Foundations of Data Science
Gastpar & Urbanke

Ecole Polytechnique Fédérale, Lausanne: Fall 2024
November 4, 2023

Problem Set 5

For the Exercise Sessions on Nov 21 and Nov 28

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

Problem 3: 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. Thit 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.

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 $\|p\|_2^2 = \sum_i p_i^2$.

Problem 5: James-Stein Estimator (a) Assume that $X \sim \mathcal{N}(0,1)$ and that $f : \mathbb{R} \to \mathbb{R}$ is such that $\mathbb{E}[|Xf(X)|] < \infty$ and $\mathbb{E}[|f'(X)|] < \infty$. Show that

$$\mathbb{E}[Xf(X)] = \mathbb{E}[f'(X)].$$

Hint 1: for the derivative of the probability density function $p(\cdot)$ of a mean zero, unit variance Gaussian random variable it holds that $p'(x) = -xp(x)$.

Hint 2: recall that integration by parts asserts that $\int_a^b u(t)v'(t)dt = u(t)v(t)|_a^b - \int_a^b u'(t)v(t)dt$.

(b) Now assume that $X \sim \mathcal{N}(\mu, \sigma^2)$ and that $f : \mathbb{R} \to \mathbb{R}$ is such that $\mathbb{E}[|(X-\mu)f(X)|] < \infty$ and $\mathbb{E}[|f'(X)|] < \infty$. Re-using the result from (a), show that

$$\mathbb{E}[(X-\mu)f(X)] = \sigma^2\mathbb{E}[f'(X)].$$

For the remainder of the problem, we are concerned with assessing the performance of estimators $\hat{\theta}$ of a mean vector $\theta \in \mathbb{R}^n$, with ℓ_2-loss and corresponding risk $\mathcal{R}(\hat{\theta}) := \mathbb{E}[\|\theta - \hat{\theta}(Z)\|_2^2]$, and with data generated according to $Z := (Z_1, Z_2, \ldots, Z_n) \sim \mathcal{N}(\theta, \sigma^2 I)$.

Assume that we write the estimator in the form $\hat{\theta}(z) = g(z) + z$ with $z = (z_1, \ldots, z_n)$ and $g(z) = (g_1(z), \ldots, g_n(z))$. Consider the expression

$$\hat{\mathcal{R}}(\hat{\theta}, z) = n\sigma^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial g_i(z)}{\partial z_i} + \sum_{i=1}^{n} g_i^2(z).$$

(c) Show that $\hat{\mathcal{R}}(\hat{\theta}, z)$ is an unbiased estimator of the risk, i.e., verify that $\mathbb{E}[\hat{\mathcal{R}}(\hat{\theta}, Z)] = \mathcal{R}(\hat{\theta})$. You can assume without proof that the technical assumptions necessary for the result in (b) are met.

Hint: $(a-b)^2 = (a-c+c-b)^2$ for any c; choosing c cleverly might help you.

The above risk estimator is called *Stein's Unbiased Risk Estimate (SURE).*

We assume from hereon for simplicity that $\sigma^2 = 1$.

In statistical inference, if one has complete knowledge about the data generating model (in our case we know that $Z \sim \mathcal{N}(\theta, \sigma^2 I)$), it is usually a safe bet to do maximum likelihood (ML) estimation. In our setting, the ML estimator is given by the simple identity map $\hat{\theta}_{ML}(z) = z$. It can be proven that for our Gaussian model and with $n = 1$, one cannot "do better" (in some precise technical sense) in terms of ℓ_2-risk than $\hat{\theta}_{ML}$. Encouraged by this fact, let us analyze its performance in the general multi-dimensional case:

(d) Assume $n \in \mathbb{N}^+$. Calculate the risk $\mathcal{R}(\hat{\theta}_{ML})$ of the maximum likelihood estimator.

A historically important result in statistics states that when one tries to jointly estimate multiple parameters ($n > 1$), it can happen that there are methods that perform strictly better than a simple component-wise application of the best scalar ($n = 1$) estimator.

One such example is provided by the James-Stein estimator, which is defined as

$$\hat{\theta}_{JS}(z) = \left(1 - \frac{n-2}{\|z\|_2^2}\right)z.$$

We assume from hereon that $n \geq 3$ (Remark: we do this since for $n = 1$, the technical assumptions necessary for the result in b) are not met; and for $n = 2$, $\hat{\theta}_{JS} = \hat{\theta}_{ML}$ which is not very interesting.).

(e) Using SURE, estimate the risk of the James-Stein estimator, i.e., calculate $\hat{\mathcal{R}}(\hat{\theta}_{JS}, Z)$.

Hint: recall the quotient rule which states that $\left(\frac{u(t)}{v(t)}\right)' = \frac{u'(t)v(t)-u(t)v'(t)}{(v(t))^2}$.

(f) Calculate the risk $\mathcal{R}(\hat{\theta}_{JS})$ – **not** by direct calculation (which is quite tedious) – but by exploiting the unbiasedness of SURE and using the result in (e). How does the risk compare to that of $\hat{\theta}_{ML}$ for $n \geq 3$?