distribution. Thus, we get

$$
\begin{aligned}
&P\left(T_1(X^N)=t_1,\ldots,T_k\left(X^N\right)=t_k\Big|N=\sum_{i=1}^{k}t_i\right)\cdot P\left(N=\sum_{i=1}^{k}t_i\right)\\
&=\binom{t_1+\cdots+t_k}{t_1,\ldots,t_k}\prod_{i=1}^{k}\left(p_i^{t_i}\right)\cdot\frac{e^{-n}n^{t_1+\cdots+t_k}}{(t_1+\cdots+t_k)!}\\
&=\frac{(t_1+\cdots+t_k)!}{t_1!t_2!\ldots t_k!}\prod_{i=1}^{k}\left(p_i^{t_i}\right)\cdot\frac{e^{-n}n^{t_1+\cdots+t_k}}{(t_1+\cdots+t_k)!}\\
&=e^{-n}n^{t_1+\cdots+t_k}\prod_{i=1}^{k}\left(\frac{p_i^{t_i}}{t_i!}\right)\\
&=\prod_{i=1}^{k}\frac{e^{-np_i}(np_i)^{t_i}}{t_i!}\\
&=\prod_{i=1}^{k}P(T_i(X^N)=t_i).
\end{aligned}
$$

So the $T_i(X^N)$ are independent.