Assignment 2

Yang Can, HKUST

Problem 1

Let $\mathcal{D} = \{X, y\}$ be the collected data, where $X \in \mathbb{R}^{n\times p}$ is the design matrix with full rank and $y \in \mathbb{R}^n$ is the vector of response. Consider the following optimization problem

$$\hat{\beta} = \arg\min\{\|b\|_2 : b \text{ minimizes } \frac{1}{2n}\|y - Xb\|_2^2\}. \tag{1}$$

(a) Show that the optimal solution of problem (1) is

$$\hat{\beta} = (X^TX)^{-1}X^Ty \text{ when } n \geq p,$$

and

$$\hat{\beta} = X^T(XX^T)^{-1}y \text{ when } n < p.$$

What is the degrees of freedom based on Stein's lemma

$$df = \mathbb{E}\left[\sum_i \frac{\partial \hat{y}_i}{\partial y_i}\right]?$$

(b) Mikhail Belkin et al. (2019) PNAS paper "Reconciling modern machine-learning practice and the classical bias–variance trade-off" demonstrated the **double decent** phenomenon for many machine learning methods. Let $\gamma = p/n$. Can you use simulation study to demonstrate the **double decent** phenomenon with the above linear model in the underparameterized regime ($\gamma < 1$), overparameterized regime ($\gamma > 1$) and the special regime ($\gamma = 1$)? It would be great if you can show the pattern of bias-variance tradeoff in these different regimes. For example, you may use the value of γ as the x-axis, and use squared bias and variance as the y-axis to visualize the bias-variance tradeoff. Of course, **the answer to this part** is quite open.

(c) Intialize $\beta^{(0)} = 0$, and gradient descent on the least square loss yields

$$\beta^k = \beta^{k-1} + \epsilon\frac{X^T}{n}(y - X\beta^{(k-1)}), \tag{2}$$

where we take $0 < \epsilon \leq 1/\lambda_{max}(X^TX/n)$ (and $1/\lambda_{max}(X^TX/n)$ is the largest eigenvalue of X^TX/n). Will the gradient descent converge to the optimal solution given in (a)? Please justify your answer.

(d) After rearranging (2), we find

$$\frac{\beta^k - \beta^{(k-1)}}{\epsilon} = \frac{X^T}{n}(y - X\beta^{(k-1)}),$$

Setting $\beta(t) = \beta(k)$ at time $t = k\epsilon$, we have the left-hand side as the discrete derivative of $\beta(t)$ at time t, which approaches its continous-time derivative as $\epsilon \to 0$:

$$\frac{d\beta(t)}{dt} = \frac{X^T}{n}(y - X\beta(t)), \tag{3}$$

over time $t \geq 0$, subject to an intial condition $\beta(0) = 0$. This is called the **gradient flow differential equation** for the least squars problem $\min \frac{1}{2n}\|y - Xb\|^2$. What is the exact solution path $\beta(t)$ to (3) for all t?

(e) Now consider the ridge regression problem

$$\min_b \frac{1}{2n}\|y - Xb\|_2^2 + \lambda\|b\|_2^2, \tag{4}$$

where $\lambda > 0$ is a tuning parameter. The closed-form solution is

$$\hat{\beta}(\lambda) = (X^T X + n\lambda I)^{-1} X^T y. \tag{5}$$

Use simulation study to investigate the differences between the solution of ridge regression $\hat{\beta}(\lambda)$ given in (5) and the solution of gradient flow $\hat{\beta}(t)$, e.g., you can compare the similarity of their solution paths, and their prediction accuracies along the solution paths. Again, **the answer to this part** is quite open.

Remark: The following optimization problem is known as *compressed sensing*

$$\hat{\beta} = \arg\min\{\|b\|_1 : b \text{ minimizes } \frac{1}{2n}\|y - Xb\|_2^2\}. \tag{6}$$

Problem 2

Consider an extension of the James-Stein estimator problem. Suppose we have z-values from groups A and B: $z_{A,i}|\mu_{A,i} \sim N(\mu_{A,i}, 1)$ and $z_{B,i}|\mu_{B,i} \sim N(\mu_{B,i}, 1)$, $i = 1, \ldots, N$. Assuming

$$\begin{pmatrix} \mu_{A,i} \\ \mu_{B,i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right), \tag{7}$$

where

$$\Sigma = \begin{pmatrix} \sigma_A^2 & \rho\sigma_A\sigma_B \\ \rho\sigma_A\sigma_B & \sigma_B^2 \end{pmatrix}$$

is a 2-by-2 positive matrix.

(a) Let $\hat{\mu}_A^{(\rho)} = [\hat{\mu}_{A,1}^{(\rho)}, \ldots, \hat{\mu}_{A,N}^{(\rho)}]^T$ and $\hat{\mu}_B^{(\rho)} = [\hat{\mu}_{B,1}^{(\rho)}, \ldots, \hat{\mu}_{B,N}^{(\rho)}]^T$ be the estimated posterior means. Derive an algorithm to estimate Σ, and obtain $\hat{\mu}_A^{(\rho)}$ and $\hat{\mu}_B^{(\rho)}$ (Clearly, it reduces to the standard JSE problem when $\rho = 0$).

(b) Conduct simulation study to compare its performance with the standard JSE in term of the following expected total squared losses:

$$\ell_A = \mathbb{E}\left[\|\hat{\mu}_A - \mu_A\|_2^2\right] \text{ and } \ell_B = \mathbb{E}\left[\|\hat{\mu}_B - \mu_B\|_2^2\right].$$

It would be better to consider some situations in presence of *model misspecification*.

(c) Make some discussions based on your simulation results.

Remark: This is an example to illustrate how to borrow inforamtion across two different tasks A and B *in a statistically rigorous manner*. A similar idea can be applied to multi-task learning problems by exploring their *correlation*.

Problem 3

Consider the dataset given by "score.txt". Suppose that the generative model is known as

$$z_{ij} = \beta_0 + \mu_i + \eta_j + \epsilon_{ij}, i = 1, \ldots, 30; j = 1, \ldots, 15, \tag{8}$$

where z_{ij} is the score of an individual who is studying the j-th subject in the i-th school, $\mu_i \sim \mathcal{N}(0, \sigma_1^2)$ is the effect of the i-th school, and $\eta_j \sim \mathcal{N}(0, \sigma_2^2)$ is the effect of the j-th subject school, and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is the independent noise.

(a) Derive an algorithm to estimate $\boldsymbol{\theta} = \{\beta_0, \sigma_1^2, \sigma_2^2, \sigma_\epsilon^2\}$.

(b) Obtain the posterior mean of μ_i and η_j, i.e., $\mathbb{E}(\mu_i|\mathcal{D}; \hat{\boldsymbol{\theta}})$ and $\mathbb{E}(\eta_i|\mathcal{D}; \hat{\boldsymbol{\theta}})$, where $\mathcal{D} = \{z_{ij}\}_{i=1,\ldots,30; j=1,\ldots,15}$.

Remark: You need to implement the algorithm by yourself rather than calling packages.

Requirement

- You need to submit a report, in which you should clearly describe your method and explain your idea. The code should also be included.
- You can use R or Python for coding.
- Your report should be in the **pdf** or **html** format, which is automatically generated by either R markdown or Jupyter notebook.
- The report is due to October 7, 2021, 11:59pm.