Topics in ML: Homework #1

Due on Thursday 2nd, 2023 at 23:59pm

Jacopo Peroni

While solving this homework I discussed some of the problems with Marco Scialanga.

Problem 1

The following serve as reminders of the definitions for population risk and empirical risk with respect to the collection S_n.

$$\mathcal{R}_\rho(\theta) = \mathbf{E}_{Z\sim\rho}[(\theta - Z)^2]$$

$$\hat{\mathcal{R}}_{S_n}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(\theta - Z_i)^2$$

Subproblem 1.1

For a fixed $\rho \in \mathcal{P}_{\sigma^2}(\mathbb{R})$, we know that the distribution has a finite second moment, so

$$\mathcal{R}_\rho(\theta) = \mathbf{E}[(\theta - Z)^2] = \mathbf{E}[\theta^2 - 2\theta Z + Z^2] = \mathbf{E}[\theta^2] + \mathbf{E}[-2\theta Z] + \mathbf{E}[Z^2] = \theta^2 - 2\theta\mathbf{E}[Z] + \mathbf{E}[Z^2]$$

Where in the last equality we used the fact that the expected value of a constant is the constant itself.
This is just a convex quadratic function in θ, so we can look for the optimal parameter value setting to 0 the gradient.

$$\nabla\mathcal{R}_\rho(\theta) = 2\theta - 2\mathbf{E}[Z] \implies \theta^* = \mathbf{E}[Z]$$

So the value of $\mathcal{R}_\rho(\theta*)$ is

$$\mathcal{R}_\rho(\theta^*) = \mathbf{E}[(\mathbf{E}[Z] - Z)^2]$$

By definition this value is $< \sigma^2$.

Subproblem 1.2

If we define a vector with entries θ (we'll write it as $\mathbb{1}_\theta$), we can rewrite the empirical risk as:

$$\hat{\mathcal{R}}_{S_n}(\theta) = \frac{1}{n}\|\mathbb{1}_\theta - S_n\|^2 = \frac{1}{n}(\|\mathbb{1}_\theta\|^2 - 2S_n^T\mathbb{1}_\theta + S_n^TS_n) = \frac{1}{n}(n\theta^2 - 2S_n^T\mathbb{1}_\theta + S_n^TS_n)$$

We can find the minimum by setting $\nabla\hat{\mathcal{R}}_{S_n}(\theta) = 0$.

$$\nabla(\frac{1}{n}(n\theta^2 - 2S_n^T\mathbb{1}_\theta + S_n^TS_n)) = \frac{1}{n}(2n\theta - 2\sum_{i=1}^{n} Z_i) = 0$$

Thus,

$$\hat{\theta}^{ERM}(S_n) = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

Subproblem 1.3

By sdefinition we have that:

$$\mathcal{R}_\rho(\theta) - \mathcal{R}_\rho(\theta^*_\rho) = \mathbf{E}[(\theta - Z)^2] - \mathbf{E}[(\theta^*_\rho - Z)^2]$$

If we expand the formula, using the fact that the expected value is linear and homogeneous and that the expected value of a constant is the constant itself, we obtain:

$$\mathbf{E}[(\theta - Z)^2] - \mathbf{E}[(\theta^*_\rho - Z)^2] = \mathbf{E}[\theta^2 - 2\theta Z + Z^2] - \mathbf{E}[\theta^{*2}_\rho - 2\theta Z + Z^2] =$$

$$= \theta^2 - 2\theta\mathbf{E}[Z] + \mathbf{E}[Z^2] - \theta^{*2}_\rho + 2\theta^{*2}_\rho\mathbf{E}[Z] - \mathbf{E}[Z^2] =$$

From subproblem 1.1 we know that $\theta^*_\rho = \mathbf{E}[Z]$. Therefore, we have

$$= \theta^2 - 2\theta\mathbf{E}[Z] - \mathbf{E}[Z]^2 + 2\mathbf{E}[Z]^2 = \theta^2 - 2\theta\mathbf{E}[Z] + \mathbf{E}[Z]^2 = (\theta - \mathbf{E}[Z])^2 = (\theta - \theta^{*2}_\rho)^2$$

Subproblem 1.4

From subproblem 1.3 we know that $\mathcal{R}_\rho(\hat{\theta}^{ERM}(S_n)) - \mathcal{R}_\rho(\theta^*_\rho) = (\hat{\theta}^{ERM}(S_n) - \theta^*_\rho)^2$. Since $\mathcal{R}_\rho(\theta^*_\rho)$ is defined as the inf of the population risk, it's coherent that this difference is non-negative.

Additionally, if Z_i are i.i.d. from a σ^2-subgaussian distribution, we have that $\mathbf{E}[\frac{1}{n}\sum_{i=1}^n Z_i] = \frac{1}{n}\sum_{i=1}^n \mathbf{E}[Z_i] = \mathbf{E}[Z]$ and that $\frac{1}{n}\sum_{i=1}^n Z_i$ is a $\frac{\sigma^2}{n}$-subgaussian distribution. Using these concepts in the problem we get

$$\mathbf{P}\left(\mathcal{R}_\rho(\hat{\theta}^{ERM}(S_n)) - \mathcal{R}_\rho(\theta^*_\rho) \geq \frac{2\sigma^2 log(2/\delta)}{n}\right) = \mathbf{P}\left((\hat{\theta}^{ERM}(S_n) - \theta^*_\rho)^2 \geq \frac{2\sigma^2 log(2/\delta)}{n}\right) =$$

Now from subproblem 1.1 and 1.2 we know that $\hat{\theta}^{ERM}(S_n) = \frac{1}{n}\sum_{i=1}^n Z_i$ and that $\theta^*_\rho = \mathbf{E}[Z]$ (let's note that $\mathbf{E}[\frac{1}{n}\sum_{i=1}^n Z_i] = \mathbf{E}[Z]$).

So we are looking for the probability of a superlever set of the square of the difference between a $\frac{\sigma^2}{n}$-subgaussian and its expected value.

$$= \mathbf{P}\left(\hat{\theta}^{ERM}(S_n) - \theta^*_\rho \geq \sqrt{\frac{2\sigma^2 log(2/\delta)}{n}}\right) + \mathbf{P}\left(\theta^*_\rho - \hat{\theta}^{ERM}(S_n) \geq \sqrt{\frac{2\sigma^2 log(2/\delta)}{n}}\right)$$

To avoid repetition, We prove that the first and second probabilities are $< \delta/2$, and by summing them we obtain the desired result.

$$\mathbf{P}\left(\hat{\theta}^{ERM}(S_n) - \theta^*_\rho \geq \sqrt{\frac{2\sigma^2 log(2/\delta)}{n}}\right) = \mathbf{P}\left(\exp\left(\lambda(\hat{\theta}^{ERM}(S_n) - \theta^*_\rho)\right) \geq \exp\left(\lambda\sqrt{\frac{2\sigma^2 log(2/\delta)}{n}}\right)\right)$$

Using Markov inequality and the property of subgaussian we get that

$$\leq \exp\left(-\lambda\sqrt{\frac{2\sigma^2 log(2/\delta)}{n}}\right)\mathbf{E}\left[\exp\left(\lambda(\hat{\theta}^{ERM}(S_n)-\theta^*_\rho)\right)\right] =$$

$$\exp\left(-\lambda\sqrt{\frac{2\sigma^2 log(2/\delta)}{n}}\right)\mathbf{E}\left[\exp\left(\lambda\left(\frac{1}{n}\sum_{i=1}^n Z_i - \mathbf{E}[Z]\right)\right)\right] \leq \exp\left(-\lambda\sqrt{\frac{2\sigma^2 log(2/\delta)}{n}}\right)\exp\left(\frac{\lambda^2\sigma^2}{2n}\right)$$

Now, choosing

$$\lambda = \frac{n}{\sigma^2}\sqrt{\frac{2\sigma^2}{n}log\left(\frac{2}{\delta}\right)}$$

concludes the proof.

Subproblem 1.5

Here we lose the subgaussian property but we still have that $\mathcal{R}_\rho(\hat{\theta}^{ERM}(S_n)) - \mathcal{R}_\rho(\theta^*_\rho) = (\hat{\theta}^{ERM}(S_n)-\theta^*_\rho)^2$.
We know that

$$\mathbf{P}\left((\hat{\theta}^{ERM}(S_n)-\theta^*_\rho)^2 \geq \frac{2\sigma^2 log(2/\delta)}{n}\right) = \mathbf{P}\left(\left(\frac{1}{n}\sum_{i=1}^n Z_i - \mathbf{E}[Z]\right)^2 \geq \frac{2\sigma^2 log(2/\delta)}{n}\right) =$$

So we can rewrite

$$= \mathbf{P}\left(\left(\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n Z_i - \mathbf{E}[Z]\right)\right)^2 \geq 2\sigma^2 log(2/\delta)\right) =$$

$$\mathbf{P}\left(\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n Z_i - \mathbf{E}[Z]\right) \geq \sqrt{2\sigma^2 log(2/\delta)}\right) + \mathbf{P}\left(\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n Z_i - \mathbf{E}[Z]\right) \leq -\sqrt{2\sigma^2 log(2/\delta)}\right)$$

We know that the Z_i are i.i.d. so they all have the same variance ($< \sigma^2$ by definition), which we denote by $\tilde{\sigma}^2$.
By taking the limit, we can make use of the central limit theorem to show that both probabilities converge to the cdf of $N \sim \mathcal{N}(0, \tilde{\sigma}^2)$. Indeed,

$$\lim_{n\to+\infty}\mathbf{P}\left(\mathcal{R}_\rho(\hat{\theta}^{ERM}(S_n)) - \mathcal{R}_\rho(\theta^*_\rho) \geq \frac{2\sigma^2 log(2/\delta)}{n}\right) = 2\mathbf{P}\left(N \geq \frac{\sigma\sqrt{2log(2/\delta)}}{\tilde{\sigma}}\right) \leq$$

$$\leq 2\mathbf{P}\left(N \geq \sqrt{2log(2/\delta)}\right)$$

Now we prove that this probability is upper bounded by δ.
We know that this normal random variable is $\tilde{\sigma}^2$-subgaussian, so we use the property of a subgaussian to upper bound the probability.

$$\mathbf{P}\left(\exp\left(\lambda N\right) \geq \exp\left(\lambda\sqrt{2log(2/\delta)}\right)\right) \leq \mathbf{E}\left[\exp(\lambda N)\right]\exp\left(-\lambda\sqrt{2log(2/\delta)}\right) \leq$$

$$\leq \exp\left(\frac{\tilde{\sigma}^2\lambda^2}{2}\right)\exp\left(-\lambda\sqrt{2log(2/\delta)}\right) = \exp\left(\frac{\tilde{\sigma}^2\lambda^2}{2} - \lambda\sqrt{2log(2/\delta)}\right)$$

So if we fix that this exponential to be $< \delta/2$ we will get the desired result. This is only possible if the following inequality has a positive solution

$$\frac{\tilde{\sigma}^2\lambda^2}{2} - \lambda\sqrt{2log(2/\delta)} \leq log\left(\frac{\delta}{2}\right)$$

If we set $\lambda = -\sqrt{2log(2/\delta)}$, we get

$$-\tilde{\sigma}^2 log\left(\frac{\delta}{2}\right) - 2log\left(\frac{\delta}{2}\right) \leq log\left(\frac{\delta}{2}\right) \implies -\tilde{\sigma}^2 - 2 \leq 1 \implies \tilde{\sigma}^2 \geq -3$$

That is always true. Consequently, we obtain that

$$2\mathbf{P}\left(N \geq \frac{\sigma\sqrt{2log(2/\delta)}}{\tilde{\sigma}}\right) \leq 2\mathbf{P}\left(N \geq \sqrt{2log(2/\delta)}\right) \leq 2\exp\left(\frac{\delta}{2}\right) = \delta \implies$$

$$\lim_{n\to+\infty}\mathbf{P}\left(\mathcal{R}_\rho(\hat{\theta}^{ERM}(S_n)) - \mathcal{R}_\rho(\theta^*_\rho) \geq \frac{2\sigma^2 log(2/\delta)}{n}\right) \leq \delta$$

Subproblem 1.6

It's useful to recall that

$$\mathbf{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Z_i - \mathbf{E}[Z]\right)^2\right] = Var\left(\frac{1}{n}\sum_{i=1}^n Z_i\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^n Z_i\right) = \frac{1}{n^2}\sum_{i=1}^n Var(Z_i) = \frac{\tilde{\sigma}^2}{n}$$

due to the fact that the Z_i are i.i.d.
So now, using Markov's inequality (or Chebyshev inequality if we take the square root of both sides)

$$\mathbf{P}\left((\hat{\theta}^{ERM}(S_n) - \theta^*_\rho)^2 \geq \frac{\sigma^2}{n\delta}\right) = \mathbf{P}\left(\left(\frac{1}{n}\sum_{i=1}^n Z_i - \mathbf{E}[Z]\right)^2 \geq \frac{\sigma^2}{n\delta}\right) \leq \frac{n\delta}{\sigma^2}\mathbf{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Z_i - \mathbf{E}[Z]\right)^2\right] =$$

$$= \frac{n\delta}{\sigma^2}\frac{\tilde{\sigma}^2}{n} \leq \frac{n\delta}{\sigma^2}\frac{\sigma^2}{n} = \delta$$

Subproblem 1.7

To prove the lower bound we use the same line of reasoning as the previous subproblem.
If we introduce the value $\frac{c\sigma^2}{n\delta}$ in the previous formula we find an upper bound of $\frac{\delta}{c}$, this tells us that if we want to lower bound the same probability with δ we need $c \leq 1$.
Even more, with this constrain we obtain that

$$\frac{n\delta}{c\sigma^2}\frac{\tilde{\sigma}^2}{n} \geq \delta \iff \tilde{\sigma}^2 \in \left[c\sigma^2, \sigma^2\right]$$

That is a non-empty condition iff $c \leq 1$.
In the previous subproblem we used the Markov inequality to upper bound the probability if we were able to make that inequality an equality, adding the lower bound just presented, we would have the result.
We have to find a distribution such that

$$\mathbf{P}\left(\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbf{E}[Z]\right)^2 \geq \frac{c\sigma^2}{n\delta}\right) = \frac{n\delta}{c\sigma^2}\mathbf{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbf{E}[Z]\right)^2\right]$$

Let's write the proof for Markov's inequality, let $X \geq 0$ and $t \geq 0$ so we have

$$\mathbf{E}[X] = \int_0^{+\infty} \mathbf{P}\left(X \geq y\right) dy \geq \int_0^{t} \mathbf{P}\left(X \geq y\right) dy \geq t\mathbf{P}\left(X \geq t\right)$$

So to have an equality where there is the first inequality we need $\mathbf{P}\left(X \geq y\right) = 0$ for $y > t$.
For the second one we need that $\mathbf{P}\left(X \geq y\right) = \text{const}$ for $y \in [0, t]$ so it means that $X \in \{0, t\}$.
In our case

$$X = \left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbf{E}[Z]\right)^2 \text{ and } t = \frac{c\sigma^2}{n\delta}$$

So a distribution that satisfies the lower bound inequality is such that if $Z_i \sim \rho_{n,\delta}$ we have that the $Var(Z_i) \in [c\sigma^2, \sigma^2]$ and

$$\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbf{E}[Z] \in \left\{-\sigma\sqrt{\frac{c}{n\delta}}, 0, \sigma\sqrt{\frac{c}{n\delta}}\right\}$$

I'm unable to find an example of a distribution with these two properties.

Subproblem 1.8

Starting from the hint we can see that $U \sim$ Binomial$(m, \frac{1}{4})$ is a sum of m Bernoulli variables (U_i) that are limited to the range $[0, 1]$

$$\mathbf{P}\left(U \geq \frac{m}{2}\right) = \mathbf{P}\left(\sum_{i=1}^{m} U_i \geq \frac{m}{2}\right) = \mathbf{P}\left(\sum_{i=1}^{m} U_i - \frac{m}{4} \geq \frac{m}{4}\right) = \mathbf{P}\left(\frac{1}{m}\sum_{i=1}^{m}\left(U_i - \frac{1}{4}\right) \geq \frac{1}{4}\right) \leq$$

By Hoeffding's inequality, we know that

$$\leq \exp\left(-\frac{2m}{16}\right) = \exp\left(-\frac{m}{8}\right)$$

Thus yielding the required inequality.
Now, we first start by seeing that

$$\hat{\theta}^{ERM}(S_n^l) = \frac{1}{k}\sum_{i=1}^{k} Z_{k(l-1)+i}$$

And so by fixing one specific S_n^j as the median we can define the event

$$A = \left\{ \left| \frac{1}{k}\sum_{i=1}^{m} Z_{k(j-1)+i} - \mathbf{E}[Z] \right| \geq \frac{2\sigma}{\sqrt{k}} \right\}$$

Which, from the subproblem 1.6, has an upper bounded probability with δ defined as

$$\frac{1}{\sqrt{\delta}} = 2 \Longrightarrow \delta = \frac{1}{4}$$

If we want that the median of the set of variables to follow this rule, we require that at least half of them follow that rule.
We can consider m times the event A with changing j, the probability of the event A occurring (at fixed j) is always the same because alle the variables are i.i.d.
The original problem is equivalent to asking that A is happening at least $\frac{m}{2}$ times.
So defining m variables like this, $i \in \{1, \ldots, m\}$

$$\tilde{U}_i = \begin{cases} 1 & \text{A occurs} \\ 0 & \text{otherwise} \end{cases}$$

We have that $\mathbf{P}(\tilde{U}_i = 1) \leq \frac{1}{4}$, so for all i there exist a $U_i \sim$ Bernoulli($\frac{1}{4}$) such that

$$\mathbf{P}(\tilde{U}_i = 1) \leq \mathbf{P}(U_i = 1) = \frac{1}{4}$$

If we define

$$U = \sum_{i=1}^{m} U_i$$

we obtain a Binomial($m, \frac{1}{4}$).
Putting it all together we have that

$$\mathbf{P}\left(\left| median\left[\hat{\theta}^{ERM}(S_n^1), \ldots, \hat{\theta}^{ERM}(S_n^m)\right] - \mathbf{E}[Z]\right| \geq \frac{2\sigma}{\sqrt{k}}\right) \leq \mathbf{P}\left(U \geq \frac{m}{2}\right) \leq \exp\left(-\frac{m}{8}\right)$$

Subproblem 1.9

As suggested we want to build an estimator starting from the solution of the previous question.

We start by noticing that

$$\mathbf{P}\left(\mathcal{R}_\rho(\hat{\theta}_{\sigma^2,\delta}(S_n)) - \mathcal{R}_\rho(\theta^*_\rho) \geq \frac{c\sigma^2 log(1/\delta)}{n}\right) = \mathbf{P}\left((\hat{\theta}_{\sigma^2,\delta}(S_n) - \theta^*_\rho)^2 \geq \frac{c\sigma^2 log(1/\delta)}{n}\right) =$$

$$\mathbf{P}\left(\left|\hat{\theta}_{\sigma^2,\delta}(S_n) - \theta^*_\rho\right| \geq \sqrt{\frac{c\sigma^2 log(1/\delta)}{n}}\right)$$

If we take $n = mk$ with m and k integers (which can always be done, in the worst case, with prime numbers, we have $n = n * 1$). We consider

$$\hat{\theta}_{\sigma^2,\delta}(S_n) = median\left[\hat{\theta}^{ERM}(S^1_n), \ldots, \hat{\theta}^{ERM}(S^m_n)\right]$$

Where we are using the same notation from before. We have a condition very similar to the one before.

In fact we can fix c such that

$$\mathbf{P}\left(\left|\hat{\theta}_{\sigma^2,\delta}(S_n) - \theta^*_\rho\right| \geq \sqrt{\frac{c\sigma^2 log(1/\delta)}{n}}\right) \leq \mathbf{P}\left(\left|\hat{\theta}_{\sigma^2,\delta}(S_n) - \theta^*_\rho\right| \geq \frac{2\sigma}{\sqrt{k}}\right)$$

The condition that we have to fix is

$$\sigma\sqrt{c\frac{log(1/\delta)}{mk}} \geq \frac{2\sigma}{\sqrt{k}} \Longrightarrow \sqrt{c\frac{log(1/\delta)}{m}} \geq 2 \Longrightarrow m \leq \frac{c}{4}log(1/\delta)$$

we will tackle the problem of m being restricted to the divisors of n later.

Now, from subproblem 1.9, we know that

$$\mathbf{P}\left(\left|\hat{\theta}_{\sigma^2,\delta}(S_n) - \theta^*_\rho\right| \geq \frac{2\sigma}{\sqrt{k}}\right) \leq \exp\left(\frac{-m}{8}\right)$$

If we enforce that the $exp(-m/8)$ has to be lower that δ we obtain

$$\exp\left(\frac{-m}{8}\right) \leq \delta \Longrightarrow m \geq 8log\left(1/\delta\right)$$

Therefore by fixing $c > 16$ there is a non-empty subset of $\mathbb{R}$ with values of m that follows both inequalities.

The only problem now is if the admissible values of m are divisors of n, we can solve this problem by defining the estimator as follows

$$\hat{\theta}_{\sigma^2,\delta}(S_n) = \begin{cases} median\left[\hat{\theta}^{ERM}(S^1_n), \ldots, \hat{\theta}^{ERM}(S^m_n)\right] & \text{if there exist admissible } m \text{ divisor of } n \\ \mathbf{E}[Z] & \text{otherwise} \end{cases}$$

Problem 2

Just a reminder about some definitions.
Let's define Hadamard matrix H_d by recurrence

$$H_1 = (1) \text{ and } H_{2d} = \begin{pmatrix} H_d & H_d \\ H_d & -H_d \end{pmatrix}$$

If we set the i-th row to be h_i we define the set

$$\mathcal{X} = \{h_1, h_2, \dots, h_d\}$$

and for $j \in \{1, 2, \dots, d\}$ we have the distribution ρ_j supported onto $\mathcal{X} \times \{-1, 1\}$ with

$$\mathbf{P}_{(X,Y)\sim\rho_j}(X = h_i, Y = h_{ij}) = \frac{1}{d} \text{ for } i = 1, \dots, d$$

Subproblem 2.1

Let's prove that the distributions have finite second moments.

$$\mathbf{E}_X[Y] = \sum_{y\in\{-1,1\}} y\mathbf{P}_X(y) = \sum_{y\in\{-1,1\}} y \sum_{x\in\mathcal{X}} \mathbf{P}(X = x, Y = y) =$$
$$= \sum_{x\in\mathcal{X}} (\mathbf{P}(X = x, Y = 1) - \mathbf{P}(X = x, Y = -1))$$
$$\mathbf{E}_Y[X] = \sum_{x\in\mathcal{X}} x\mathbf{P}_Y(x) = \sum_{x\in\mathcal{X}} x \sum_{y\in\{-1,1\}} \mathbf{P}(X = x, Y = y) =$$
$$= \sum_{x\in\mathcal{X}} x\left(\mathbf{P}(X = x, Y = 1) + \mathbf{P}(X = x, Y = -1)\right) = \sum_{x\in\mathcal{X}} x\frac{1}{d}$$

Where in the X case we are using a component-wise definition of multiplication.

$$Var(Y) = \mathbf{E}\left[Y^2\right] - \mathbf{E}[Y]^2 =$$
$$= \sum_{y\in\{-1,1\}} y^2 \sum_{x\in\mathcal{X}} \mathbf{P}(X = x, Y = y) - \mathbf{E}[Y]^2 =$$
$$= \sum_{x\in\mathcal{X}} (\mathbf{P}(X = x, Y = 1) + \mathbf{P}(X = x, Y = -1)) - \mathbf{E}[Y]^2 = 1 - \mathbf{E}[Y]^2 \leq 1$$

Where in the last equality we used the fact that if $x_j = 1$ we have that $\mathbf{P}(X = x, Y = 1) = 1/d$ and $\mathbf{P}(X = x, Y = -1) = 0$ and vice versa, so summing for $i = 1$ we have d times $1/d$. This we have that the variance of Y is limited.

We'll do the same with X. With an abuse of notation, we consider the following operations as component-wise.

$$Var(X) = \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2 = \sum_{x\in\mathcal{X}} x^2 \frac{1}{d} - \left(\sum_{x\in\mathcal{X}} x\frac{1}{d}\right)^2 = \frac{1}{d}\mathbb{1}_d - \left(\sum_{x\in\mathcal{X}} x\frac{1}{d}\right)^2 \leq \frac{1}{d}\mathbb{1}_d$$

Thus the $Var(X) is limited.$ So we can say that the distribution has finite second moments.

Now we can solve the subproblem using "bias-variance decomposition", $\forall z \in \mathbb{R}$

$$\mathcal{R}_{\rho_j}(z|x) = \mathbf{E}\left[(Y-z)^2\,|X=x\right] =$$
$$\mathbf{E}\left[(Y-\mathbf{E}\left[Y|X=x\right]+\mathbf{E}\left[Y|X=x\right]-z)^2\,|X=x\right] =$$
$$\mathbf{E}\left[(Y-\mathbf{E}\left[Y|X=x\right])^2\,|X=x\right] + 2\mathbf{E}\left[(Y-\mathbf{E}\left[Y|X=x\right])\left(\mathbf{E}\left[Y|X=x\right]-z\right)|X=x\right]$$
$$+\left(\mathbf{E}\left[Y|X=x\right]-z\right)^2$$

Since the second term is equal to zero, the minimization is for $z = \mathbf{E}\left[Y|X=x\right]$.
Consequently,

$$f_j^*(h_i) = \operatorname{argmin}_f \mathcal{R}_{\rho_j}(f) = \mathbf{E}\left[Y|X=h_i\right]$$

Now let's compute it

$$\mathbf{E}\left[Y|X=h_i\right] = \sum_{y\in\{-1,1\}} y\mathbf{P}_{Y|X}\left(Y=y|X=h_i\right) = \sum_{y\in\{-1,1\}} y\frac{\mathbf{P}\left(X=h_i, Y=y\right)}{\mathbf{P}_X(h_i)} =$$
$$= h_{ij}\frac{1}{d}\frac{1}{\mathbf{P}_X(h_i)} = h_{ij}$$

Because

$$\mathbf{P}_X(h_i) = \sum_{y\in\{-1,1\}} \mathbf{P}(X=h_i, Y=y) = \frac{1}{d}$$

To conclude let's compute $\mathcal{R}_{\rho_j}(f_j^*)$

$$\mathcal{R}_{\rho_j}(f_j^*) = \mathbf{E}_{(X,Y)\sim\rho_j}\left[\left(f_j^*(X)-Y\right)^2\right] = \mathbf{E}_X\left[\mathbf{E}_Y\left[\left(f_j^*(X)-Y\right)^2|X=h_i\right]\right] =$$
$$= \mathbf{E}_X\left[\sum_{y\in\{-1,1\}}\left(f_j^*(X)-y\right)^2\frac{\mathbf{P}\left(X=h_i,Y=y\right)}{\mathbf{P}_X(h_i)}\right] = \mathbf{E}_X\left[\left(f_j^*(X)-h_{ij}\right)^2\right] =$$
$$= \sum_{i=1}^d \left(f_j^*(h_i)-h_{ij}\right)^2\mathbf{P}_X(h_i) = \sum_{i=1}^d \left(h_{ij}-h_{ij}\right)^2\mathbf{P}_X(h_i) = 0$$

Subproblem 2.2

By definition, if $\hat{\theta} \in \mathcal{A}$ there exist a vector α such that

$$\hat{\theta}(S_n) = \sum_{i=1}^{n} \alpha_i \phi(x_i)$$

We want to prove that $\theta_\lambda^{KRR} \in \mathcal{A}$ where θ_λ^{KRR} is defined as follows

$$\hat{\theta}_\lambda^{KRR}(S_n) = \text{argmin}_{\theta \in \mathbb{R}} \sum_{i=1}^{n} \left(f_\theta(x_i) - y_i\right)^2 + \lambda \|\theta\|_2^2$$

We can use theorem 2.1 (Representer theorem) from the lectures notes and prove the result. In fact, knowing that $\mathbb{R}^m$ is an Hilbert space, and that the subspace $\tilde{\mathcal{A}} = \text{span}\{\phi(x_1), \dots, \phi(x_n)\}$ is closed in $\mathbb{R}^m$, if $\theta \in \mathbb{R}^m$ it is possible to write it as $\theta = \theta_{\tilde{\mathcal{A}}} + \theta^\perp$ with $\theta_{\tilde{\mathcal{A}}}$ being the orthogonal projection of θ onto $\tilde{\mathcal{A}}$.
Thus,

$$\langle \theta, \phi(x_i) \rangle = \langle \theta_{\tilde{\mathcal{A}}}, \phi(x_i) \rangle + \langle \theta^\perp, \phi(x_i) \rangle = \langle \theta_{\tilde{\mathcal{A}}}, \phi(x_i) \rangle$$

$$\|\theta\|_2^2 = \|\theta_{\tilde{\mathcal{A}}}\|_2^2 + \|\theta^\perp\|_2^2 \geq \|\theta_{\tilde{\mathcal{A}}}\|_2^2$$

So for every θ its projection $\theta_{\tilde{\mathcal{A}}}$ has smaller objective value.
Thus $\hat{\theta}_\lambda^{KRR} \in \mathcal{A}$.

Subproblem 2.3

Let's write $\hat{\theta}(S_n)$ and f_j^* more explicitly

$$\hat{\theta}(S_n) = \sum_{i=1}^{n} \alpha_i \phi(x_i)$$

$$f_{\hat{\theta}(S_n)}(h_k) = \sum_{i=1}^{n} \alpha_i \langle \phi(x_i), \phi(h_k) \rangle$$

$$f_j^*(h_k) = h_{kj} = (He_j)_k$$

Given that $\hat{\theta}$ is a learning algorithm, the coefficients $\alpha_1, \dots, \alpha_n$ are dependent on S_n.
Using the fact that $\mathcal{R}_{\rho_j}\left(f_j^*\right) = 0$ we have

$$\mathcal{R}_{\rho_j}\left(f_{\hat{\theta}(S_n)}\right) = \mathbf{E}_{(X,Y)\sim\rho_j}\left[\left(f_{\hat{\theta}(S_n)}(X) - Y\right)^2\right]$$

We know that by the law of total expectation computing $\mathbf{E}_{(X,Y)\sim\rho_j}[\cdot]$ is the same as doing $\mathbf{E}_X[\mathbf{E}_Y[\cdot|X = x]]$

$$\mathbf{E}_X\left[\mathbf{E}_Y\left[\left(f_{\hat{\theta}(S_n)}(X) - Y\right)^2 |X = \tilde{h}\right]\right] =$$

$$= \mathbf{E}_X \left[\sum_{y\in\{-1,1\}} \left(f_{\hat{\theta}(S_n)}(X) - y \right)^2 \frac{\mathbf{P}\left(X = \tilde{h}, Y = y \right)}{\mathbf{P}_X(\tilde{h})} \right] =$$

$$= \mathbf{E}_X \left[\left(\sum_{i=1}^{n} \alpha_i \langle \phi(x_i), \phi(\tilde{h}) \rangle - \tilde{h}_j \right)^2 \right] = \sum_{k=1}^{d} \left(\sum_{i=1}^{n} \alpha_i \langle \phi(x_i), \phi(h_k) \rangle - h_{kj} \right)^2 \mathbf{P}_X(h_k) =$$

$$= \frac{1}{d} \sum_{k=1}^{d} \left(\sum_{i=1}^{n} \alpha_i \langle \phi(x_i), \phi(h_k) \rangle - h_{kj} \right)^2 = \frac{1}{d} \sum_{k=1}^{d} \left(\sum_{i=1}^{n} \alpha_i \phi(x_i)^T \phi(h_k) - (H_d e_j)_k \right)^2 =$$

Given the matrix Φ as defined by the problem and

$$\tilde{\Phi} = \begin{pmatrix} \phi(x_1)^T \\ \vdots \\ \phi(x_n)^T \end{pmatrix} \; a = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}$$

We have

$$= \frac{1}{d} \| \Phi \tilde{\Phi}^T a - H_d e_j \|_2^2$$

Subproblem 2.4

Let's write the definition of a Frobenius norm for $A \in \mathbb{R}^{n\times m}$

$$\|A\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^2}$$

Considering that we have shown in the first subproblem that $\mathcal{R}_{\rho_j}\left(f_j^*\right) = 0$, we know need to prove that

$$\mathbf{E}_{J\sim\text{Uniform}\{1,\ldots,d\}} \left[\mathcal{R}_{\rho_J}\left(f_{\hat{\theta}(S_n^J)} \right) \right] = \sum_{j=1}^{d} \mathcal{R}_{\rho_J}\left(f_{\hat{\theta}(S_n^J)} \right) \frac{1}{d} =$$

From the result of previous exercise, and since the coefficients are dependent on the S_n^j, we have

$$= \frac{1}{d} \sum_{j=1}^{d} \frac{1}{d} \| \Phi \tilde{\Phi}^T a_j - H_d e_j \|_2^2 = \frac{1}{d^2} \sum_{j=1}^{d} \sum_{i=1}^{d} \left(\left(\Phi \tilde{\Phi}^T a_j \right)_i - \left(H_d e_j \right)_i \right)^2 =$$

$$= \frac{1}{d^2} \sum_{j=1}^{d} \sum_{i=1}^{d} \left(\left(\Phi \tilde{\Phi}^T a_j - H_d e_j \right)_i \right)^2 = \frac{1}{d^2} \| B - H_d \|_F^2$$

Where B is a matrix defined like this

$$B = \begin{pmatrix} \Phi \tilde{\Phi}^T a_1 & \ldots & \Phi \tilde{\Phi}^T a_d \end{pmatrix}$$

Subproblem 2.5

Let's write the SVD of B and of H_d.

$$B = U\Sigma V^T \text{ and } H_d = A\Gamma W^T$$

We have that $H_d = \sum_{i=1}^{d} \gamma_i a_i w_i^T$. Truncating the sum to the n-th term yields

$$H_{d,n} = \sum_{i=1}^{n} \gamma_i a_i w_i^T$$

Or in matrix terms

$$A = \begin{pmatrix} A_n & A_{d-n} \end{pmatrix} \; \Gamma = \begin{pmatrix} \Gamma_n & 0 \\ 0 & \Gamma_{d-n} \end{pmatrix} \; W = \begin{pmatrix} W_n & W_{d-n} \end{pmatrix}$$

$$H_{d,n} = A_n \Gamma_n W_n^T$$

Let's note also that the Frobenius norm is unitary, so

$$\text{Tr}\left(\left(U\Sigma V^T\right)^T \left(U\Sigma V^T\right)\right) = \text{Tr}\left(V\Sigma^T U^T U\Sigma V^T\right) = \text{Tr}\left(V\Sigma^T\Sigma V^T\right) = \text{Tr}\left(\Sigma^T\Sigma V^T V\right) =$$

$$= \text{Tr}\left(\Sigma^T\Sigma\right)$$

Consider the inf of the Frobenius norm

$$\|B - H_d\|_F^2 = \text{Tr}\left(\left(B^T - H_d^T\right)(B - H_d)\right) = \text{Tr}\left(B^T B\right) - 2\text{Tr}\left(B^T H_d\right) + \text{Tr}\left(H_d^T H_d\right) =$$

$$= \text{Tr}\left(\Sigma^T\Sigma\right) - 2\text{Tr}\left(B^T H_d\right) + \text{Tr}\left(\Gamma^T\Gamma\right) =$$

$$= \text{Tr}\left(\Sigma^T\Sigma\right) - 2\text{Tr}\left(B^T H_d\right) + \text{Tr}\left(\Gamma_n^T\Gamma_n\right) + \text{Tr}\left(\Gamma_{d-n}^T\Gamma_{d-n}\right) \geq$$

$$\geq \text{Tr}\left(\Sigma^T\Sigma\right) - 2\text{Tr}\left(\Sigma^T\Gamma_n\right) + \text{Tr}\left(\Gamma_n^T\Gamma_n\right) + \text{Tr}\left(\Gamma_{d-n}^T\Gamma_{d-n}\right) = \|\Sigma - \Gamma_n\|_F^2 + \|\Gamma_{d-n}\|_F^2$$

Where the only inequality is the Von Neumann's trace inequality.
Thus the matrix B which minimizes the objective is $H_{d,n}$.
So we know that

$$\inf_{B\in\mathbb{R}^{d\times d}:\ \text{rank}(B)\leq n} \|B - H_d\|_F^2 \geq \|H_{d,n} - H_d\|_F^2$$

Computing the difference gives

$$H_d - H_{d,n} = \sum_{i=n+1}^{d} \gamma_i a_i w_i^T$$

Thus, given that $H_d^T H_d = dI_d$.

$$\inf_{B\in\mathbb{R}^{d\times d}:\ \text{rank}(B)\leq n} \|B - H_d\|_F^2 \geq \|H_{d,n} - H_d\|_F^2 = \|H_d - H_{d,n}\|_F^2 = \sum_{i=n+1}^{d} \gamma_i^2 = d(d-n)$$

Subproblem 2.6

Starting from the fact that $\mathcal{R}_{\rho_J}\left(f_J^*\right)=0$ we have

$$\mathbf{E}_{J\sim\text{Uniform}\{1,\dots,d\}}\mathbf{E}_{S_n\sim\rho_J^{\otimes n}}\left[\mathcal{R}_{\rho_J}\left(f_{\hat{\theta}(S_n^J)}\right)-\mathcal{R}_{\rho_J}\left(f_J^*\right)\right]=$$

$$=\mathbf{E}_{J\sim\text{Uniform}\{1,\dots,d\}}\mathbf{E}_{S_n\sim\rho_J^{\otimes n}}\left[\mathcal{R}_{\rho_J}\left(f_{\hat{\theta}(S_n^J)}\right)\right]=$$

Now we can expand the expected values

$$=\frac{1}{d}\sum_{j=1}^{d}\mathbf{E}_{S_n\sim\rho_J^{\otimes n}}\left[\mathcal{R}_{\rho_J}\left(f_{\hat{\theta}(S_n^J)}\right)\right]=\frac{1}{d}\sum_{j=1}^{d}\sum_{S_n}\mathcal{R}_{\rho_J}\left(f_{\hat{\theta}(S_n^J)}\right)\mathbf{P}\left(S_n\right)=$$

Now we swap the sums given that there are a finite number of possible S_n for all j

$$=\sum_{S_n}\frac{1}{d}\sum_{j=1}^{d}\mathcal{R}_{\rho_J}\left(f_{\hat{\theta}(S_n^J)}\right)\mathbf{P}\left(S_n\right)=\sum_{S_n}\frac{1}{d^2}\|B_{S_n}-H_d\|_F^2\mathbf{P}\left(S_n\right)\geq\frac{1}{d^2}d(d-n)\sum_{S_n}\mathbf{P}\left(S_n\right)=1-\frac{n}{d}$$

Where we wrote B_{S_n} because the matrix Φ, used to define B, dependens on S_n.

Subproblem 2.7

(a) Let's start by analyzing the term $\mathcal{R}_{\rho_J}\left(f_{\tilde{\theta}_{1,\phi}^{KRR}(S_n^J)}\right)-\mathcal{R}_{\rho_J}\left(f_J^*\right)$ inside the expected values of the function $\mathcal{E}(\phi,n)$.
We know that for all $j\in\{1,\dots,d\}$ $\mathcal{R}_{\rho_J}\left(f_J^*\right)=0$.
We know that by fixing $j\in\{1,\dots,d\}$ and $S_n^j=\{\left(x_i,f_j^*(x_i)\right\})_{i=1}^n$ we have that

$$\tilde{\theta}_{1,\phi}^{KRR}(S_n^j)=\text{argmin}_{\theta\in\mathbb{R}^d}\sum_{i=1}^{n}\left(f_\theta(x_i)-f_j^*(x_i)\right)^2+\|\theta\|_2^2$$

And we know from subproblem 2.2 that $\tilde{\theta}_{1,\phi}^{KRR}\in\mathcal{A}$. So $\tilde{\theta}_{1,\phi}^{KRR}$ can be written as

$$\tilde{\theta}_{1,\phi}^{KRR}(S_n^j)=\sum_{i=1}^{n}\alpha_i f(x_i)$$

It's important to notice that α_i is dependent on S_n^j.
Now through subproblem 2.3 we know that

$$\mathcal{R}_{\rho_J}\left(f_{\tilde{\theta}_{1,\phi}^{KRR}(S_n^J)}\right)=\frac{1}{d}\|\Phi\tilde{\Phi}^Ta-H_de_j\|_F^2$$

Where $\tilde{\Phi}$ and a are dependent on S_n^j.
We can sample $Z^{\phi,n}$ by sampling $j\in\{1,\dots,d\}$ uniformly and n times $x\in\{h_1,\dots,h_d\}$ uniformly.

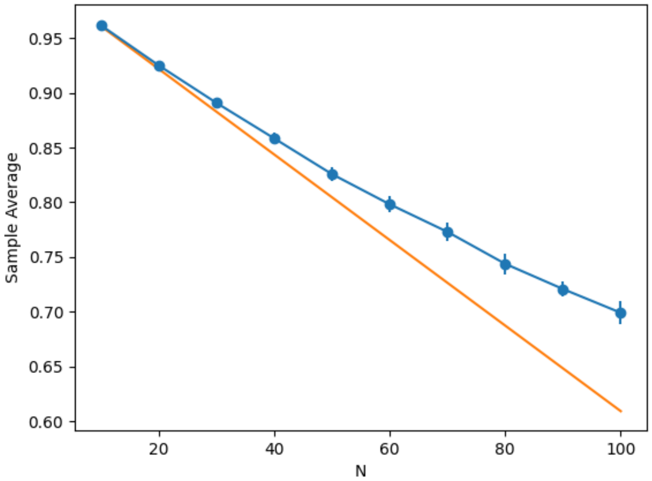
We also need to compute Φ, $\tilde{\Phi}$, a, and $\frac{1}{d}\|\Phi\tilde{\Phi}^T a - H_d e_j\|_F^2$.
From exercise sheet 5.2.a, applied to our specific case, we have that

$$a = (K + nI_n)^{-1}y \text{ with } K = \tilde{\Phi}\tilde{\Phi}^T$$

Given that the random components of this $Z^{\phi,n}$ follow the same distribution of J and S_n. We have that

$$\mathbb{E}\left[Z^{\phi,n}\right] = \mathcal{E}(\phi, n)$$

(b) Sample average of 50 i.i.d. random variables sampled as in (a), plotted with 1 std deviation error bar (in blue) and the lower bound (in orange).



(c) The main problem with the implementation of this feature map is the creation of the index set. This has a size related to the ways you can choose 100 spots (the maximum number of non-zero spots in a partition of 100) in a 256-length multi-index and the number of partitions of 100, so the product the size of J is on the order of some factorial.
What we can do is work around the problem by looking at where we are going to use the function ϕ.
The function is used to define the matrices Φ and $\tilde{\Phi}$, to compute the vector a and $\frac{1}{d}\|\Phi\tilde{\Phi}^T a - H_d e_j\|_F^2$.
What we can note is that if we define

$$K := \Phi\Phi^T \in \mathbb{R}^{d\times d}$$

we can connect the high-dimensional problem of computing Φ and $\tilde{\Phi}$ to some operations on K.
To be more precise, once we fix a $S_n = \{(x_i, y_i)\}_{i=1}^n$ we know that $\forall i \in \{1, \ldots, n\}$

$\exists j \in \{1, \ldots, d\}$ s.t. $x_i = h_j$, so

$$\left(\Phi\tilde{\Phi}^T\right)_{j_1 i_2} = \langle \phi\left(h_{j_1}\right), \phi\left(x_{i_2}\right)\rangle = \langle \phi\left(h_{j_1}\right), \phi\left(h_{j_2}\right)\rangle = K_{j_1 j_2}$$
$$\left(\tilde{\Phi}\tilde{\Phi}^T\right)_{i_1 i_2} = \langle \phi\left(x_{i_1}\right), \phi\left(x_{i_2}\right)\rangle = \langle \phi\left(h_{j_1}\right), \phi\left(h_{j_2}\right)\rangle = K_{j_1 j_2}$$

Computing the risk or a is an operation that can be traced down to some computation with K.
So the problem of dealing with a high dimension space like $R^{|J|}$ is now converted into computing K.
If we change a bit the notation of ϕ to

$$\phi(x_m) = (a_l^{x_m})_{l\in J}$$

We get

$$K_{mt} = \langle \phi\left(h_m\right), \phi\left(h_t\right)\rangle = \sum_{l_1+\cdots+l_d=100;\ l_1,\ldots,l_d\geq 0} a_j^{h_m} a_j^{h_t} =$$
$$= \sum_{l_1+\cdots+l_d=100;\ l_1,\ldots,l_d\geq 0} \frac{100!}{\prod_{i=1}^d l_{ji}!} \prod_{i=1}^d \left(\frac{h_{mi}}{\sqrt{d}}\right)^{l_{ji}} \left(\frac{h_{ti}}{\sqrt{d}}\right)^{l_{ji}} =$$
$$= \sum_{l_1+\cdots+l_d=100;\ l_1,\ldots,l_d\geq 0} \frac{100!}{\prod_{i=1}^d l_{ji}!} \frac{1}{d^{100}} \prod_{i=1}^d \left(h_{mi}h_{ti}\right)^{l_{ji}} = \frac{1}{d^{100}} \left(\sum_{i=1}^d h_{mi}h_{ti}\right)^{100}$$

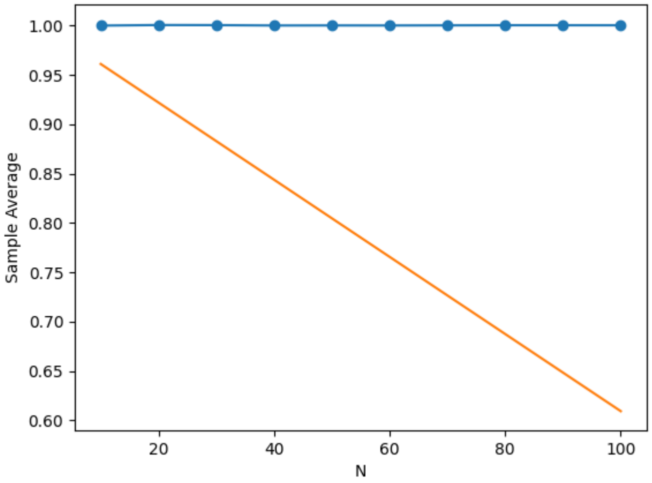
Where the last equality comes from the multinomial theorem.
Now we observe that

$$\sum_{i=1}^d h_{mi}h_{ti} = \langle h_m, h_t\rangle = \left(H_d H_d^T\right)_{mt} = d\delta_{mt}$$

Where δ_{mt} is the Kronecker delta.
Thus,

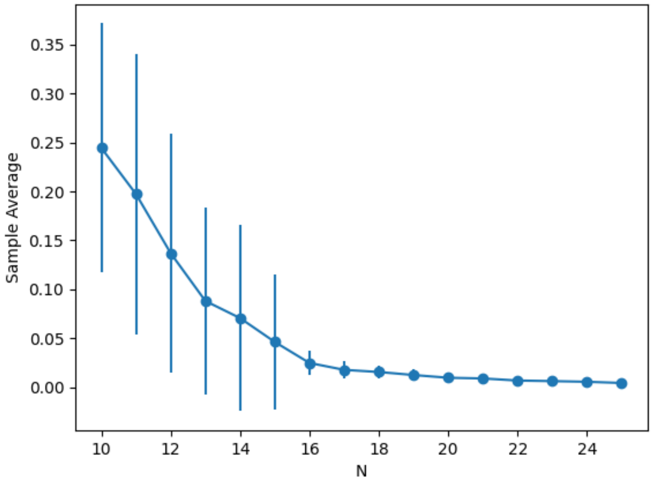
$$\frac{1}{d^{100}} \left(\sum_{i=1}^d h_{mi}h_{ti}\right)^{100} = \frac{1}{d^{100}} \left(d\delta_{mt}\right)^{100} = \delta_{mt}$$

So we have found that $K = I_d$.
Through this method we get that computing $\Phi\tilde{\Phi}$ has a complexity of $O(dn)$ and computing $\tilde{\Phi}\tilde{\Phi}$ has a complexity of $O(n^2)$.
On the next page there is the plot generated by the simulation.



Subproblem 2.8

Sample average of 50 i.i.d. random variables sampled from a uniform distribution for the index j, n times the distribution ρ_j for S_n, with $\theta(S_n)$ defined as in the exercise and $\phi(x) = x$. Plotted with 1 std deviation error bar.



A consideration that we can do is that this graph is always lower the lower bound line (in orange in the previous plot) so it means that this $\theta \notin \mathcal{A}$.