ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

Foundations of Data Science
Fall 2023

Assignment date: Friday, January 19, 2024, 15:15
Due date: Friday, January 19, 2024, 18:15

Final Exam – GCA 330/331

This exam is open book. No electronic devices of any kind are allowed. There are 5 problems. Choose the ones you find easiest and collect as many points as possible. We do not necessarily expect you to finish all of them. Good luck!

Name: _____

Problem 1	/ 10
Problem 2	/ 5
Problem 3	/ 10
Problem 4	/ 12
Problem 5	/ 10
Total	/47

Problem 1. *(Subgaussian)*[10 pts]

In this problem, we develop an alternative proof of the fact that bounded random variables are subgaussian.

(a) [1 pt] Prove (*difficult*) (or simply take for granted and move on...) the following inequality:

$$\cosh(x) = (e^x + e^{-x})/2 \leq e^{x^2/2}.$$

(b) [1 pt] Using the previous inequality, give an upper bound on the moment generating function $\mathbb{E}[e^{\lambda S}]$ of a random variable S that only takes the values $+1$ and -1, with equal probability.

(c) [2 pts] Consider any random variable X and let X' be a random variable *independent of* X, but with exactly the same distribution. Show that

$$\mathbb{E}_X[e^{\lambda(X-\mathbb{E}[X])}] \leq \mathbb{E}_{X,X'}[e^{\lambda(X-X')}].$$

Hint: Start with the right hand side expression and consider for starters only the ("inner") expectation over X'.

(d) [1 pt] Show that the random variables $(X - X')$ and $S(X - X')$, where S is as in Part (b) and assumed independent of X and X', have the same distribution.

(e) [2 pts] From the previous part, we thus know that

$$\mathbb{E}_{X,X'}[e^{\lambda(X-X')}] = \mathbb{E}_{S,X,X'}[e^{\lambda S(X-X')}].$$

Now assume that X is a bounded random variable, $X \in [a, b]$. Condition on $X = x$ and $X' = x'$, and take expectation over S. Observe that $(x - x')^2 \leq (b - a)^2$. Use this and your result from Part (b) to further upper bound $\mathbb{E}_{S,X,X'}[e^{\lambda S(X-X')}]$.

(f) [2 pts] Combine your results to give an upper bound on the moment generating function of a centered bounded random variable $X - \mathbb{E}[X]$, where $X \in [a, b]$.

(g) [1 pt] Compare your result to Lemma 2.4 in the lecture notes. Discuss the differences. *Note: Lemma 2.4 in the lecture notes states that a zero-mean random variable with* $X \in [a, b]$ *is* $(b - a)^2/4$*-subgaussian.*

Solution 1.

(a) To prove this inequality, we can proceed via the expansions of the exponential function. Specifically,

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!} + \dots$$
$$e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} - \frac{x^5}{5!} + \frac{x^6}{6!} - \dots$$

Adding up and dividing by 2,

$$\cosh(x) = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} \dots + \frac{x^{2n}}{(2n)!} + \dots$$

Expanding $e^{x^2/2}$ via the standard expansion for e^y,

$$e^{x^2/2} = 1 + \frac{x^2}{2} + \frac{x^4}{4 \cdot 2!} + \frac{x^6}{8 \cdot 3!} \dots + \frac{x^{2n}}{2^n n!} + \dots$$

A term-by-term comparison and noting that $2^n n! \leq (2n)!$ gives the claimed bound.

(b) Simply write out

$$\mathbb{E}\left[e^{\lambda S}\right] = \sum_s p_S(s) e^{\lambda s} = \frac{1}{2}e^{\lambda} + \frac{1}{2}e^{-\lambda} \leq e^{\lambda^2/2}.$$

(c) Write out using the independence of X and X'

$$\mathbb{E}_{X,X'}[e^{\lambda(X-X')}] = \mathbb{E}_X\left[\mathbb{E}_{X'}[e^{\lambda(X-X')}]\right].$$

Now, for the inner expectation, we apply Jensen's inequality, noting that the exponential function is convex:

$$\mathbb{E}_{X'}[e^{\lambda(X-X')}] \geq e^{\mathbb{E}_{X'}[\lambda(X-X')]} = e^{\lambda(X - \mathbb{E}_{X'}[X'])} = e^{\lambda(X - \mathbb{E}[X])},$$

where we have used the linearity of expectation and the fact that X and X' have the same distribution.

(d) For example, we can argue via the CDF. First, we observe that since X and X' are indistinguishable, we must have for any real number y

$$\mathbb{P}(X - X' \leq y) = \mathbb{P}(X' - X \leq y)$$

But then, by conditioning, we must have for any real number y

$$\begin{aligned}\mathbb{P}(S(X - X') \leq y) &= \mathbb{P}(S = 1)\mathbb{P}((X - X') \leq y) + \mathbb{P}(S = -1)\mathbb{P}(-(X - X') \leq y) \\ &= \mathbb{P}(X - X' \leq y),\end{aligned}$$

which proves the claim.

(e) Following the instruction, we write

$$\mathbb{E}_{S,X,X'}[e^{\lambda S(X-X')}] = \mathbb{E}_{X,X'}\left[\mathbb{E}_S[e^{\lambda S(X-X')}]|X, X'\right].$$

Now, for any fixed values $X = x$ and $X' = x'$, we have, using Part (b),

$$\mathbb{E}_S[e^{\lambda S(x-x')}] \leq e^{\lambda^2(x-x')^2/2} \leq e^{\lambda^2(b-a)^2/2}.$$

Hence,

$$\mathbb{E}_{X,X'}[e^{\lambda(X-X')}] = \mathbb{E}_{S,X,X'}[e^{\lambda S(X-X')}] \leq e^{\lambda^2(b-a)^2/2}.$$

(f) Combining everything, we have

$$\begin{aligned}\mathbb{E}_X[e^{\lambda(X-\mathbb{E}[X])}] &\leq \mathbb{E}_{X,X'}[e^{\lambda(X-X')}] \\ &= \mathbb{E}_{S,X,X'}[e^{\lambda S(X-X')}] \\ &\leq e^{\lambda^2(b-a)^2/2}.\end{aligned}$$

(g) In the lecture notes, we have shown that for a bounded random variable X, the moment generating function satisfies

$$\mathbb{E}_X[e^{\lambda(X-\mathbb{E}[X])}] \leq e^{\frac{\lambda^2(b-a)^2}{8}}.$$

So, the proof above establishes also that bounded random variables are subgaussian, but with a suboptimal parameter: the argument developed here says that bounded random variables are $(b-a)^2$-subgaussian, where with the more intricate argument from your homework, you can actually show that they are $(b-a)^2/4$-subgaussian. Needless to say, for many proofs, these two results are equally interesting, and there is only a small gain to be had from the factor of 4 improvement in the exponent.

Problem 2. *(Entropy and variance)*[5 pts]

Let X be a continuous random variable with density $p(x)$ and support in $\mathbb{R}$. Let $h(X)$ denote the (differential) entropy of X, where (in nats)

$$h(x) = \int -p(x) \ln p(x) dx.$$

Let $\mathrm{Var}(X)$ denote the variance of X, where

$$\mathrm{Var}(X) = \int p(x)(x - \mathbb{E}[X])^2 dx.$$

Assume that you are told that $h(X) \geq h > 0$. What bound can you conclude on $\mathrm{Var}(X)$ in terms of h?

Solution 2.

Recall that the random variable of second moment ρ^2 that has maximum entropy is a Gaussian with mean zero and variance ρ^2. Since the mean does not change the entropy, it is also true that the random variable with a given mean and a given variance that maximizes the entropy is a Gaussian. Further, for a Gaussian of mean μ and variance σ^2 we have

$$\begin{aligned} h(X) &= \int p(x)[\ln(\sqrt{2\pi\sigma^2}) + \frac{(x-\mu)^2}{2\sigma^2}]dx \\ &= \frac{1}{2}\ln(2\pi e\sigma^2). \end{aligned}$$

We conclude that for any random variable we have $h \leq h(X) \leq \frac{1}{2}\ln(2\pi e\mathrm{Var}(X))$. Inverting this inequality, we get that for any random variable we have $\mathrm{Var}(X) \geq \frac{e^{2h}}{2\pi e}$.

Problem 3. *(Johnson-Lindenstrauss)*[10 pts]

(a) [2 pts] In preparation for this problem, establish the following facts:

- If U is a subexponential random variable with parameters (ν, b), then αU (where we assume $\alpha > 0$) is a subexponential random variable with parameters $(\alpha\nu, \alpha b)$.
- If U and V are independent subexponential random variables with parameters (ν_u, b_u) and (ν_v, b_v), respectively, then $U + V$ is a subexponential random variable with parameters $(\sqrt{\nu_u^2 + \nu_v^2}, \max(b_u, b_v))$.

In this problem, we reconsider the Johnson-Lindenstrauss Lemma from Section 10.2.2 of the Lecture Notes. The only change is that inside the real-valued $k \times d$ matrix X in the proof of the Lemma, we no longer assume that the entries are independent Gaussians. We still assume the entries X_{ij} to be independent. We also still assume that they each have mean zero and variance 1. But beyond this, we only assume that they are subgaussian with variance proxy σ^2.

To proceed, exactly as in the lecture notes, consider an arbitrary real-valued vector u of length d. As in the lecture notes, we define, for $i = 1, 2, \ldots, k$,

$$Z_i = \frac{1}{\|u\|_2} \sum_{j=1}^{d} u_j X_{ij}.$$

(b) [2 pts] Show the following facts (short justifications are sufficient, and you may refer freely to the lecture notes)

- The random variables Z_i are independent of each other.
- Each Z_i is subgaussian. Find the corresponding variance proxy.
- We have $\mathbb{E}[Z_i^2] = 1$.

To continue, we will need the following theorem that was alluded to a few times during the lectures:

Theorem. If Y is subgaussian with variance proxy σ^2, then Y^2 with mean $\mathbb{E}[Y^2]$ is subexponential with parameters $(\sqrt{32}\sigma^2, 4\sigma^2)$.

(c) [2 pts] Exactly as in the lecture notes, we next need to analyze $S = \frac{1}{k}\sum_{i=1}^{k} Z_i^2$. Leveraging the theorem, show that S is subexponential with mean 1 and find the corresponding parameters.

(d) [2 pts] Give a concentration bound, that is, an upper bound of the form

$$\mathbb{P}\left\{\left|\frac{1}{k}\sum_{i=1}^{k} Z_i^2 - 1\right| > \delta\right\} \leq \ldots$$

(e) [2 pts] Discuss the differences of the resulting lemma with respect to what is proved in the lecture notes.

Solution 3.

(a) We prove the two facts in turn, noting that the proof argument are identical to the proof of Parts (ii) and (iii) of Lemma 2.1 in the lecture notes:

- First, observe that the mean of αU is simply $\alpha\mu_u$, where μ_u denotes the mean of U. Hence, we need to study

$$\mathbb{E}[e^{\lambda(\alpha U - \alpha\mu_u)}] = \mathbb{E}[e^{\lambda\alpha(U-\mu_u)}].$$

  Now, since U is a subexponential random variable with parameters (ν, b), we know that we can upper bound this as

$$\mathbb{E}[e^{\lambda(\alpha U - \alpha\mu_u)}] = \mathbb{E}[e^{\lambda\alpha(U-\mu_u)}] \leq e^{(\lambda\alpha)^2\nu^2/2},$$

  as long as $|\lambda\alpha| < 1/b$. Now, we rewrite this just slightly. Namely, we can upper bound

$$\mathbb{E}[e^{\lambda(\alpha U - \alpha\mu_u)}] \leq e^{\lambda^2(\alpha\nu)^2/2},$$

  as long as $|\lambda| < 1/(\alpha b)$. Which is exactly the same as saying "αU is a subexponential random variable with parameters $(\alpha\nu, \alpha b)$".
- First, observe that the mean of $U + V$ is simply the sum of the means of U and V. Hence, looking at the definition of subexponential, we need to study

$$\mathbb{E}[e^{\lambda(U+V-\mu_u-\mu_v)}] = \mathbb{E}[e^{\lambda(U-\mu_u)}e^{\lambda(V-\mu_v)}] = \mathbb{E}[e^{\lambda(U-\mu_u)}]\mathbb{E}[e^{\lambda(V-\mu_v)}],$$

  where the last step follows because U and V are independent. Next, since U and V are subexponential, we can upper bound the two factors as

$$\mathbb{E}[e^{\lambda(U+V-\mu_u-\mu_v)}] = \mathbb{E}[e^{\lambda(U-\mu_u)}]\mathbb{E}[e^{\lambda(V-\mu_v)}] \leq e^{\nu_u^2\lambda^2/2}e^{\nu_v^2\lambda^2/2},$$

  which holds whenever $|\lambda| < 1/b_u$ and at the same time also $|\lambda| < 1/b_v$. Arranging terms, we can thus conclude that

$$\mathbb{E}[e^{\lambda(U+V-\mu_u-\mu_v)}] \leq e^{\left(\sqrt{\nu_u^2+\nu_v^2}\right)^2\lambda^2/2},$$

  whenever $|\lambda| < \min(1/b_u, 1/b_v)$. Which is exactly the same as saying "$U + V$ is a subexponential random variable with parameters $(\sqrt{\nu_u^2 + \nu_v^2}, \max(b_u, b_v))$".

(b) We take up the claims in turn:

- The random variables Z_i are independent of each other because Z_i are merely (weighted) sums of the X_{ij}, no X_{ij} appears in more than one of the Z_i, and the X_{ij} are by assumption independent of each other.

- Each Z_i is subgaussian simply because it is a (weighted) sum of subgaussian random variables, see Lemma 2.1 in the lecture notes. From that same lemma, we directly find that the variance proxy of Z_i is σ^2.
- We have $\mathbb{E}[Z_i^2] = \frac{1}{\|u\|_2^2} \sum_{j=1}^d u_j^2 \mathbb{E}[X_{ij}^2]$, since the X_{ij} are independent of each other and have mean zero. Moreover, we have $\mathbb{E}[X_{ij}^2] = 1$ for all i and j, and thus, $\mathbb{E}[Z_i^2] = 1$.

(c) We know that each Z_i is subgaussian with variance proxy σ^2. Therefore, using the theorem, we know that each Z_i^2 is subexponential with mean 1 and parameters $(\sqrt{32}\sigma^2, 4\sigma^2)$. Moreover, all the Z_i^2 are independent of each other. Now, using the second half of Part (a), we can observe that $\sum_{i=1}^k Z_i^2$ is subexponential with mean k and parameters $(\sqrt{k}\sqrt{32}\sigma^2, 4\sigma^2)$. Then, using the first half of Part (a), we can observe that $S = \frac{1}{k}\sum_{i=1}^k Z_i^2$ is subexponential with mean 1 and parameters $(\frac{\sqrt{32}\sigma^2}{\sqrt{k}}, \frac{4\sigma^2}{k})$.

Alternatively, we could directly observe that the mean of S is 1 and write out, leveraging the fact that the Z_i^2 are independent of each other:

$$\begin{aligned}
\mathbb{E}[e^{\lambda(S-1)}] &= \mathbb{E}[e^{\lambda(\frac{1}{k}\sum_{i=1}^k Z_i^2 - 1)}] \\
&= \prod_{i=1}^k \mathbb{E}[e^{\frac{\lambda}{k}(Z_i^2-1)}] \\
&\leq \left(e^{16\sigma^4\left(\frac{\lambda}{k}\right)^2}\right)^k = e^{\frac{16\sigma^4\lambda^2}{k}},
\end{aligned}$$

which holds for $|\frac{\lambda}{k}| < \frac{1}{4\sigma^2}$. That is, S with mean 1 is subexponential with parameters $(\frac{\sqrt{32}\sigma^2}{\sqrt{k}}, \frac{4\sigma^2}{k})$.

(d) Here, we can directly leverage Lemma 2.6 from the Lecture Notes to conclude

$$\mathbb{P}\left\{\left|\frac{1}{k}\sum_{i=1}^k Z_i^2 - 1\right| > \delta\right\} \leq 2e^{-\frac{\delta^2 k}{64\sigma^4}},$$

which holds whenever

$$\delta \leq \frac{\left(\frac{\sqrt{32}\sigma^2}{\sqrt{k}}\right)^2}{\frac{4\sigma^2}{k}} = 8\sigma^2.$$

More precisely, we apply Lemma 2.6 from the Lecture Notes separately to the positive and to the negative deviations. Since these are disjoint events, we can just add up the probabilities, which leads to the leading factor of 2 in our expression. This is an argument we have seen several times in the class.

(e) Discuss the differences of the resulting lemma with respect to what is proved in the lecture notes.

Problem 4. *(James-Stein Estimator)*[12 pts]

(a) [2 pts] 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) [2 pts] 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) [3 pts] 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) [1 pts] 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) [2 pts] 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) [2 pts] 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 4.

(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 $n \geq 3$.

Problem 5. *(l_2 Estimation)*[10 pts]

Assume that we have two distributions p and q on $\{1, \cdots K\}$. Let $n \in \mathbb{N}$. Let $N \sim \text{Poi}(n)$. We are given N iid samples from each, call them $\{X_j\}_{j=1}^N$ and $\{Y_j\}_{j=1}^N$, respectively. Let $t_k(X^N)$, $k = 1, \cdots, K$, respectively, $t_k(Y^N)$, denote the empirical counts. E.g.,

$$t_k(x^n) = |\{j \in \{1, \dots n\} : x_j = k\}|.$$

We want to estimate $\|p - q\|_2^2$.

Define $Z = \sum_{k=1}^K (t_k(X^N) - t_k(Y^N))^2 - t_k(X^N) - t_k(Y^N)$. We claim that Z/n^2 is a good estimator for $\|p - q\|_2^2$.

(a) [4pts] Show that Z is an unbiased estimator of $n^2\|p - q\|_2^2$.

Hint: The expression for Z should look somewhat familiar. The notes are your best friend.

(b) [3pts] Assuming that $\|p\|_2^2 \le b$ and $\|q\|_2^2 \le b$ show that the variance of Z can be upper bounded in the following way:

$$\begin{aligned}\text{Var}(Z) &\overset{(i)}{=} \sum_{k=1}^K 4n^3(p_k - q_k)^2(p_k + q_k) + 2n^2(p_k + q_k)^2 \\ &\overset{(ii)}{\le} \sum_{k=1}^K 8n^3(p_k - q_k)^2 + 2n^2(p_k^2 + q_k^2 + 2p_k q_k) \\ &\overset{(iii)}{\le} 8n^3\|p - q\|_2^2 + 8n^2 b.\end{aligned}$$

Justify each of the three steps.

Hint: Define $R = (U - V)^2 - U - V$, where $U \sim \text{Poi}(\lambda)$ and $V \sim \text{Poi}(\mu)$. A straighforward but tedious calculation shows that $\text{Var}(R) = 4(\lambda - \mu)^2(\lambda + \mu) + 2(\lambda + \mu)^2$.

(c) [3pts] Show that $\mathbb{P}\{|Z/n^2 - \|p - q\|_2^2| \ge \epsilon\} \le \frac{8n\|p-q\|_2^2 + 8b}{n^2\epsilon^2}$.

Solution 5.

(a) Define $Z_k = (t_k(X^N) - t_k(Y^N))^2 - t_k(X^N) - t_k(Y^N)$. Note that $t_k(X^N)$ is distributed according to $\text{Poi}(np_k)$ and that $t_k(Y^N)$ is distributed according to $\text{Poi}(nq_k)$. Hence,

$$\begin{aligned}
\mathbb{E}[Z_k] &= \mathbb{E}[(t_k(X^N) - t_k(Y^N))^2 - t_k(X^N) - t_k(Y^N)] \\
&= \mathbb{E}[t_k(X^N)(t_k(X^N) - 1) + t_k(Y^N)(t_k(Y^N) - 1) - 2t_k(X^N)t_k(Y^N)] \\
&= n^2p_k^2 + n^2q_k^2 - 2n^2p_kq_k \\
&= n^2(p_k - q_k)^2,
\end{aligned}$$

where in the one-before-last line we have used the fact that for a Poisson random variable of parameter λ, $\mathbb{E}[X(X-1)] = \lambda^2$ as well as the independence of $t_k(X^N)$ and $t_k(Y^N)$.

Since $Z = \sum_{k=1}^K Z_k$, it follows that $\mathbb{E}[Z] = n^2 \sum_{k=1}^K (p_k - q_k)^2 = n^2\|p-q\|_2^2$, as claimed.

(b)

$$\begin{aligned}
\text{Var}(Z) &\overset{(i)}{=} \sum_{k=1}^K 4n^3(p_k - q_k)^2(p_k + q_k) + 2n^2(p_k + q_k)^2 \\
&\overset{(ii)}{\leq} \sum_{k=1}^K 8n^3(p_k - q_k)^2 + 2n^2(p_k^2 + q_k^2 + 2p_kq_k) \\
&\overset{(iii)}{\leq} 8n^3\|p - q\|_2^2 + 8n^2b.
\end{aligned}$$

To see step (i) note that we have

$$\text{Var}(Z) = \sum_{k=1}^K \text{Var}(Z_k)$$

since the random variables Z_k are independent. For step (ii) we use $p_k + q_k \leq 2$. For step (iii) use the bounds $\|p\|_2^2 \leq b$ and $\|q\|_2^2 \leq b$ as well as $\sum_{k=1}^K p_kq_k \leq \|p\|_2\|q\|_2 \leq \beta$ (Cauchy Schwartz).

(c) This is just the Chebyshev inequality.