Problem Set 5

For the Exercise Session on Nov 7

Last name	First name	SCIPER Nr	Points

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 $p \in \Delta_k$.

Problem 2: 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?

Problem 3: 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 $d_{\mathrm{TV}}(p,q) = \frac{1}{2}\|p-q\|_1$.

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 $\Theta(\sqrt{k}/\kappa)$ 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 $\|p\|_2^2 = \sum_i p_i^2$.