Problem Set 4

Distribution Estimation and Property Testing

For the Exercise Sessions of Oct 13-29 — **Due: Tue, November 3, 10am, on Moodle**

1 The Homework

The following three problems are the core of this homework. You work out solutions and turn them in. Problem 1 is suitable for the exercise session October 13. Problem 2 is suitable for October 29.

Problem 1: χ^2 Divergence Distance Measure

In class we defined the ℓ_1 distance, the ℓ_2 distance, as well as the KL divergence measure. But there are other distance measures that are important and used in practice.

One of those is the χ^2 divergence distance measure. It is defined as

$$\chi^2(p,q) = \sum_{i=1}^{k} \frac{(p_i - q_i)^2}{q_i}.$$

Recall the definition of the min-max loss for a given distance measure L, an alphabet size of k and assuming we have n iid samples:

$$r_{k,n}^{L} = \min_{q} \max_{p \in \Delta_k} \mathbb{E}_{X^n \sim p}[L(p, q(X^n))].$$

(i) [2 pt] Show that, alternatively, $\chi^2(p,q) = \sum_{i=1}^{k} \frac{p_i^2}{q_i} - 1$.

(ii) [8 pts] Show that for $k \geq 2$ and $n \geq 1$, $r_{k,n}^{\chi^2} \leq \frac{k-1}{n+1}$. [Note: There is also a corresponding lower bound that is fairly close to this upper bound showing that this upper bound is relatively tight, but we will only be concerned with the upper bound.]

HINT: We are looking for an upper bound on the min-max loss. Hence, we are free to consider any particular estimator. The add+1 estimator is your friend. Also, remember that $\frac{\binom{n}{t}}{t+1} = \frac{\binom{n+1}{t+1}}{n+1}$.

Solution 1.

(i) Since $\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} q_i = 1$, we have:

$$\chi^2(p,q) = \sum_{i=1}^{k} \frac{(p_i - q_i)^2}{q_i} = -1 + \sum_{i=1}^{k} \frac{p_i^2}{q_i}.$$

(ii) As suggested by the hint, we will use the add+1 estimator. Hence we have:

$$\begin{aligned} r_{k,n}^L &= \min_q \max_{p\in\Delta_k} \mathbb{E}_{X^n\sim p}[L(p, q(X^n))] \\ &\le \max_{p\in\Delta_k} \mathbb{E}_{X^n\sim p}\left[\chi^2(p, q^{+1}(X^n))\right] \\ &= \max_{p\in\Delta_k} \mathbb{E}_{X^n\sim p}\left[-1 + \sum_{i=1}^k \frac{p_i^2}{\frac{T_i(X^n)+1}{n+k}}\right] \\ &= \max_{p\in\Delta_k} -1 + (n+k)\sum_{i=1}^k p_i^2 \mathbb{E}_{X^n\sim p}\left[\frac{1}{T_i(X^n)+1}\right]. \end{aligned}$$

Now note that

$$\begin{aligned} p_i^2 \mathbb{E}_{X^n\sim p}\left[\frac{1}{T_i(X^n)+1}\right] &= p_i^2 \sum_{t=0}^n \frac{1}{t+1}\binom{n}{t} p_i^t (1-p_i)^{n-t} \\ &= \frac{p_i}{n+1}\sum_{t=0}^n \binom{n+1}{t+1} p_i^{t+1}(1-p_i)^{n-t} \\ &= \frac{p_i}{n+1}\sum_{t=1}^{n+1} \binom{n+1}{t} p_i^{t}(1-p_i)^{n+1-t} \\ &= \frac{p_i(1-(1-p_i)^{n+1})}{n+1} \le \frac{p_i}{n+1}, \end{aligned}$$

where the first equality is true since $T_i(X^n)$ is distributed according to a binomial random variable with parameter p_i and hence the probability of $T_i(X^n)$ being t is equal to $\binom{n}{t}p_i^t(1-p_i)^{n-t}$.

Therefore,

$$r_{k,n}^L \le \max_{p\in\Delta_k} \mathbb{E}_{X^n\sim p}[\chi^2(p, q^{+1}(X^n))] \le -1 + (n+k)\sum_{i=1}^k \frac{p_i}{n+1} = \frac{k-1}{n+1}.$$

The right-hand side is independent of p and hence this expression is also the desired upper bound.

Problem 2: Property Testing: Variance

A colleague claims to have implemented an algorithm which outputs i.i.d. samples distributed according to a discrete distribution P that has unit variance. Your task is to design a statistic to test whether this is indeed true.

Let Δ_k be the set of probability distributions on the alphabet $\mathcal{X} = \{1, \cdots, k\}$. Assume that $P \in \mathcal{P} \cup \mathcal{Q}$ with $\mathcal{P} := \{P \in \Delta_k : P \text{ has variance } 1\}$ and $\mathcal{Q} := \{P \in \Delta_k : P \text{ has variance} \in [0, 1-\epsilon] \cup [1+\epsilon, \infty)\}$, where $0 < \epsilon < 1$. You are given n samples $\{X_i\}_{i=1}^n$, where the X_i are independent copies sampled according to P.

Remark: For the following three questions we do not ask you to write down a proof (or explicit calculation) that your proposed solution works.

a) We say that an estimator $e : S^n \mapsto \Pi$ on a sample S^n of length n (ϵ, δ)-learns a parameter $p \in \Pi$ if for any $(\epsilon, \delta) \in (0,1)^2$, given sufficiently many samples n, we have that $\mathbb{P}(\{|e(S^n) - p| > \epsilon\}) < \delta$. Give a brief explanation (one sentence, no calculations) why the empirical estimator of the second moment $\hat{\mu}_{X^2} := \frac{1}{n}\sum_{i=1}^n X_i^2$ can (ϵ, δ)-learn the second moment in our setting.

b) First, assume that a genie tells you that X has zero mean. Design a simple test statistic and give a threshold in order to check for the above mentioned unit variance property.

Hint: Use the claim in a).

c) Now consider the more general case where X can have arbitrary mean. Again, design a simple test statistic and give a threshold.

Hint: You can assume that $\hat{\mu}_X^2$ (ϵ, δ)*-learns* $\mathbb{E}[X]^2$.

Solution 2. a) The empirical mean of squares $\hat{\mu}_{X^2}$ concentrates strongly around the true second moment due to the boundedness of X^2 together with Hoeffding's inequality.

Alternatively: finite expectation of distribution P + (weak) law of large numbers.

b) The approach is analogous to the property test for uniformity as presented in the lecture. We first note that the variance is equal to the second moment. Compute the empirical mean of X^2, i.e., $\hat{\mu}_{X^2} := \frac{1}{n}\sum_{i=1}^n X_i^2$ with sufficiently many samples such that we have that $|\hat{\mu}_{X^2} - var(X)| < \epsilon/2$ with high probability due to Hoeffding's inequality (which applies due to the boundedness of X^2). Then consider the statistic $T(S) := |\hat{\mu}_{X^2} - 1|$, compare it to the threshold $\tau = \epsilon/2$ (with ϵ as given in the definition of $\mathcal{Q}$) and choose $\mathcal{P}$ if we are below the threshold and $\mathcal{Q}$ else. If $P \in \mathcal{P}$, by assumption $|\hat{\mu}_{X^2} - 1| < \epsilon/2$ with high probability and we successfully pick $\mathcal{P}$. If $P \in \mathcal{Q}$, then by assumption $|var(X) - 1| > \epsilon$ and furthermore $|\hat{\mu}_{X^2} - var(X)| < \epsilon/2$. Due to the triangle inequality we have that $|\hat{\mu}_{X^2} - 1| \geq |var(X) - 1| - |\hat{\mu}_{X^2} - var(X)| > \epsilon/2$ with high probability and we correctly select $\mathcal{Q}$. (*Remark: we did not expect the above derivation in the case* $P \in \mathcal{Q}$ *from you.*)

c) We additionally compute the empirical mean of X, i.e., $\hat{\mu}_X := \frac{1}{n}\sum_{i=1}^n X_i$.

We consider the statistic $T(S) := |\hat{\mu}_{X^2} - \hat{\mu}_X^2 - 1|$, compare it to the threshold $\tau = \epsilon/2$ (with ϵ as given in the definition of $\mathcal{Q}$) and choose $\mathcal{P}$ if we are below the threshold and $\mathcal{Q}$ else. The proof that this strategy works is analogous to the on presented in b).

Remark: the following is a rigorous justification as to why learnability of the first two moments implies learnability of the variance. Again, we did not expect this from you. Assume that we pick sufficiently many samples such that we $(\epsilon/4, \delta/2)$-learn the first two moments. By applying the triangle inequality and union bound we get that

$$\begin{aligned}
\mathbb{P}(\{|\hat{\mu}_{X^2} - \hat{\mu}_X^2 - var(X)| > \epsilon/2\}) &\leq \mathbb{P}(\{|\hat{\mu}_{X^2} - \mu_{X^2}| + |\hat{\mu}_X^2 - \mu_X^2| > \epsilon/2\}) && (1)\\
&\leq \mathbb{P}(\{|\hat{\mu}_{X^2} - \mu_{X^2}| > \epsilon/4\} \cup \{|\hat{\mu}_X^2 - \mu_X^2| > \epsilon/4\}) && (2)\\
&\leq \mathbb{P}((\{|\hat{\mu}_{X^2} - \mu_{X^2}| > \epsilon/4\}) + \mathbb{P}(\{|\hat{\mu}_X^2 - \mu_X^2| > \epsilon/4\}) && (3)\\
&\leq \delta/2 + \delta/2 = \delta && (4)
\end{aligned}$$

hence $\hat{\mu}_{X^2} - \hat{\mu}_X^2$ $(\epsilon/2, \delta)$-learns $var(X)$.

2 Additional Problems

For the next six problems, you do not need to turn in solutions. For the exam, we do expect you know how to solve these problems. So, you may keep these problems around and tackle them in preparation for the Midterm and/or Final exam. Or you may solve them now.

Problem 3: 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 $n \geq 1$,

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

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

Solution 3. 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 $\beta \le \sqrt{n}/k$, the expected loss is maximized when the square sum is $1/k$, and for $\beta \ge \sqrt{n}/k$, when the square sum is 1, yielding

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

For $\beta \le \sqrt{n}/k$, 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 4: Bayes Risk

Consider the task of parameter estimation, where the parameter $\underline{W} \in \mathbb{R}^d$ is distributed uniformly on the ball $\mathcal{W} = \{\underline{w} \in \mathbb{R}^d : \|\underline{w}\|_2 \le a\}$. We get samples $\underline{X}_i = \underline{W} + \underline{Z}_i$ with $\underline{Z}_i \sim \mathcal{N}(0, \sigma^2 I_d)$ independent of $\underline{W}$ for $i = 1, \dots, n$, and let $\ell(\underline{w}, \hat{\underline{w}}) = \|\underline{w} - \hat{\underline{w}}\|_2$ be the loss function.

We define the Bayes risk as $R_B = \min_{\hat{\underline{W}}} \mathbb{E}_{\underline{W},\hat{\underline{W}}}[\ell(\underline{W}, \hat{\underline{W}})]$. Show that the Bayes risk is upperbounded as $R_B \le \sqrt{\frac{\sigma^2 d}{n}}$.

Hint: Use that the estimator $\hat{\underline{W}} = \frac{1}{n}\sum_i \underline{X}_i$ achieves the desired upperbound.

Solution 4. We show the the sample mean estimator $\hat{\underline{W}} = \frac{1}{n}\sum_i \underline{X}_i$ achieves this risk.

First solution: We have

$$
\begin{aligned}
\mathbb{E}_{\underline{W},\underline{\hat{W}}}\left[\|\underline{W}-\underline{\hat{W}}\|_2^2\right] &= \mathbb{E}_{\underline{W},\underline{\hat{W}}}\left[\left\|\underline{W}-\frac{1}{n}\sum_{i=1}^{n}\underline{X_i}\right\|^2\right] \\
&= \mathbb{E}_{\underline{W},\underline{\hat{W}}}\left[\left\|\underline{W}-\frac{1}{n}\sum_{i=1}^{n}\left(\underline{W}+\underline{Z_i}\right)\right\|^2\right] \\
&= \mathbb{E}_{\underline{W},\underline{\hat{W}}}\left[\left\|\underline{W}-\underline{W}-\frac{1}{n}\sum_{i=1}^{n}\underline{Z_i}\right\|^2\right] \\
&= \mathbb{E}_{\underline{\hat{W}}}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\underline{Z_i}\right\|^2\right] \\
&= \mathbb{E}_{\underline{\hat{W}}}\left[\frac{1}{n^2}\sum_{i,j=1}^{n}\underline{Z_i}^\top\underline{Z_j}\right] \\
&= \frac{1}{n^2}\sum_{i,j=1}^{n}\mathbb{E}_{\underline{\hat{W}}}\left[\underline{Z_i}^\top\underline{Z_j}\right] \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{E}_{\underline{\hat{W}}}\left[\underline{Z_i}^\top\underline{Z_i}\right]+\sum_{i\neq j=1}^{n}\mathbb{E}_{\underline{\hat{W}}}\left[\underline{Z_i}^\top\underline{Z_j}\right]\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{E}_{\underline{\hat{W}}}\left[\underline{Z_i}^\top\underline{Z_i}\right]+\sum_{i\neq j=1}^{n}\mathbb{E}_{\underline{\hat{W}}}\left[\underline{Z_i}^\top\right]\mathbb{E}_{\underline{\hat{W}}}\left[\underline{Z_j}\right]\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}_{\underline{\hat{W}}}\left[\underline{Z_i}^\top\underline{Z_i}\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}d\sigma^2 \\
&= \frac{\sigma^2 d}{n}.
\end{aligned}
$$

Alternatively, one can also find this result using

Second solution: Note that conditioned on $\underline{W}=\underline{w}$, we have that $\underline{\hat{W}}\sim\mathcal{N}(\underline{w},(\sigma^2/n)I_d)$. Hence, we get

$$\mathbb{E}_{\underline{W},\underline{\hat{W}}}\left[\|\underline{W}-\underline{\hat{W}}\|_2^2\right] = \mathbb{E}_{\underline{W},\underline{\hat{W}}}\left[(\underline{W}-\underline{\hat{W}})^\top(\underline{W}-\underline{\hat{W}})\right] \tag{5}$$

$$= \mathbb{E}_{\underline{W},\underline{\hat{W}}}\left[\underline{W}^\top\underline{W} - \underline{\hat{W}}^\top\underline{W} - \underline{W}^\top\underline{\hat{W}} + \underline{\hat{W}}^\top\underline{\hat{W}}\right] \tag{6}$$

$$= \mathbb{E}_{\underline{W},\underline{\hat{W}}}\left[\|\underline{W}\|_2^2\right] - \mathbb{E}_{\underline{W},\underline{\hat{W}}}\left[\underline{\hat{W}}^\top\underline{W}\right] - \mathbb{E}_{\underline{W},\underline{\hat{W}}}\left[\underline{W}^\top\underline{\hat{W}}\right] + \mathbb{E}_{\underline{W},\underline{\hat{W}}}\left[\|\underline{\hat{W}}\|_2^2\right] \tag{7}$$

$$= \mathbb{E}_{\underline{W}}\left[\|\underline{W}\|_2^2\right] - \mathbb{E}_{\underline{W}}\mathbb{E}_{\underline{\hat{W}}|\underline{W}}\left[\underline{\hat{W}}^\top\underline{W}\right] - E_{\underline{W}}\mathbb{E}_{\underline{\hat{W}}|\underline{W}}\left[\underline{W}^\top\underline{\hat{W}}\right] + E_{\underline{W}}\mathbb{E}_{\underline{\hat{W}}|\underline{W}}\left[\|\underline{\hat{W}}\|_2^2\right] \tag{8}$$

$$= \mathbb{E}_{\underline{W}}\left[\|\underline{W}\|_2^2\right] - \mathbb{E}_{\underline{W}}\left[\mathbb{E}_{\underline{\hat{W}}|\underline{W}}\left[\underline{\hat{W}}^\top\right]\underline{W}\right] - E_{\underline{W}}\left[\underline{W}^\top\mathbb{E}_{\underline{\hat{W}}|\underline{W}}\left[\underline{\hat{W}}\right]\right] + E_{\underline{W}}\mathbb{E}_{\underline{\hat{W}}|\underline{W}}\left[\|\underline{\hat{W}}\|_2^2\right] \tag{9}$$

$$= \mathbb{E}_{\underline{W}}\left[\|\underline{W}\|_2^2\right] - \mathbb{E}_{\underline{W}}\left[\underline{W}^\top\underline{W}\right] - E_{\underline{W}}\left[\underline{W}^\top\underline{W}\right] + E_{\underline{W}}\mathbb{E}_{\underline{\hat{W}}|\underline{W}}\left[\|\underline{\hat{W}}\|_2^2\right] \tag{10}$$

$$= \mathbb{E}_{\underline{W}}\mathbb{E}_{\underline{\hat{W}}|\underline{W}}\left[\|\underline{\hat{W}}\|_2^2\right] - \mathbb{E}_{\underline{W}}\left[\|\underline{W}\|_2^2\right] \tag{11}$$

$$= \mathbb{E}_{\underline{W}}\mathbb{E}_{\underline{\hat{W}}|\underline{W}}\left[\|\underline{\hat{W}}\|_2^2\right] - \mathbb{E}_{\underline{W}}\left[\|\underline{W}\|_2^2\right] \tag{12}$$

$$= \mathbb{E}_{\underline{W}}\left[\|\underline{W}\|_2^2\right] + \frac{\sigma^2 d}{n} - \mathbb{E}_{\underline{W}}\left[\|\underline{W}\|_2^2\right] \tag{13}$$

$$= \frac{\sigma^2 d}{n}, \tag{14}$$

where from equations (12) to (13) we used that

$$\begin{aligned}
\mathbb{E}_{\underline{W}}\mathbb{E}_{\underline{\hat{W}}|\underline{W}}\left[\|\underline{\hat{W}}\|_2^2\right] &= \mathbb{E}_{\underline{W}}\mathbb{E}_{\underline{\hat{W}}|\underline{W}}\left[\sum_{i=1}^d \hat{W}_i^2\right] \\
&= \mathbb{E}_{\underline{W}}\left[\sum_{i=1}^d \mathbb{E}_{\hat{W}_i|W_i}[\hat{W}_i^2]\right] \\
&= \mathbb{E}_{\underline{W}}\left[\sum_{i=1}^d \left(W_i^2 + \frac{\sigma^2}{n}\right)\right] \\
&= \mathbb{E}_{\underline{W}}\left[\|\underline{W}\|_2^2\right] + \frac{\sigma^2 d}{n}.
\end{aligned}$$

Now that we know that $\mathbb{E}_{\underline{W},\underline{\hat{W}}}\left[\|\underline{W}-\underline{\hat{W}}\|_2^2\right] = \frac{\sigma^2 d}{n}$, we can use Jensen's inequality to get

$$\mathbb{E}_{\underline{W},\underline{\hat{W}}}[\|\underline{W}-\underline{\hat{W}}\|_2]^2 \le \mathbb{E}_{\underline{W},\underline{\hat{W}}}[\|\underline{W}-\underline{\hat{W}}\|_2^2] = \frac{\sigma^2 d}{n} \iff \mathbb{E}_{\underline{W},\underline{\hat{W}}}[\|\underline{W}-\underline{\hat{W}}\|_2] \le \sqrt{\frac{\sigma^2 d}{n}}.$$

Hence we find that

$$R_B = \min_{\underline{\hat{V}}} \mathbb{E}_{\underline{W},\underline{\hat{V}}}[\|\underline{W}-\underline{\hat{V}}\|_2] \le \mathbb{E}_{\underline{W},\underline{\hat{W}}}[\|\underline{W}-\underline{\hat{W}}\|_2] \le \sqrt{\frac{\sigma^2 d}{n}}$$

as desired.

Problem 5: ℓ_1 versus Total Variation

In class we defined the ℓ_1 distance as

$$\|p-q\|_1 = \sum_{i=1}^{k} |p_i - q_i|.$$

Another important distance is the total variation distance $d_{\mathrm{TV}}(p,q)$. It is defined as

$$d_{\mathrm{TV}}(p,q) = \max_{S \subseteq \{1,\cdots,k\}} |\sum_{i \in S}(p_i - q_i)|.$$

Show that if p, q are two probability mass vectors (i.e. elements of the simplex) we have that $d_{\mathrm{TV}}(p,q) = \frac{1}{2}\|p-q\|_1$.

Solution 5. Let $S = \{i \in \{1,\cdots,k\} : p_i \geq q_i\}$. Then

$$|\sum_{i \in S}(p_i - q_i)| = \sum_{i \in S} |p_i - q_i|.$$

And further, since both are probability distributions

$$\begin{aligned}
|\sum_{i \in S}(p_i - q_i)| &= |[\sum_{i \in S} p_i] - \sum_{i \in S} q_i]| \\
&= |1 - [\sum_{i \in \{1,\cdots,k\}\setminus S} p_i] - 1 + [\sum_{i \in \{1,\cdots,k\}\setminus S} q_i]| \\
&= |\sum_{i \in \{1,\cdots,k\}\setminus S} (q_i - p_i)| \\
&= \sum_{i \in \{1,\cdots,k\}\setminus S} |p_i - q_i|.
\end{aligned}$$

It follows that $\|p-q\|_1 \leq 2d_{\mathrm{TV}}(p,q)$. But this also shows that $|\sum_{i \in S}(p_i - q_i)| = 2d_{\mathrm{TV}}(p,q)$, and hence we have equality.

Problem 6: 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

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

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

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

Using the concept of conditional probability, we have

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

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 $\mathrm{Poi}(np_i)$.

(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.

Problem 7: Estimating Support Size

You are attending Balelec. You want to estimate how many people are attending. Let this number be m. Here is a very simple algorithm. You walk around randomly. Every 5 minutes you take a picture of the person who is right next to at this moment. Assume that 5 minutes is sufficiently long so that in this manner you sample participants at Balelec with uniform probability. Assume further that during the whole time you do your experiment no person joins or leaves Balelec.

You do this N times, where N is a Poisson random variable with mean $n = 100$. Once you are done you look at the photos. Assume that in total you have encountered $K = 102$ distinct people. Out of those 102, 100 you have seen only once, one you saw twice, and one you saw three times. Give an estimate of the number of people attending Balelec (the support size of the distribution). Call this number $\hat{m}$. We do not expect a number as answer since the estiate might involve an optimization step which might not be trivial to do by hand. Simplify as far as you can and then write down how you would get final answer.

Hint: Follow your own path or answer the question according to the following steps.

1. Assume that there are m people attending Balelec. Take a specific person at Balelec. Call this person "1". Given the procedure outlined above, what is the probability that this person appears c_1, $c_1 \geq 0$, times on your photos?

2. Now take two specific people. Call them "1" and "2". What is the probability that they appear $\{c_i\}_{i=1}^2$ times on your photos?

3. Now consider all people all Balelec together. Assume as before that each has a specific identity. What is the probability that the m people appear $\{c_i\}_{i=1}^m$ times on your photos?

4. Assume again that m people attend Balelec and also as before that we have the counts $\{c_i\}_{i=1}^m$. But this time we do not know who has what count, i.e., we do not know the identites of the people. All we know is the counts themselves. What is the probability of getting the counts $\{c_i\}_{i=1}^m$? [Note: What we see are the non-zero counts, but since we also assume that we know m, we know in fact all counts.]

5. How can you use the last expression to derive an estimate?

Solution 7. 1. Every specific person is sampled a Poisson number of times with mean $100/m$.

2. If we look at two specific people then their counts are independent due to the Poisson sampling and each count follows again a Poisson distribution with mean $100/m$.

3. In general, all counts are independent random variables and follow a Poisson distribution with mean $100/m$.

4. Now assume that we are again given the counts but do not know identities. Note that in general there are many ways to get the same count.

   By assumption, $K = 102$ people are sampled a non-zero number of times. Out of those 100 you saw exactly once, one person you saw twice, and one person you saw three times.

   Let us write down the likelihood of this observation. Let $\lambda = n/m = 100/m$. The likelihood of the observation of these multiplicities given a particular number m of participants is then equal to

$$\begin{aligned}
&\binom{m}{\underbrace{0,0,\ldots,0}_{m-102 \text{ times}},\underbrace{1,1,\ldots,1}_{100 \text{ times}},2,3}\left(\frac{\lambda^0 e^{-\lambda}}{0!}\right)^{m-102}\left(\frac{\lambda^1 e^{-\lambda}}{1!}\right)^{100}\frac{\lambda^2 e^{-\lambda}}{2!}\frac{\lambda^3 e^{-\lambda}}{3!}\\
=&\frac{m!}{100!(m-102)!2!3!}\frac{\lambda^{105}e^{-\lambda m}}{2!3!}\\
=&\frac{m(m-1)\cdots(m-100)(m-101)}{m^{105}}\frac{100^{105}e^{-100}}{(2!3!)^2 100!}
\end{aligned}$$

   You now get the estimate by maximizing wrt to the parameter m. I.e., we need to maximize

$$\frac{m(m-1)\cdots(m-101)}{m^{105}}.$$

   This is in principle simple but does not give a nice compact solution.