Problem Set 7

For the Exercise Session on Dec 15

Last name	First name	SCIPER Nr	Points

Problem 1: 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 $d_{\mathrm{TV}}(p,q) = \frac{1}{2}\|p-q\|_1$.

Solution

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

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

But also,

$$\begin{aligned} 0 = 1 - 1 &= \sum_i p_i - \sum_i q_i = \sum_i (p_i - q_i) = \sum_{i \in S}(p_i - q_i) + \sum_{i \in S^c}(p_i - q_i) \\ &= \sum_{i \in S}(p_i - q_i) - \sum_{i \in S^c}(q_i - p_i) = \sum_{i \in S}|p_i - q_i| - \sum_{i \in S^c}|q_i - p_i|, \end{aligned}$$

i.e., $\sum_{i \in S}|p_i - q_i| = \sum_{i \in S^c}|p_i - q_i|$. Therefore,

$$\|p-q\|_1 = 2\sum_{i \in S}|p_i - q_i| \leq 2 \max_{S \subseteq \{1,\cdots,k\}} |\sum_{i \in S}(p_i - q_i)| = 2 d_{\mathrm{TV}}(p,q).$$

But it is also clear from this derivation that the inequality is in fact an equality: We cannot do better than selecting all pairs with a positive (or all pairs with a negative) difference.

Problem 2: 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

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

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.

Show that

- $T_i(X^N)$ is distributed according to a Poisson random variable with mean $p_i n$.
- The $T_i(X^N)$ are independent.
- Conditioned on $N = n$, the induced distribution of the Poisson sampling scheme is equal to the distribution of the *original scheme*.

Solution

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

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

Using the concept of conditional probability, we have

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

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) To show the independence, it is enough to show that

$$\begin{aligned}
\mathbb{P}(T_1(X^N)=t_1,\dots,T_k(X^N)=t_k) &= \mathbb{P}\left(N=\sum_i t_i\right)\mathbb{P}\left(T_1(X^N)=t_1,\dots,T_k(X^N)=t_k \middle| N=\sum_i t_i\right)\\
&= e^{-n}\frac{n^{\sum_i t_i}}{(\sum_i t_i)!}\binom{\sum_i t_i}{t_1,\dots,t_k}\prod_i p_i^{t_i}\\
&= e^{-n\sum_i p_i}\frac{n^{\sum_i t_i}}{(\sum_i t_i)!}\frac{(\sum_i t_i)!}{t_1!\dots t_k!}\prod_i p_i^{t_i}\\
&= \prod_i\left[e^{-np_i}\frac{n^{t_i}}{t_i!}p_i^{t_i}\right]\\
&= \prod_i \mathbb{P}(T_i(X^N)=t_i)
\end{aligned}$$

(3) Under the condition that $N=n$,

$$\mathbb{P}(T_i(X^N)=t_i|N=n)=\binom{n}{t_i}p_i^{t_i}(1-p_i)^{n-t_i}=\mathbb{P}(T_i=t_i)$$

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

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,

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

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

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

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 4: 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 $\Theta(\sqrt{k}/\kappa^2)$ 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 no smaller than $\binom{n}{2}(1/k + \kappa/(2k))$ when we get $\Theta(\sqrt{k}/\kappa^2)$ 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

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), $1 \leq i < j \leq n$, 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,

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

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 $\Theta(\sqrt{k}/\kappa^2)$ samples.