Foundations of Data Science
Gastpar & Urbanke

Ecole Polytechnique Fédérale, Lausanne: Fall 2023
November 6, 2023

# Problem Set 5

For the Exercise Session on Nov 7

| Last name | First name | SCIPER Nr | Points |
|---|---|---|---|
| | | | |

### Problem 1: Add-$\beta$ Estimator

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

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

where $T_i \overset{\text{def}}{=} T_i(X^n) \overset{\text{def}}{=} \sum_{j=1}^{n} \mathbf{1}(X_j = i)$. 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 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-$\beta$ 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-$\beta$ 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 $\beta$ increases, and for $\beta > \sqrt{n}/k$, it increases as $\beta$ 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: 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 2.**

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

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

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.

**Problem 3: $\ell_1$ versus Total Variation**

In class we defined the $\ell_1$ distance as

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

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

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

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

**Solution 3.** 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 $\|p-q\|_1 \leq 2d_{\text{TV}}(p,q)$. But this also shows that $|\sum_{i \in S}(p_i - q_i)| = 2d_{\text{TV}}(p,q)$, and hence we have equality.

**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 $\epsilon$ was the $\ell_1$ distance.

Let us now derive a test from scratch. To make things simple let us consider the $\ell_2^2$ 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 $\ell_2^2$ distance from $U$ of at least $\epsilon$, 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

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

   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 $\ell_2^2$ 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 $\ell_2^2$ distance at least $\kappa/k$ 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 4.** 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.