Physics 786 Spring 2023
Homework 6

Note: A Colab notebook accompanies this homework assignment. You are required to report answers to questions marked as (Notebook) in the notebook.

1. *Playing with estimators.*

   (a) Consider a random variable X, governed by the exponential distribution $\Pr(X = x) = p_X(x;\theta) = \theta e^{-\theta x}$, where $x \geq 0$. This distribution has mean $\mathbb{E}[X] = \frac{1}{\theta}$. Now consider N independently and identically distributed random variables $X_1, \cdots, X_N$, each governed by p_X. What is the maximum likelihood estimator $\hat{\theta}_{MLE}$ of the distribution p_X?

   (b) Compute the bias and variance of the MLE you found above: $\text{Bias}(\hat{\theta}_{MLE})$ and $\text{Var}(\hat{\theta}_{MLE})$?

   (c) Compute the Fisher information, considering just the single parameter θ.

   (d) Is the estimator of the mean efficient? (Recall that an estimator is efficient if its variance saturates the Cramer-Rao lower bound)

   (e) (Notebook) Numerically generate a set of N samples, $S = \{x_1, \cdots, x_N\}$. Next, generate M such sets $S_1, \cdots, S_M$. For each set of samples S_i, compute the MLE estimator $\hat{\theta}_{MLE}$. This will give M different numerical estimates $\hat{\theta}_1, \cdots, \hat{\theta}_M$. Now compute the sample variance of this set of estimates $\{\hat{\theta}_1, \cdots, \hat{\theta}_M\}$. This gives us a numerical estimate $\widehat{\text{Var}}(\hat{\theta}_{MLE})$ of the variance of the MLE, $\text{Var}(\hat{\theta}_{MLE})$. Compare this numerical estimate to the Fisher information. Try different values of N and M. For example, try $N = 10$, $M = 1000$; this should give a noisy estimate for $\hat{\theta}_{MLE}$, but since we have a large set M of samples, our variance $\widehat{\text{Var}}(\hat{\theta}_{MLE})$ should be close to the true variance $\text{Var}(\hat{\theta}_{MLE})$.

   (f) Repeat the above problems for the MLE estimator $\hat{\mu}_{MLE}$ of the mean of the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$.

2. Consider a teacher-student setup, where the teacher generates samples (x, y) with

$$y = f^*(x) = (w^*)^T x + \epsilon, \tag{1}$$

   where w^* is a fixed vector, $x \sim \mathcal{N}(0, I_{d_{in}})$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Then, the probability distribution for $y \in \mathbb{R}$ follows $\mathcal{N}(f^*(x), \sigma^2)$.

   (a) Write down the 2×2 Fisher information matrix for the distribution $\mathcal{N}(\mu, \sigma^2)$ for parameters μ and σ^2.

   (b) Suppose we fit a linear model of the form $f(x) = w^T x$ to the dataset using MSE loss, where w is the weight parameter. Show that the population loss in this case can be written as

$$L_{pop} = \mathbb{E}_{x,y}\left[\frac{1}{2}\|w^T x - y\|^2\right] = \frac{1}{2}\|w - w^*\|^2 + \frac{1}{2}\sigma^2 \tag{2}$$

   Use the Cramer-Rao bound to give a lower bound the population loss for linear regression.

   (c) (Notebook) Test the lower bound numerically by considering a dataset of N samples, using it to estimate the parameters of the linear model. Plot the average test loss and the Cramer-Rao lower bound for various values of N.

3. Consider a teacher-student setup, where the teacher generates samples (x, y) with

$$y = (w^*)^T x + \epsilon, \tag{3}$$

where ϵ has an exponential distribution, $p(\epsilon) = \frac{1}{2}\theta e^{-\theta|\epsilon|}$.

   (a) (Notebook) Perform linear regression assuming a MSE loss, and compute the generalization error.

   (b) Derive the loss using the maximum likelihood method for the above problem.

   (c) (Notebook) Now perform linear regression using the loss derived from maximum likelihood, and compute the generalization error. Hint: Use sub-gradient decent. How does the generalization error, in this case, compare to the case where MSE loss was used?

4. We saw that the naive sample variance $\hat{\sigma}^2_{naive} = \frac{1}{N}\sum_{i=1}^{N}(X_i - (\frac{1}{N}\sum_{i=1}^{N} X_i))^2$ is a biased estimator of the sample variance. Then we saw that an unbiased estimator of the variance includes the Bessel correction, $\hat{\sigma}^2_{unbiased} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - (\frac{1}{N}\sum_{i=1}^{N} X_i))^2$. But just because it is unbiased, is it the best that we can do? Let us consider a family of estimators

$$\hat{\sigma}^2_A = \frac{1}{A}\sum_{i=1}^{N}\left(X_i - \left(\frac{1}{N}\sum_{i=1}^{N} X_i\right)\right)^2. \tag{4}$$

(Notebook) Suppose that we want an estimator $\hat{\sigma}^2_A$ that minimizes the mean squared error $\mathbb{E}[(\hat{\sigma}^2_A - \sigma^2)^2]$. For normally distributed random numbers, $X \sim \mathcal{N}(0, \sigma^2)$, find the optimal value of A numerically.

Bonus problem: Show the above result analytically.

5. *James-Stein estimator* The James-Stein estimator is an estimator for the mean of a multi-variate Gaussian distribution. More specifically, let X be an m-component random variable, meaning X takes values in $\mathbb{R}^m$. The m-dimensional Gaussian distribution is

$$p(x) = (2\pi\sigma^2)^{-m/2} e^{-\frac{||x-\mu||^2}{2\sigma^2}}. \tag{5}$$

The mean is an m-component vector $\mu \in \mathbb{R}^m$, and we take the covariance matrix to be diagonal, with σ^2 the variance of each component. $||\cdot||$ refers to the L_2 norm.

The formula for the James-Stein estimator is

$$\hat{\mu}_{JS} = \left(1 - \frac{(m-2)\sigma^2}{||X||^2}\right) X. \tag{6}$$

   (a) First we will prove that for $m \geq 3$, the James-Stein estimator has smaller MSE than the usual MLE of the mean.

      i. Using the following decomposition

$$(\hat{\mu}_{JS,i} - \mu_i)^2 = (X_i - \hat{\mu}_{JS,i})^2 - (X_i - \mu_i)^2 + 2(\hat{\mu}_{JS,i} - \mu_i)(X_i - \mu_i), \tag{7}$$

      show that

$$\mathbb{E}\left[\|\hat{\mu}_{JS} - \mu\|^2\right] = \mathbb{E}\left[\|X_i - \hat{\mu}_{JS}\|^2\right] - m\sigma^2 + 2\sigma^2\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial \hat{\mu}_{JS,i}}{\partial X_i}\right] \tag{8}$$

      ii. Next, using the definition of $\hat{\mu}_{JS}$, simplify the averages $\mathbb{E}\left[\|X_i - \hat{\mu}_{JS}\|^2\right]$ and $\mathbb{E}\left[\frac{\partial \hat{\mu}_{JS,i}}{\partial x_i}\right]$ to arrive to

$$\mathbb{E}\left[\|\hat{\mu}_{JS} - \mu\|^2\right] = \mathbb{E}\left[\|\hat{\mu}_{MLE} - \mu\|^2\right] - \sigma^4\mathbb{E}\left[\frac{(m-2)^2}{\|X\|^2}\right]. \tag{9}$$

   Check that for $m \geq 3$, $\mathbb{E}\left[\|\hat{\mu}_{JS} - \mu\|^2\right] < \mathbb{E}\left[\|\hat{\mu}_{MLE} - \mu\|^2\right]$ for all μ.

(b) (Notebook) Generate a sample of X and compute the JS estimator numerically. Now generate N such samples. Using the estimator of the variance $\hat{\sigma}^2$ instead of σ^2 in Eqn. 6, show empirically that, on average, the James-Stein estimator is closer to the true mean than the MLE.

6. (Notebook)*Experimenting with bias and variance for polynomial regression* In this problem, we will compute the bias and variance in the polynomial regression problem with samples (x, y) generated using the relation

$$y = 2x^3 - x^2 + x + 1 + \eta, \tag{10}$$

where $\eta \sim \mathcal{N}(0, 1)$ is random noise.

Randomly generate a training dataset S of size N and a test example (x', y'). Fit a p degree polynomial, $f_{p,S}(x)$, to the randomly generated set S and evaluate it on the test example (x', y'). Repeat this process K times and estimate the bias $\left(\mathbb{E}_{S,(x',y')}[f_{p,S}(x') - y']\right)^2$ and variance $\mathbb{E}_{S,(x',y')}[f_{p,S}(x') - \mathbb{E}_{S,(x',y')} f_{p,S}(x')]^2$. Repeat this for different values of p. Plot the bias, variance, and test loss as a function of the degree p.

7. (Notebook)*Double descent* In this problem, we will demonstrate the sample-wise double descent phenomenon for linear regression for synthetic dataset consisting of (x, y) pairs related as

$$y = Wx + \epsilon, \tag{11}$$

where $x \sim \mathcal{N}(0, I_{d_{in}})$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.1$.

Generate subsets of the training dataset of size n and fit a linear model to this subset using the exact solution to linear regression with MSE loss. Next, plot the test loss as a function of n , and mark the underparameterized and overparameterized regimes.

1. (a) $\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} \left\{ \sum_{i=1}^{N} [\log\theta - \theta x_i] \right\}$

$$\Rightarrow \frac{N}{\theta} = \sum_{i=1}^{N} x_i \quad \hat{\theta}_{MLE} = \frac{N}{\sum_{i=1}^{N} x_i}$$

(b) $\text{bias}(\hat{\theta}_{MLE}) = \mathbb{E}[\hat{\theta}_{MLE}] - \theta = N\,\mathbb{E}\left[\frac{1}{\sum_{i=1}^{N} x_i}\right] - \theta$

$$= N\mathbb{E}\left[\frac{1}{S}\right] - \theta \quad S \sim \text{Gamma}(n, \tfrac{1}{\theta})$$

$$= \frac{N\theta}{N-1} - \theta = \frac{1}{n-1}\theta$$

$$\text{Var}(\hat{\theta}_{MLE}) = \mathbb{E}[(\hat{\theta}_{MLE} - \langle\hat{\theta}_{MLE}\rangle)^2]$$

$$= \mathbb{E}\left[\frac{N^2}{S^2}\right] - \left(\frac{N\theta}{N-1}\right)^2 = \frac{N^2\theta^2}{(N-1)(N-2)} - \left(\frac{N\theta}{N-1}\right)^2$$

$$\frac{1}{N-1}\cdot\left(\frac{1}{N-2} - \frac{1}{N-1}\right) = \frac{1}{(N-1)^2 N-2}$$

$$= \frac{N^2\theta^2}{(N-1)^2(N-2)}$$

(c) $S_{p\|q} = \sum_i p_i \log\frac{p_i}{q_i}$

$$= N\int dx\ \theta\exp(-\theta x)\log\frac{\exp(-\theta x)\,\theta}{\exp(-\theta' x)\,\theta'}$$

$$= N\int dx\ \theta\exp(-\theta x)\left(\log\frac{\theta}{\theta'} - (\theta - \theta')\right)$$

$$\frac{\partial^2 S_{p\|q}}{\partial\theta'^2} = N\int dx\ \theta\exp(-\theta x)\left(\frac{1}{\theta'^2}\right) = \frac{N}{\theta'^2}$$

$$\Rightarrow F = \frac{N}{\theta^2}$$

(d) $\dfrac{\theta^2 N^2}{(N-1)^2(N-2)} > \dfrac{N}{\theta^2}$ $\therefore$ not efficient

(f) $\hat{\mu}_{MLE} = \underset{\mu}{\operatorname{argmax}} \left\{ \sum_{i=1}^{N} \left[\log \frac{1}{\sqrt{2\pi\sigma}} - \frac{(x_i - \mu)^2}{2\sigma^2} \right] \right\}$

$$\Rightarrow \frac{N}{\theta} = \sum_{i=1}^{N} x_i \quad \hat{\theta}_{MLE} = \frac{N}{\sum_{i=1}^{N} x_i}$$

$$0 = \sum_{i=1}^{N} \frac{\mu - x_i}{2\sigma^2} \Rightarrow \hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\text{bias}(\hat{\mu}_{MLE}) = \mathbb{E}[\hat{\theta}_{MLE}] - \mu = 0$$

$$\text{Var}(\hat{\mu}_{MLE}) = \mathbb{E}[(\hat{\mu}_{MLE} - \langle \hat{\mu}_{MLE} \rangle)^2]$$

$$= \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N} x_i - \mu\right)^2\right]$$

$$= \mathbb{E}\left[\frac{1}{N^2}\sum_{i,j} x_i x_j - \frac{2\mu}{N}\sum_i x_i + \mu^2\right]$$

$$= \mathbb{E}\left[\frac{1}{N^2}\sum_{i,j} x_i x_j\right] - \mu^2$$

$$= \frac{1}{N}(\mu^2 + \sigma^2) + \frac{1}{N^2}\sum_{i \neq j} E[x_i x_j] - \mu^2$$

$$= \frac{1}{N}(\mu^2 + \sigma^2) + \frac{N(N-1)}{N^2}\mu^2 - \mu^2 = \frac{\sigma^2}{N}.$$

$$F = \text{Var}\left[\frac{\partial}{\partial \mu}\sum_{i=1}^{N}\left(\log\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x_i - \mu)^2}{2\sigma^2}\right)\right]$$

$$= \text{Var}\left[\sum_{i=1}^{N}\frac{x_i - \mu}{\sigma^2}\right) = \frac{1}{\sigma^4}\text{Var}\left(\sum_{i=1}^{N}(x_i - \mu)\right) = \frac{N}{\sigma^2}$$

$$\partial_\mu \mathbb{E}[\hat{\mu}_{MLE}] = 1$$

$\Rightarrow$ CRLB $= \frac{\sigma^2}{N}$: efficient

2. (a) $F = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$

(b) $L_{pop} = \mathbb{E}_{x,y}\left[\frac{1}{2} \| w^T x - y \|^2 \right]$

$$= \int d^d x \, dy \ \frac{1}{2} \| w^T x - y \|^2 \frac{1}{\sqrt{2\pi}^d} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - f^*(x))^2}{2\sigma^2}}$$

$$= \frac{1}{2} \int d^d x \, dy \ \left(y^2 - 2 y w^T x + (w^T x)^2 \right) \frac{1}{\sqrt{2\pi}^d} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y - f^*(x))^2}{2\sigma^2}}$$

$$= \frac{1}{2} \int d^d x \ \left(f^*(x)^2 + \sigma^2 - 2 f^*(x) + (w^T x)^2 \right) \frac{1}{\sqrt{2\pi}^d} e^{-x^2/2}$$

$$= \frac{\sigma^2}{2} + \frac{1}{2} \| w - w^* \|^2 \qquad f^*(x) = w^{*T} x$$

$$\langle (\hat{y} - y)^2 \rangle = Var(\hat{y}) + Bias(\hat{y}) + \sigma^2$$

$$L_{pop} = \frac{1}{2} Var(\hat{y}) + \frac{1}{2} \sigma^2 \qquad Var(\hat{y}) \geq \frac{\sigma^2}{N} \text{ from 1. (f)}$$

$$\Rightarrow \quad L_{pop} \geq \frac{1}{2} \frac{\sigma^2}{N} + \frac{1}{2} \sigma^2 = \frac{1}{2} \frac{N+1}{N} \sigma^2$$

3. (b) $L = \frac{1}{N} \sum_{i=1}^{N} \log p(y_i | x_i, \theta)$

$\leftarrow$ constant

$= \frac{1}{N} \sum_{i=1}^{N} \left[\theta | y_i - w^{*T} x_i | - N \log \frac{\theta}{2} \right]$

$\Rightarrow L = \frac{1}{N} \sum_{i=1}^{N} | y_i - w^{*T} x_i |$

4. $L = E[(\hat{\sigma}_A^2 - \sigma^2)^2]$

$$= E\left[\frac{1}{A^2}\left(\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{j=1}^{N}x_j\right)^2\right)^2\right]$$

$$- 2\frac{\sigma^2}{A} E\left[\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{j=1}^{N}x_j\right)^2\right] + \sigma^4$$

$$= \frac{1}{A^2}\sigma^4(N^2-1) - 2\frac{\sigma^2}{A}\sigma^2(N-1) + \sigma^4$$

$$\frac{\partial L}{\partial A} = -\frac{2}{A^3}\sigma^4(N^2-1) + \frac{2\sigma^4}{A^2}(N-1) = 0$$

$$\Rightarrow A = N+1$$

3. (a) (i) $(\hat{\mu}_{JS,i} - \mu_i)^2 = (X_i - \hat{\mu}_{JS,i})^2 - (X_i - \mu_i)^2 + 2(\hat{\mu}_{JS,i} - \mu_i)(X_i - \mu_i)$

$$E[\|\hat{\mu}_{JS} - \mu\|^2] = E[\|X - \hat{\mu}_{JS}\|^2] - E(\|X-\mu\|^2) + 2\sum_{i=1}^{m} E((\hat{\mu}_{JS,i} - \mu_i)(X_i - \mu_i))]$$

$$= E[\|X - \hat{\mu}_{JS}\|^2] - m\sigma^2 + 2\sigma^2 \sum_{i=1}^{m} E\left[\frac{\partial \hat{\mu}_{JS,i}}{\partial X_i}\right]$$

by Stein's Lemma

(ii) $$X - \hat{\mu}_{JS} = -\frac{(m-2)\sigma^2}{\|X\|^2} X$$

$$E[\|X - \hat{\mu}_{JS}\|^2] = (m-2)^2\sigma^4 E\left[\frac{1}{\|X\|^2}\right]$$

$$\frac{\partial \hat{\mu}_{JS,i}}{\partial x_i} = 1 - \frac{(m-2)\sigma^2}{\sum_j x_j^2} + \frac{2(m-2)\sigma^2}{(\sum_j x_j^2)^2} x_i^2$$

$$\sum_{i=1}^{m} E\left[\frac{\partial \hat{\mu}_{JS}}{\partial x_i}\right] = m - m(m-2)\sigma^2 E\left[\frac{1}{\|X\|^2}\right] + 2(m-2)\sigma^2 E\left[\frac{1}{\|X\|^2}\right]$$

$$= m - (m-2)^2\sigma^2 E\left[\frac{1}{\|X\|^2}\right]$$

$$E[\|\hat{\mu}_{JS} - \mu\|^2] = E[\|X - \hat{\mu}_{JS}\|^2] - m\sigma^2 + 2\sigma^2 \sum_{i=1}^{m} E\left[\frac{\partial \hat{\mu}_{JS,i}}{\partial x_i}\right]$$

$$= (m-2)^2\sigma^4 E\left[\frac{1}{\|X\|^2}\right] - m\sigma^2 + 2\sigma^2 m - 2(m-2)^2\sigma^4 E\left[\frac{1}{\|X\|^2}\right]$$

$$= m\sigma^2 - (m-2)^2\sigma^4 E\left[\frac{1}{\|X\|^2}\right]$$

$$= E[\|\hat{\mu}_{MLE} - \mu\|^2] - (m-2)^2\sigma^4 E\left[\frac{1}{\|X\|^2}\right]$$

$\Rightarrow \quad E[\|\hat{\mu}_{JS} - \mu\|^2] < E[\|\hat{\mu}_{MLE} - \mu\|^2]$ for $m \geq 3$.