Problem 1.1: Linear Regression

We can have two approaches for solving this problem. Both of them are correct.

Matrix way

Part (a) *[5 points]*

Since the noise terms are independent, we can write the covariance matrix of them Σ as a diagonal matrix with elements $(\sigma_1^2, \cdots, \sigma_N^2)$. The inverse of this matrix is another diagonal matrix with elements $(\sigma_1^{-2}, \cdots, \sigma_N^{-2})$. Thus, using the pdf of multivariate normal distribution, we can write the likelihood of the data as follows:

$$P(D) = (2\pi)^{-\frac{N}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta})^\top \Sigma^{-1}(\boldsymbol{y} - X\boldsymbol{\beta})\right).$$

The negative log likelihood can be written as:

$$\begin{aligned} -\log P(D) &= \frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta})^\top \Sigma^{-1}(\boldsymbol{y} - X\boldsymbol{\beta}) + \text{const.}, \\ &= \frac{1}{2}\left\|H(\boldsymbol{y} - X\boldsymbol{\beta})\right\|_2^2, \end{aligned}$$

where in the last step H is a diagonal matrix with $(\sigma_1^{-1}, \cdots, \sigma_N^{-1})$ on its diagonal.

Part (b) *[5 points]*

Now, observing the fact that by defining $\tilde{\boldsymbol{y}} = H\boldsymbol{y}$ and $\tilde{X} = HX$, we have a ordinary regression problem, we can use its solution and write:

$$\widehat{\boldsymbol{\beta}} = (\tilde{X}^\top \tilde{X})^{-1}\tilde{X}^\top \tilde{\boldsymbol{y}} = (X^\top H^\top H X)^{-1} X^\top H^\top H \boldsymbol{y} = (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} \boldsymbol{y}.$$

Summation way

Part (a)

$$y_n = \boldsymbol{x}_n^\top \boldsymbol{\beta} + \varepsilon_n$$

Here, $\varepsilon \sim \mathcal{N}(0, \sigma_n)$, σ_n not given to be equal for all n. First, we note that each y_n is taken from the distribution $\mathcal{N}(\boldsymbol{x}_n\boldsymbol{\beta}, \sigma_n)$. Thus:

$$P(y_n|\boldsymbol{\beta}, \boldsymbol{x}_n) = (2\pi\sigma_n^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma_n^2}(y_n - \boldsymbol{x}_n^\top\boldsymbol{\beta})^2\right\}$$

$$P(D) = \prod_n^N P(y_n|\boldsymbol{\beta}, \boldsymbol{x}) = \prod_n^N (2\pi\sigma_n^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma_n^2}(y_n - \boldsymbol{x}_n^\top\boldsymbol{\beta})^2\right\}$$

$$\log P(D) = \sum_n^N \log(P(y_n|\boldsymbol{\beta}, \boldsymbol{x})) = \sum_n^N \left[-\frac{1}{2}\log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2}(y_n - \boldsymbol{x}_n^\top\boldsymbol{\beta})^2\right]$$

Part (b)

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log P(D) = \sum_n^N \frac{1}{\sigma_n^2}\left(y_n - \boldsymbol{x}_n^\top\boldsymbol{\beta}\right)\boldsymbol{x}_n^\top$$

Set equal to zero and rearrange.

$$\sum_n^N \frac{1}{\sigma_n^2}\boldsymbol{x}_n^\top\boldsymbol{\beta}\boldsymbol{x}^\top = \sum_n^N \frac{1}{\sigma_n^2} y_n \boldsymbol{x}_n^\top$$

$$\boldsymbol{\beta}^\top \sum_n^N \frac{1}{\sigma_n^2} \boldsymbol{x}_n \boldsymbol{x}_n^\top = \sum_n^N \frac{1}{\sigma_n^2} \boldsymbol{x}_n^\top y_n$$

$$\hat{\boldsymbol{\beta}}^\top = \left[\sum_n^N \frac{1}{\sigma_n^2} \boldsymbol{x}_n \boldsymbol{x}_n^\top \right]^{-1} \sum_n^N \frac{1}{\sigma_n^2} \boldsymbol{x}_n^\top y_n$$

Problem 1.2: Smooth Coefficients

Part (a)

The regularizer representing $(\beta_i - \beta_{i+1})^2$ is given by: $\sum_i^{p-1} (\beta_i - \beta_{i+1})^2$. This can be rearranged into vector form. Define matrix $D \in \mathbb{R}^p$ as the following

$$D = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 1 & -1 & 0 \\ 0 & \cdots & 0 & 1 & -1 \\ 0 & \cdots & \cdots & 0 & 0 \end{bmatrix}$$

The regularizer is then *[5 points]* :

$$\boldsymbol{\beta}^\top D^\top D \boldsymbol{\beta} = \|D\boldsymbol{\beta}\|_2^2$$

The full optimization problem is *[5 points]* :

$$L(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 + \mu \|D\boldsymbol{\beta}\|_2^2$$

$$L(\boldsymbol{\beta}) = (\boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta} + \mu \boldsymbol{\beta}^\top D^\top D \boldsymbol{\beta}$$

for some λ and μ hyper-parameters.

Part (b) *[5 points]*

$$\nabla L(\boldsymbol{\beta}) = 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{X}^\top \boldsymbol{y} + 2\lambda\boldsymbol{\beta} + 2\mu D^\top D\boldsymbol{\beta}$$

Set equal to zero.

$$(\boldsymbol{X}^\top \boldsymbol{X} + \lambda I_p + \mu D^\top D)\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{y}$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda I_p + \mu D^\top D)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

Problem 1.3: Constrained Linear Regression

Here we take the immediate jump to the L_2 minimization problem, as presented in class.

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2 \quad \text{s.t.} \quad A\boldsymbol{\beta} = \boldsymbol{b}$$

Write the Lagrangian.

$$\mathcal{L}(\boldsymbol{\beta}, \lambda) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2 + \boldsymbol{\lambda}^\top (A\boldsymbol{\beta} - \boldsymbol{b})$$

Take the derivative with respect to β.

$$\frac{\partial}{\partial \boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta},\lambda) = 2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{X}^\top\boldsymbol{y} + A^\top\boldsymbol{\lambda}$$

Set equal to zero and solve for β:

$$\boldsymbol{\beta} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} - \frac{1}{2}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}A^\top\boldsymbol{\lambda}$$

Use the constraint by applying A to both sides.

$$A\boldsymbol{\beta} = \boldsymbol{b} = A(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y} - \frac{1}{2}A(\boldsymbol{X}^\top\boldsymbol{X})^{-1}A^\top\boldsymbol{\lambda}$$

Solve for λ.

$$A(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} - \boldsymbol{b} = \frac{1}{2}A(\boldsymbol{X}^\top\boldsymbol{X})^{-1}A^\top\boldsymbol{\lambda}$$

$$\boldsymbol{\lambda} = 2(A(\boldsymbol{X}^\top\boldsymbol{X})^{-1}A^\top)^{-1}(A(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} - \boldsymbol{b})$$

Plug $\boldsymbol{\lambda}$ into the derivative.

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}A^\top(A(\boldsymbol{X}^\top\boldsymbol{X})^{-1}A^\top)^{-1}(A(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} - \boldsymbol{b})$$

Problem 2: Perceptron

Conditions: From any current step parameters $\boldsymbol{w}_i$ we want to update the classifier such that $\text{sign}(\boldsymbol{w}_{i+1}^\top\boldsymbol{x}_{i+1}) = y_{i+1}$. However, we also would like $\|\boldsymbol{w}_{i+1} - \boldsymbol{w}_i\|_2$ to be small.

Solution: If $y_{i+1} = \text{sign}(\boldsymbol{w}_{i+1}^\top\boldsymbol{x}_{i+1})$, then let $\boldsymbol{w}_{i+1} = \boldsymbol{w}_i$ (do nothing). Otherwise we need the smallest amount of movement such that then point $\boldsymbol{x}_{i+1}$ is on the correct side of the plane. Do the following:

$$\boldsymbol{w}_{i+1} = \arg\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_i\|_2^2 \quad \text{s.t.} \quad \boldsymbol{w}^\top\boldsymbol{x}_{i+1}y_{i+1} = 0$$

Writing the Lagrangian, yields

$$\mathcal{L}(\boldsymbol{w},\lambda) = \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_i)^\top(\boldsymbol{w} - \boldsymbol{w}_i) + \lambda\boldsymbol{w}^\top\boldsymbol{x}_{i+1}y_{i+1}$$

Take a derivative w.r.t. $\boldsymbol{w}$

$$\frac{\partial}{\partial \boldsymbol{w}}\mathcal{L}(\boldsymbol{w},\lambda) = (\boldsymbol{w} - \boldsymbol{w}_i) - \lambda\boldsymbol{x}_{i+1}y_{i+1} = 0$$

$$\boldsymbol{w} = \lambda\boldsymbol{x}_{i+1}y_{i+1} + \boldsymbol{w}_i$$

Transpose and multiply by $\boldsymbol{x}_{i+1}y_{i+1}$ on both sides, then apply the equality $\boldsymbol{w}^\top\boldsymbol{x}_{i+1}y_{i+1} = 0$:

$$\boldsymbol{w}^\top\boldsymbol{x}_{i+1}y_{i+1} = 0 = \lambda(\boldsymbol{x}_{i+1}y_{i+1})^\top(\boldsymbol{x}_{i+1}y_{i+1}) + \boldsymbol{w}_i^\top(\boldsymbol{x}_{i+1}y_{i+1})$$

$$\lambda = -\frac{\boldsymbol{w}_i^\top(\boldsymbol{x}_{i+1}y_{i+1})}{\|\boldsymbol{x}_{i+1}\|_2^2}$$

Plug back in, and let this be the update rule.

$$\boldsymbol{w}_{i+1} = \boldsymbol{w}_i - \frac{\boldsymbol{w}_i^\top \boldsymbol{x}_{i+1}}{\|\boldsymbol{x}_{i+1}\|_2^2}\boldsymbol{x}_{i+1}$$

Geometrically this is the same as finding some vector that is perpendicular to $\boldsymbol{x}_{i+1}$ and projecting $\boldsymbol{w}_i$ onto it, taking the projection as the new normal vector.

Problem 3

Part (a) *[5 points]* : Given:

$$K_3 = a_1 K_1 + a_2 K_2$$

where $a_1, a_2 \geq 0$ and K_1, K_2 positive semi-definite. For any $\mathbf{x} \in \mathbb{R}^N$:

$$\mathbf{x}^\top K_3 \mathbf{x} = \mathbf{x}^\top (a_1 K_1 + a_2 K_2)\mathbf{x} = a_1 \mathbf{x}^\top K_1 \mathbf{x} + a_2 \mathbf{x}^\top K_2 \mathbf{x}$$

By assumption for any $\mathbf{x} \in \mathbb{R}^N$ both $\mathbf{x}^\top K_1 \mathbf{x} \geq$ and $\mathbf{x}^\top K_2 \mathbf{x} \geq 0$ (definition of positive semi-definite). Thus, the non-negative combination of the two is also ≥ 0. So $\mathbf{x}^\top K_3 \mathbf{x} \geq 0$.

Part (b) *[5 points]* : Given:

$$K_4 : k_4(\boldsymbol{x}, \boldsymbol{x}') = f(\boldsymbol{x})f(\boldsymbol{x}')$$

for any real valued function f.

Let $\boldsymbol{f} = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_1))$. We can write $K_4 = \boldsymbol{f}\boldsymbol{f}^\top$, thus $\mathbf{x}^\top K_4 \mathbf{x} = (\mathbf{x}^\top \boldsymbol{f})^2 \geq 0$ for any vector $\mathbf{x} \in \mathbb{R}^N$.

Part (c) *[5 points]* : Given:

$$K_5 : k_5(\boldsymbol{x}, \boldsymbol{x}') = k_1(\boldsymbol{x}, \boldsymbol{x}')k_2(\boldsymbol{x}, \boldsymbol{x}')$$

We can also see that

$$K_5 = K_1 \circ K_2$$

where $\circ$ is the elementwise product. K_5 follows from the following proof. We also use an identity of the elementwise product, namely $\boldsymbol{x}^\top (A \circ B)\boldsymbol{y} = \operatorname{tr}(A \operatorname{diag}(\boldsymbol{x}) B \operatorname{diag}(\boldsymbol{y}))$.

$$\boldsymbol{x}^\top K_1 \circ K_2 \boldsymbol{x} = \operatorname{tr}(K_1 \operatorname{diag}(\boldsymbol{x}) K_2 \operatorname{diag}(\boldsymbol{x}))$$

K_1 and K_2 are assumed to be positive semi-definite, so they admit a root. $K_1 = (K_1^{\frac{1}{2}})(K_1^{\frac{1}{2}})$, $K_2 = (K_2^{\frac{1}{2}})(K_2^{\frac{1}{2}})$.

$$\operatorname{tr}((K_1^{\frac{1}{2}})(K_1^{\frac{1}{2}}) \operatorname{diag}(\boldsymbol{x})(K_2^{\frac{1}{2}})(K_2^{\frac{1}{2}}) \operatorname{diag}(\boldsymbol{x})) = \operatorname{tr}((K_1^{\frac{1}{2}}) \operatorname{diag}(\boldsymbol{x})(K_2^{\frac{1}{2}})(K_2^{\frac{1}{2}}) \operatorname{diag}(\boldsymbol{x})(K_1^{\frac{1}{2}}))$$

This is equivalent to $\operatorname{tr}(A^\top A)$ for some matrix A, which is the trace of a gram matrix, which is always greater than zero. Thus $K_1 \circ K_2$ is positive semi-definite.

—Solutions based on interpretation of this kernel as the covariance of a random vector which is obtained by elementwise product of two independent random vector with covariances K_1 and K_2 is also acceptable. See the Wikipedia page for Schur product theorem.

Problem 4

Part (a) *[3 points]* The closed form solution for $\widehat{\boldsymbol{\beta}}_\lambda$ can be written as follows:

$$\widehat{\boldsymbol{\beta}}_\lambda = (X^\top X + \lambda I)^{-1} X^\top \mathbf{y} = (X^\top X + \lambda I)^{-1} X^\top (X\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon})$$

Given the theorem about affine transformation of Gaussian random vectors, we can see that $\widehat{\boldsymbol{\beta}}_\lambda$ will be a Gaussian random vector with the following mean and variance:

$$\widehat{\boldsymbol{\beta}}_\lambda \sim \mathcal{N}\left((X^\top X + \lambda I)^{-1} X^\top X \boldsymbol{\beta}^\star, (X^\top X + \lambda I)^{-1} X^\top X (X X^\top + \lambda I)^{-1}\right).$$

Part (b) *[5 points]* Using part (a), we can write the bias as follows:

$$\mathbb{E}[\mathbf{x}^\top \widehat{\boldsymbol{\beta}}_\lambda - \mathbf{x}^\top \boldsymbol{\beta}^\star] = \mathbf{x}^\top \mathbb{E}\left[\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^\star\right] = \mathbf{x}^\top \left((X^\top X + \lambda I)^{-1} X^\top X - I\right)\boldsymbol{\beta}^\star.$$

Part (c) *[5 points]* For the variance part, using the theorem about affine transformation of Gaussian random vectors again, we realize that $\mathbf{x}^\top(\widehat{\boldsymbol{\beta}}_\lambda - \mathbb{E}[\widehat{\boldsymbol{\beta}}_\lambda])$ is a zero-mean Gaussian random variable with variance $\mathbf{x}^\top (X^\top X + \lambda I)^{-1} X^\top X (X X^\top + \lambda I)^{-1}\mathbf{x} = \|X(XX^\top + \lambda I)^{-1}\mathbf{x}\|_2^2$. Thus, because the square of a Gaussian variable is a χ^2 random variable, we can use the mean of χ^2 random variable to conclude:

$$\mathbb{E}\left[\left(\mathbf{x}^\top(\widehat{\boldsymbol{\beta}}_\lambda - \mathbb{E}[\widehat{\boldsymbol{\beta}}_\lambda])\right)^2\right] = \|X(XX^\top + \lambda I)^{-1}\mathbf{x}\|_2^2.$$

Part (d) *[2 points]* The bias and variance trade-off can be written as:

$$\mathbb{E}\left[\left(\mathbf{x}^\top \widehat{\boldsymbol{\beta}}_\lambda - \mathbf{x}^\top \boldsymbol{\beta}^\star\right)^2\right] = (\mathbf{x}^\top \left((X^\top X + \lambda I)^{-1} X^\top X - I\right)\boldsymbol{\beta}^\star)^2 + \|X(XX^\top + \lambda I)^{-1}\mathbf{x}\|_2^2 + \text{const}.$$

It is clear that as λ increases, the bias term increases and the variance term decreases.