# Problem Set 5

For the Exercise Sessions on Nov 05 and Nov 12

| 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,

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

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 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: $\ell_1$ 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 $d_{\mathrm{TV}}(p,q)$. 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 2.** 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 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.

**Solution 3.** (1) Recall that the pmf of a Poi$(n)$ is:

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

Using the concept of conditional probability, we have

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

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 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

$$\begin{aligned}
P\left(T_1(X^N) = t_1, \ldots, T_k(X^N) = t_k\right) &= \sum_{m \geq 0} P\left(T_1(X^N) = t_1, \ldots, T_k(X^N) = t_k | N = m\right) P(N = m) \\
&= P\left(T_1(X^N) = t_1, \ldots, T_k(X^N) = t_k \middle| N = \sum_{i=1}^{k} t_i\right) P\left(N = \sum_{i=1}^{k} t_i\right)
\end{aligned}$$

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

$$
\begin{aligned}
&P\left(T_1(X^N)=t_1,\ldots,T_k\left(X^N\right)=t_k\middle|N=\sum_{i=1}^k t_i\right)\cdot P\left(N=\sum_{i=1}^k t_i\right)\\
&=\binom{t_1+\cdots+t_k}{t_1,\ldots,t_k}\prod_{i=1}^k\left(p_i^{t_i}\right)\cdot\frac{e^{-n}n^{t_1+\cdots+t_k}}{(t_1+\cdots+t_k)!}\\
&=\frac{(t_1+\cdots+t_k)!}{t_1!t_2!\ldots t_k!}\prod_{i=1}^k\left(p_i^{t_i}\right)\cdot\frac{e^{-n}n^{t_1+\cdots+t_k}}{(t_1+\cdots+t_k)!}\\
&=e^{-n}n^{t_1+\cdots+t_k}\prod_{i=1}^k\left(\frac{p_i^{t_i}}{t_i!}\right)\\
&=\prod_{i=1}^k\frac{e^{-np_i}(np_i)^{t_i}}{t_i!}\\
&=\prod_{i=1}^k P(T_i(X^N)=t_i).
\end{aligned}
$$

So the $T_i(X^N)$ are independent.

**Problem 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

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

   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

$$\binom{n}{2}\frac{(k-1)(2n-3)}{k^2} \le \frac{n^2}{2k}.$$

*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, $\|p\|_2^2 = \langle p, p\rangle \ge \langle p, u\rangle^2/\langle u, u\rangle = 1/k$.

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.

**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 $\ell_2$-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 $\ell_2$-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 $n \geq 3$?

**Solution 5.**

(a) Integrating by parts and using the hint yields

$$\mathbb{E}[f'(X)] = \int_{-\infty}^{\infty} f'(t)p(t)dt = f(t)p(t)\big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} tf(t)p(t)dt.$$

We observe that the first term on the RHS is zero whereas the second term equals $\mathbb{E}[Xf(X)]$.

(b) Define $\tilde{X} := \frac{X-\mu}{\sigma}$ and $\tilde{f}(x) := f(\sigma x + \mu)$. Note that it holds $\tilde{X} \sim \mathcal{N}(0,1)$ and $f(X) = \tilde{f}(\tilde{X})$ and hence $\mathbb{E}[(X-\mu)f(X)] = \sigma\mathbb{E}[\tilde{X}\tilde{f}(\tilde{X})]$. The result now follows from applying (a) and noting that $\tilde{f}'(x) = \sigma f'(x)$ by the chain rule.

(c)

$$\begin{aligned}
\mathcal{R}(\hat{\theta}) &= \sum_{i=1}^{n} \mathbb{E}[(\hat{\theta}_i - \theta_i)^2] \\
&= \sum_{i=1}^{n} \mathbb{E}[(\hat{\theta}_i - Z_i + Z_i - \theta_i)^2] \\
&= \sum_{i=1}^{n} \mathbb{E}[(Z_i - \theta_i)^2] + 2\sum_{i=1}^{n} \mathbb{E}[(\hat{\theta}_i - Z_i)(Z_i - \theta_i)] + \sum_{i=1}^{n} \mathbb{E}[(\hat{\theta}_i - Z_i)^2] \\
&= n\sigma^2 + 2\sum_{i=1}^{n} \mathbb{E}[g_i(Z)(Z_i - \theta_i)] + \sum_{i=1}^{n} \mathbb{E}[(\hat{\theta}_i - Z_i)^2] \\
&= n\sigma^2 + 2\sigma^2\sum_{i=1}^{n} \mathbb{E}[\frac{\partial g_i(Z)}{\partial z_i}] + \sum_{i=1}^{n} \mathbb{E}[(\hat{\theta}_i - Z_i)^2] \qquad (1) \\
&= \mathbb{E}[\hat{\mathcal{R}}(Z)]
\end{aligned}$$

where (1) follows from applying (b) to $\mathbb{E}[g_i(Z)(Z_i - \theta_i)|Z_i]$.

(d) $\mathcal{R}(\hat{\theta}_{ML}) = \mathbb{E}[\|Z - \theta\|^2] = n$.

(e) First note that that the James-Stein's estimator can be written in the form $\hat{\theta}_{JS}(z) = z + g(z)$ with $g(z) = -\frac{n-2}{\|z\|_2^2}z$. Applying SURE, we get

$$\begin{aligned}
\hat{\mathcal{R}}(\hat{\theta}_{JS}) &= n + 2\sum_{i=1}^{n} \frac{-(n-2)\|Z\|_2^2 + (n-2)Z_i \cdot 2Z_i}{\|Z\|_2^4} + \sum_{i=1}^{n} \frac{(n-2)^2}{\|Z\|_2^4} Z_i^2 \\
&= n - \frac{2n(n-2)}{\|Z\|_2^2} + \frac{4(n-2)}{\|Z\|_2^2} + \frac{(n-2)^2}{\|Z\|_2^2} \\
&= n - \frac{(n-2)^2}{\|Z\|_2^2}.
\end{aligned}$$

(f) Next, we get the true risk by recalling that SURE is unbiased. Hence we can get the true risk by simply taking the expectation of SURE:

$$\mathcal{R}(\hat{\theta}_{JS}) = \mathbb{E}[\hat{\mathcal{R}}(\hat{\theta}_{JS})] = n - (n-2)^2\mathbb{E}[\frac{1}{\|Z\|_2^2}].$$

The risk is strictly smaller than $\mathcal{R}(\hat{\theta}_{ML})$ for all $n \geq 3$.