Introduction to Machine Learning
Fall 2023
University of Science and Technology of China

Lecturer: Jie Wang — Homework 2
Posted: Oct. 10, 2023 — Due: Oct. 19, 2023

Notice, to get the full credits, please present your solutions step by step.

Exercise 1: Projection

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{x} \in \mathbb{R}^m$. Define

$$\mathbf{P}_{\mathbf{A}}(\mathbf{x}) = \underset{\mathbf{z}\in\mathbb{R}^m}{\mathbf{argmin}} \{\|\mathbf{x}-\mathbf{z}\|_2 : \mathbf{z} \in \mathcal{C}(\mathbf{A})\}.$$

We call $\mathbf{P}_{\mathbf{A}}(\mathbf{x})$ the projection of the point $\mathbf{x}$ onto the column space of $\mathbf{A}$.

1. Please prove that $\mathbf{P}_{\mathbf{A}}(\mathbf{x})$ is unique for any $\mathbf{x} \in \mathbb{R}^m$.
2. Let $\mathbf{v}_i \in \mathbb{R}^n$, $i = 1, \ldots, d$ with $d \le n$, which are linearly independent.
   (a) For any $\mathbf{w} \in \mathbb{R}^n$, please find $\mathbf{P}_{\mathbf{v}_1}(\mathbf{w})$, which is the projection of $\mathbf{w}$ onto the subspace spanned by $\mathbf{v}_1$.
   (b) Please show $\mathbf{P}_{\mathbf{v}_1}(\cdot)$ is a linear map, i.e.,
   $$\mathbf{P}_{\mathbf{v}_1}(\alpha\mathbf{u} + \beta\mathbf{w}) = \alpha\mathbf{P}_{\mathbf{v}_1}(\mathbf{u}) + \beta\mathbf{P}_{\mathbf{v}_1}(\mathbf{w}),$$
   where $\alpha, \beta \in \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}^n$.
   (c) Please find the projection matrix corresponding to the linear map $\mathbf{P}_{\mathbf{v}_1}(\cdot)$, i.e., find the matrix $\mathbf{H}_1 \in \mathbb{R}^{n\times n}$ such that
   $$\mathbf{P}_{\mathbf{v}_1}(\mathbf{w}) = \mathbf{H}_1\mathbf{w}.$$
   (d) Let $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_d)$, and $\mathbf{v}_1, \ldots, \mathbf{v}_d$ are linearly independent.
      i. For any $\mathbf{w} \in \mathbb{R}^n$, please find $\mathbf{P}_{\mathbf{V}}(\mathbf{w})$ and the corresponding projection matrix $\mathbf{H}$.
      ii. Please find $\mathbf{H}$ if we further assume that $\mathbf{v}_i^\top \mathbf{v}_j = 0$, $\forall\, i \neq j$.
3. (a) Suppose that
   $$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$
   What are the coordinates of $\mathbf{P}_{\mathbf{A}}(\mathbf{x})$ with respect to the column vectors in $\mathbf{A}$ for any $\mathbf{x} \in \mathbb{R}^2$? Are the coordinates unique?
   (b) Suppose that
   $$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix}.$$
   What are the coordinates of $\mathbf{P}_{\mathbf{A}}(\mathbf{x})$ with respect to the column vectors in $\mathbf{A}$ for any $\mathbf{x} \in \mathbb{R}^2$? Are the coordinates unique?

4. (Optional) A matrix $\mathbf{P}$ is called a projection matrix if $\mathbf{Px}$ is the projection of $\mathbf{x}$ onto $\mathcal{C}(\mathbf{P})$ for any $\mathbf{x}$.

   (a) Let λ be the eigenvalue of $\mathbf{P}$. Show that λ is either 1 or 0. (*Hint: you may want to figure out what the eigenspaces corresponding to $\lambda = 1$ and $\lambda = 0$ are, respectively.*)

   (b) Show that $\mathbf{P}$ is a projection matrix if and only if $\mathbf{P}^2 = \mathbf{P}$ and $\mathbf{P}$ is symmetric.

5. (Optional) Let $\mathbf{B} \in \mathbb{R}^{m\times s}$ and $\mathcal{C}(\mathbf{B})$ be its column space. Suppose that $\mathcal{C}(\mathbf{B})$ is a proper subspace of $\mathcal{C}(\mathbf{A})$. Is $\mathbf{P_B}(\mathbf{x})$ the same as $\mathbf{P_B}(\mathbf{P_A}(\mathbf{x}))$? Please show your claim rigorously.

Solution: 1. When $\|\mathbf{x}-\mathbf{z}\|_2^2$ reaches its minimum, $\|\mathbf{x}-\mathbf{z}\|_2$ also reaches its minimum. Next we can just analyse $\|\mathbf{x}-\mathbf{z}\|_2^2$.

$$\frac{\partial}{\partial \mathbf{z}}\|\mathbf{x}-\mathbf{z}\|_2^2 = -2(\mathbf{x}-\mathbf{Az})^\top \mathbf{A} = 0$$

We can get $\mathbf{z} = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{x}$. This is one of the extreme point of function $\|\mathbf{x}-\mathbf{z}\|_2$. And because $\|\mathbf{x}-\mathbf{z}\|_2$ is a convex function, it should get its minimum value in this point. It means the solution is $\mathbf{z} = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{x}$ and it is unique for any $x \in \mathbb{R}^m$.

2. (a)

$$\mathbf{P}_{\mathbf{v}_1}(\mathbf{w}) = \underset{\mathbf{z}\in\mathbb{R}^n}{\mathbf{argmin}}\{\|\mathbf{w}-\mathbf{z}\|_2 : \mathbf{z} = \lambda\mathbf{v_1}\}.$$

For convenience, we consider $\|\mathbf{w}-\mathbf{z}\|_2^2$,

$$\begin{aligned}\frac{\partial}{\partial\lambda}\|\mathbf{x}-\mathbf{z}\|_2^2 &= \frac{\partial}{\partial\lambda}(\mathbf{w}\top\mathbf{w} - 2\lambda\mathbf{w}^\top\mathbf{v}_1 + \lambda^2\mathbf{v_1}^\top\mathbf{v_1}) \\ &= -2\mathbf{w}^\top\mathbf{v}_1 + 2\lambda\mathbf{v}_1^\top\mathbf{v}_1 \\ &= 0\end{aligned}$$

We can get $\lambda = \frac{\mathbf{w}^\top\mathbf{v_1}}{\mathbf{v}_1^\top\mathbf{v}_1}$. It means $\|\mathbf{w}-\mathbf{z}\|_2$ can get its minimum value at this point. Therefore, $\mathbf{P}_{\mathbf{v}_1}(\mathbf{w}) = \frac{\mathbf{w}^\top\mathbf{v_1}}{\mathbf{v}_1^\top\mathbf{v}_1}\mathbf{v}_1$.

(b)

$$\begin{aligned}\mathbf{P}_{\mathbf{v}_1}(\alpha\mathbf{u}+\beta\mathbf{w}) &= \frac{(\alpha\mathbf{u}+\beta\mathbf{w})^\top\mathbf{v_1}}{\mathbf{v}_1^\top\mathbf{v}_1}\mathbf{v}_1 \\ &= \alpha\frac{\mathbf{u}^\top\mathbf{v_1}}{\mathbf{v}_1^\top\mathbf{v}_1}\mathbf{v}_1 + \beta\frac{\mathbf{w}^\top\mathbf{v_1}}{\mathbf{v}_1^\top\mathbf{v}_1}\mathbf{v}_1 \\ &= \alpha\mathbf{P}_{\mathbf{v}_1}(\mathbf{u}) + \beta\mathbf{P}_{\mathbf{v}_1}(\mathbf{w}),\end{aligned}$$

(c) According to the result of (a), we know

$$\lambda = \frac{\mathbf{w}^\top \mathbf{v_1}}{\mathbf{v}_1^\top \mathbf{v}_1} = \frac{\mathbf{v}_1^\top \mathbf{w}}{\mathbf{v}_1^\top \mathbf{v}_1}$$

Therefore, $\mathbf{P}_{\mathbf{v}_1}(\mathbf{w}) = \mathbf{v}_1\lambda = \frac{\mathbf{w}^\top \mathbf{v_1}}{\mathbf{v}_1^\top \mathbf{v}_1} = \frac{\mathbf{v}_1\mathbf{v}_1^\top}{\mathbf{v}_1^\top \mathbf{v}_1}\mathbf{w} = \mathbf{H}_1\mathbf{w}$. It means $\mathbf{H}_1 = \frac{\mathbf{v}_1\mathbf{v}_1^\top}{\mathbf{v}_1^\top \mathbf{v}_1}$.

(d) i. We can use the result of 1.:$\mathbf{P}_\mathbf{V}(\mathbf{w}) = (\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}^\top\mathbf{w} = \mathbf{H}\mathbf{w}$. Therefore, we can derive $\mathbf{H} = (\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}^\top$.

ii.

$$\mathbf{V}^\top\mathbf{V} = \begin{pmatrix} \mathbf{v_1}^\top \\ \mathbf{v_2}^\top \\ \vdots \\ \mathbf{v_d}^\top \end{pmatrix} \begin{pmatrix} \mathbf{v_1} & \mathbf{v_2} & \cdots & \mathbf{v_d} \end{pmatrix} = \begin{pmatrix} \mathbf{v_1}^\top\mathbf{v_1} & 0 & 0 & \cdots \\ 0 & \mathbf{v_2}^\top\mathbf{v_2} & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots \end{pmatrix}$$

It is a diagonal matrix. We can derive its inverse matrix easily:

$$(\mathbf{V}^\top\mathbf{V})^{-1} = \begin{pmatrix} \frac{1}{\mathbf{v_1}^\top\mathbf{v_1}} & 0 & 0 & \cdots \\ 0 & \frac{1}{\mathbf{v_2}^\top\mathbf{v_2}} & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots \end{pmatrix}$$

$$\begin{aligned} \mathbf{H} &= (\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}^\top \\ &= \begin{pmatrix} \frac{\mathbf{v_1}^\top}{\mathbf{v_1}^\top\mathbf{v_1}} \\ \frac{\mathbf{v_2}^\top}{\mathbf{v_2}^\top\mathbf{v_2}} \\ \vdots \\ \frac{\mathbf{v_d}^\top}{\mathbf{v_d}^\top\mathbf{v_d}} \end{pmatrix} \end{aligned}$$

3. (a) $\mathbf{P}_\mathbf{A}(\mathbf{x}) = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{x} = \mathbf{x}$. Therefore, the coordinates of $\mathbf{x}$ won't change. It's unique.

(b) We let $\mathbf{x} = (a, b)$. $\mathcal{C}(A) = \{\lambda\mathbf{v}_1 | \lambda \in \mathbb{R}, \mathbf{v}_1 = (1, 1)\}$. According to the definition of projection, we know

$$\begin{aligned} \mathbf{P}_\mathbf{A}(\mathbf{x}) &= \underset{\mathbf{z}\in\mathbb{R}^n}{\mathbf{argmin}}\{\|\mathbf{w} - \mathbf{z}\|_2 : \mathbf{z} = \lambda\mathbf{v_1}\} \\ &= \frac{\mathbf{v}_1\mathbf{v}_1^\top}{\mathbf{v}_1^\top\mathbf{v}_1}\mathbf{x} \\ &= (\frac{a+b}{2}, \frac{a+b}{2}) \end{aligned}$$

It is unique.

■

Exercise 2: Projection to a Matrix Space

Let $\mathbb{R}^{n\times n}$ be the linear space of $n \times n$ matrices. The inner product in this space is defined as

$$\langle A, B\rangle = \mathrm{tr}(A^T B).$$

1. Show that the set of diagonal matrices in $\mathbb{R}^{n\times n}$ forms a linear space. Besides, please find the the projection of any matrix onto the space of diagonal matrices.
2. Prove that the set of symmetric matrices, denoted S^n, in $\mathbb{R}^{n\times n}$ forms a linear space. Also, determine the dimension of this linear space.
3. Show that the inner product of any symmetric matrix and skew-symmetric matrix is zero. Moreover, prove that any matrix can be decomposed as the sum of a symmetric matrix and a skew-symmetric matrix.
4. Find the projection of any matrix onto the space of symmetric matrices.

Solution:

1. We denote the set of diagonal matrices in $\mathbb{R}^{n\times n}$ as $\mathcal{V}$.
   $\forall \mathbf{U}, \mathbf{V} \in \mathcal{V}$, $\mathbf{U} + \mathbf{V} = \mathbf{V} + \mathbf{U}$(commutative), and $(\mathbf{V} + \mathbf{U}) + \mathbf{W} = \mathbf{V} + (\mathbf{U} + \mathbf{W})$(associative).
   $\exists \mathbf{0} \in \mathcal{V}$, $\forall \mathbf{U} \in \mathcal{V}$, $\mathbf{0} + \mathbf{U} = \mathbf{U} + \mathbf{0} = \mathbf{U}$(zero vector).
   $\forall \mathbf{U} \in \mathcal{V}, \exists -\mathbf{U} \in \mathcal{V}$, such that $\mathbf{U} + (-\mathbf{U}) = (-\mathbf{U}) + \mathbf{U} = 0$(additive inverse).
   $\forall \mathbf{U} \in \mathcal{V}, \forall a, b \in \mathbb{R}, (ab)\mathbf{U} = a(b\mathbf{U})$(compatible).
   Identity matrix $\mathbf{I}$ is also a diagonal matrix, $\mathbf{I} \in \mathcal{V}$, $\forall \mathbf{U} \in \mathcal{V}, \mathbf{I}\mathbf{U} = \mathbf{U}$(multiplicative identity).
   $\forall a \in \mathbb{R}, \forall \mathbf{U}, \mathbf{V} \in \mathcal{V}, a(\mathbf{U} + \mathbf{V}) = a\mathbf{U} + a\mathbf{V}$(distributive).
   $\forall a, b \in \mathbf{R}, \forall \mathbf{U} \in \mathcal{V}, (a + b)\mathbf{U} = a\mathbf{U} + b\mathbf{V}$. Therefore we can say $\mathcal{V}$ is a linear space.
   According to the definition of projection, we can write:

$$\begin{aligned}\mathbf{P}_{\mathcal{V}}(\mathbf{x}) &= \underset{\mathbf{Z}\in\mathcal{V}}{\mathbf{argmin}} \, \|\mathbf{X} - \mathbf{Z}\|_2 \\ &= \underset{\lambda\in\mathbb{R}}{\mathbf{argmin}} \, \|\mathbf{X} - \lambda\mathbf{I}\|_2\end{aligned}$$

$$\frac{\partial}{\partial\lambda}\|\mathbf{X} - \lambda\mathbf{I}\|_2^2 = \frac{\partial}{\partial\lambda}(\mathbf{X}^\top\mathbf{X} - \lambda\mathbf{X} - \lambda\mathbf{X}^\top + \lambda^2\mathbf{I}) = -\mathbf{X} - \mathbf{X}^\top + 2\lambda\mathbf{I} = 0$$

   We can get $\lambda = \frac{\mathbf{X}+\mathbf{X}^\top}{2}$. It means $\mathbf{P}_{\mathcal{V}}(\mathbf{x}) = \frac{\mathbf{X}+\mathbf{X}^\top}{2}\mathbf{I}$.

2. $\forall \mathbf{U}, \mathbf{V} \in S^n$, $\mathbf{U} + \mathbf{V} = \mathbf{V} + \mathbf{U}$(commutative), and $(\mathbf{V} + \mathbf{U}) + \mathbf{W} = \mathbf{V} + (\mathbf{U} + \mathbf{W})$(associative).
   $\mathbf{0}$ is a symmetric matrices. Thus $\exists \mathbf{0} \in S^n$,$\forall \mathbf{U} \in S^n$, $\mathbf{0} + \mathbf{U} = \mathbf{U} + \mathbf{0}$(zero vector).
   If $\mathbf{U}$ is a symmetric matrix, $-\mathbf{U}$ must be a symmetric matrix. Thus $\forall \mathbf{U} \in S^n$, $\exists -\mathbf{U} \in S^n$, $\mathbf{U} + (-\mathbf{U}) = (-\mathbf{U}) + \mathbf{U} = \mathbf{0}$(additive inverse).
   $\forall \mathbf{U} \in S^n, \forall a, b \in \mathbb{R}, (ab)\mathbf{U} = a(b\mathbf{U})$(compatible).
   Identity matrix $\mathbf{I}$ is also a symmetric matrix, $\mathbf{I} \in S^n$, $\forall \mathbf{U} \in S^n, \mathbf{I}\mathbf{U} = \mathbf{U}$(multiplicative identity).
   $\forall a \in \mathbb{R}, \forall \mathbf{U}, \mathbf{V} \in S^n, a(\mathbf{U} + \mathbf{V}) = a\mathbf{U} + a\mathbf{V}$(distributive).

$\forall a, b \in \mathbf{R}, \forall \mathbf{U} \in S^n, (a+b)\mathbf{U} = a\mathbf{U} + b\mathbf{V}$.
Therefore we can say $\mathcal{V}$ is a linear space.
We use $\mathbf{E}_{ij}(1 \leq i \leq j \leq n)$ to represent those matrices in which the elements in the i-th row, j-th column and j-th row and i-column are 1, and the remaining elements are all 0. It is obvious that they are linear independent and any symmetric matrices in S^n can be written as the linear combination of $\mathbf{E}_{ij}$. It means it is a group of basis. $n + \frac{(n-2)n}{2} = \frac{n^2}{2}$. In other words, the dimension of this linear space is $\frac{n^2}{2}$.

3. Suppose that A is a symmetric matrix and B is a skew-symmetric matrix.

$$\begin{aligned} < A, B > &= \mathrm{tr}(A^\top B) = \mathrm{tr}(AB) \\ < B, A > &= \mathrm{tr}(B^\top A) = \mathrm{tr}(-BA) \\ &= -\mathrm{tr}(AB) \\ &= - < A, B > \end{aligned}$$

According to the definition of trace, we know

$$\mathrm{tr}(A^\top B) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ij} = \mathrm{tr}(B^\top A)$$

It means $< A, B > = < B, A > = - < A, B >$. In other words $< A, B > = 0$.
$\forall C \in \mathbb{R}^{n\times n}$, we let $A = \frac{C+C^\top}{2}$ and $B = \frac{C-C^\top}{2}$. Because $A^\top = \frac{C^\top+C}{2} = A$ and $B^\top = \frac{C^\top-C}{2} = -B$, we know A is a symmetric matrix and B is a skew-symmetric matrix. Notice that $C = A + B$.

4. We know any matrices can be decomposed as the sum of a symmetric matrix and a skew-symmetric matrix. And the inner product of any symmetric matrix and skew-symmetric matrix is zero. Thus, a symmetric matrix and a skew-symmetric matrix can be a group of basis of $\mathbb{R}^{n\times n}$ linear space.
According to the definition of projection, we know the projection of a matrix onto the space of symmetric matrices is its symmetrical components: $\mathbf{P}_{S^n}(X) = \frac{X+X^\top}{2}$.

■

Exercise 3: Projection to a Function Space

1. Suppose X and Y are both random variables defined in the same sample space Ω with finite second-order moment, i.e. $\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$.

   (a) Let $L^2(\Omega) = \{Z : \Omega \to \mathbb{R} \mid \mathbb{E}[Z^2] < \infty\}$ be the set of random variables with finite second-order moment. Please show that $L^2(\Omega)$ is a linear space, and $\langle X, Y\rangle := \mathbb{E}[XY]$ defines an inner product in $L^2(\Omega)$. Then find the projection of Y on the subspace of $L^2(\Omega)$ consisting of all constant variables.

   (b) Please find a real constant $\hat{c}$, such that

   $$\hat{c} = \mathop{\mathbf{argmin}}_{c\in\mathbb{R}} \mathbb{E}[(Y - c)^2].$$

   [Hint: you can solve it by completing the square.]

   (c) Please find the necessary and sufficient condition where $\min_{c\in\mathbb{R}} \mathbb{E}[(Y - c)^2] = \mathbb{E}[Y^2]$. Then give it a geometric interpretation using inner product and projection.

2. Suppose X and Y are both random variables defined in the same sample space Ω and all the expectations exist in this problem. Consider the problem

   $$\min_{f:\mathbb{R}\to\mathbb{R}} \mathbb{E}[(f(X) - Y)^2].$$

   (a) Please solve the above problem by completing the square.

   (b) We let $\mathcal{C}(X)$ denote the subspace $\{f(X) \mid f(\cdot) : \mathbb{R} \to \mathbb{R}, \mathbb{E}[f(X)^2] < \infty\}$ of $L^2(\Omega)$. Please show that the solution of the above problem is the projection of Y on $\mathcal{C}(X)$.

   (c) Please show that question 1 is a special case of question 2. Please give a geometric interpretation of conditional expectation.

Solution: 1. (a) $\forall X, Y \in L^2(\Omega) \forall a, b \in \mathbb{R}, \mathbb{E}[(aX + bY)^2] = a^2\mathbb{E}[X^2] + b^2\mathbb{E}[Y^2] + 2ab\mathbb{E}[XY]$. We know that $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$,According to Cauchy-Schwarz inequality, we know $\mathbb{E}[XY] \le \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]} < \infty$. Thus, $\mathbb{E}[(X + Y)^2] < \infty$. In other words, $aX + bY \in L^2(\Omega)$.
For random variables, other properties of linear spaces(i.e. commutative, associative, zero vector...) are obvious. $L^2(\Omega)$ is a linear space.
$< X, X >= \mathbb{E}[X^2] \ge 0, \forall X \in L^2(\Omega)$ and $< X, X >= 0$ if and only if $X = 0$(nonnegative and definite).
$\forall X, Y \in L^2(\Omega), < X, Y >= \mathbb{E}[XY] = \mathbb{E}[YX] =< Y, X >$(symmetric).
$\forall X, Y, Z \in L^2(\Omega), < aX+bY, Z >= \mathbb{E}[(aX+bY)Z] = \mathbb{E}[aXZ+bYZ] = a\mathbb{E}[XZ]+b\mathbb{E}[YZ] = a < X, Z > +b < Y, Z >$.The same can be said, $< X, aY + bZ >= a < X, Y > +b < X, Z >$(bilinear). In summary, $\langle X, Y\rangle := \mathbb{E}[XY]$ defines an inner product.
$\mathbf{P}_{L^2(\Omega)}(Y) = \mathbb{E}[Y]$, because $\forall c \in \mathbb{R}, < c, Y - \mathbb{E}[Y] >= \mathbb{E}[cY - c\mathbb{E}[Y]] = c\mathbb{E}[Y] - c\mathbb{E}[Y] = 0$. It means all elements in $\mathbb{R}$ are orthogonal to $Y - \mathbb{E}[Y]$. Thus $\mathbb{E}[Y]$ is the projection of Y on the subspace of $L^2(\Omega)$ consisting of all

constant variables.

(b) $\mathbb{E}[(Y-c)^2] = \mathbb{E}[Y^2] - 2c\mathbb{E}[Y] + c^2$.We differentiate this expression with respect to c:

$$\frac{\partial}{\partial c}(\mathbb{E}[Y^2] - 2c\mathbb{E}[Y] + c^2) = -2\mathbb{E}[Y] + 2c = 0$$

From above, we know it can get its minimum value when $c = \mathbb{E}[Y]$. In other words, $\hat{c} = \mathbb{E}[Y]$.

(c) According to the result of (b), we know it can get it minimum value when $c = \mathbb{E}[Y]$. In other words, $min_{c\in\mathbb{R}}\mathbb{E}[(Y-c)^2] = \mathbb{E}[(Y-\mathbb{E}[Y])^2] = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2$. It means that $\min_{c\in\mathbb{R}} \mathbb{E}[(Y-c)^2] = \mathbb{E}[Y^2]$ if and only if $\mathbb{E}[Y] = 0$.
From a geometric interpretation point of view, $\mathbb{E}[Y] = 0$ means the projection of Y onto the space $\mathbb{R}$ is zero, or we can say vector Y is orthogonal to flat $\mathbb{R}$.

2. (a)

$$\begin{aligned}(f(X)-Y)^2 &= (f(X) - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - Y)^2 \\ &= (f(X) - \mathbb{E}[Y|X])^2 + 2(f(X) - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - Y) \\ &+ (\mathbb{E}[Y|X] - Y)^2\end{aligned}$$

$$\begin{aligned}\mathbb{E}[(f(x)-y)^2] = &\int (f(x) - \mathbb{E}[y|x])^2 p(x)dx \\ &+ 2\int (f(x) - \mathbb{E}[y|x])[\int (\mathbb{E}[y|x] - y)p(x,y)dy]dx \\ &+ \iint (\mathbb{E}[y|x] - y)^2 p(x,y)dxdy\end{aligned}$$

For the second term, we notice that

$$\begin{aligned}\int (\mathbb{E}[y|x] - y)p(x,y)dy &= \mathbb{E}[y|x]\int p(x,y)dy - \int yp(x,y)dy \\ &= \mathbb{E}[y|x]p(x) - p(x)\int yp(y|x)dy \\ &= \mathbb{E}[y|x]p(x) - \mathbb{E}[y|x]p(x) \\ &= 0\end{aligned}$$

Therefore, $J[f] = \mathbb{E}[(f(x)-y)^2] = \int (f(x) - \mathbb{E}[y|x])^2 p(x)dx + \iint (\mathbb{E}[y|x] - y)^2 p(x,y)dxdy$. We can see that the second term of the above equation are constant for f and the first term is non-negative. To obtain the minimum, we have to let the first term equals zero. It means $f^*(X) = \mathbb{E}[Y|X]$.$\min_{f:\mathbb{R}\to\mathbb{R}} \mathbb{E}[(f(X)-Y)^2] = \iint (\mathbb{E}[Y|X] - Y)^2 p(X,Y)dXdY$.

(b)

$$\begin{aligned}\forall f(X) \in \mathcal{C}(X), < Y - f^*(X), f(X) > &= \mathbb{E}[f(X)Y - f(X)\mathbb{E}[Y|X]] \\ &= \mathbb{E}[f(X)Y] - \mathbb{E}[f(X)\mathbb{E}[Y|X]]\end{aligned}$$

$$\mathbb{E}[f(X)Y] = \mathbb{E}[\mathbb{E}[f(X)Y|X]] = \mathbb{E}[f(X)\mathbb{E}[Y|X]]$$

Thus, $< Y - f^*(X), f(X) >= 0$, it means $f^*(X) = \mathbb{E}(Y|X)$ is the projection of Y on $\mathcal{C}(X)$.

(c) When we limit the function f to map a random variable to a constant($f(X) = c$ and c is a constant), $\mathcal{C}(X) = \mathbb{R}$. At this time, $\mathbb{E}[Y|X] = \mathbb{E}[Y]$ because X is a constant. Thus question 1 is a special case of question 2.
The conditional expectation is the projection of Y onto the space $\mathcal{C}(X)$.

■

Exercise 4: Regularized least squares

Suppose that $\mathbf{X} \in \mathbb{R}^{n\times d}$.

1. Please show that $\mathbf{X}^\top\mathbf{X}$ is always positive semi-definite. Moreover, $\mathbf{X}^\top\mathbf{X}$ is positive definite if and only if $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d$ are linearly independent.

2. Please show that $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ is always invertible, where $\lambda > 0$ and $\mathbf{I} \in \mathbb{R}^{d\times d}$ is an identity matrix.

3. (Optional) Consider the regularized least squares linear regression and denote

$$\mathbf{w}^*(\lambda) = \mathop{\mathbf{argmin}}_{\mathbf{w}} L(\mathbf{w}) + \lambda\Omega(\mathbf{w}),$$

where $L(\mathbf{w}) = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$ and $\Omega(\mathbf{w}) = \|\mathbf{w}\|_2^2$. For regular parameters $0 < \lambda_1 < \lambda_2$, please show that $L(\mathbf{w}^*(\lambda_1)) < L(\mathbf{w}^*(\lambda_2))$ and $\Omega(\mathbf{w}^*(\lambda_1)) > \Omega(\mathbf{w}^*(\lambda_2))$. Explain intuitively why this holds.

Solution:

1. $\forall\mathbf{y} \in \mathbb{R}^d, \mathbf{y} \neq 0, \mathbf{y}^\top(\mathbf{X}^\top\mathbf{X})\mathbf{y} = (\mathbf{X}\mathbf{y})^\top\mathbf{X}\mathbf{y}$. Let $\mathbf{w} = \mathbf{X}\mathbf{y}$, we have $\mathbf{y}^\top(\mathbf{X}^\top\mathbf{X})\mathbf{y} = \|\mathbf{w}\|_2^2 \geq 0$. Thus, $\mathbf{X}^\top\mathbf{X}$ is always positive semi-definite.
   $\mathbf{y}^\top(\mathbf{X}^\top\mathbf{X})\mathbf{y} = 0 \Leftrightarrow \mathbf{w} = 0 \Leftrightarrow \exists\mathbf{y} \neq 0, \mathbf{X}\mathbf{y} = y_1\mathbf{x}_1 + y_2\mathbf{x}_2 + \cdots + y_d\mathbf{x}_d = 0 \Leftrightarrow$ $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d$ are linearly dependent. In other words, $\mathbf{X}^\top\mathbf{X}$ is positive$\Leftrightarrow\forall\mathbf{y} \in \mathbb{R}^n$, $\mathbf{y}^\top(\mathbf{X}^\top\mathbf{X})\mathbf{y} > 0 \Leftrightarrow \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d$ are linearly independent.

2. $\forall\mathbf{y} \in \mathbb{R}^d, \mathbf{y} \neq 0, \mathbf{y}^\top(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\mathbf{y} = \|\mathbf{X}\mathbf{y}\|_2^2 + \lambda\|\mathbf{y}\|_2^2$. Because $\|\mathbf{X}\mathbf{y}\|_2^2 \geq 0$, $\lambda > 0$ and $\|\mathbf{y}\|_2^2 > 0$, $\mathbf{y}^\top(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\mathbf{y} > 0$. In other words, $\mathbf{y}^\top(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\mathbf{y}$ is a positive definite matrix, it must be invertible($\det(\mathbf{y}^\top(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\mathbf{y}) > 0$).

■

Exercise 5: Bias-Variance Trade-off (Programming Exercise)

We provide you with $L = 100$ data sets, each having $N = 25$ points:

$$\mathcal{D}^{(l)} = \{(x_n, y_n^{(l)})\}_{n=1}^{N}, \quad l = 1, 2, \cdots, L,$$

where x_n are uniformly taken from $[-1, 1]$, and all points $(x_n, y_n^{(l)})$ are independently from the sinusoidal curve $h(x) = \sin(\pi x)$ with an additional disturbance.

1. For each data set $\mathcal{D}^{(l)}$, consider fitting a model with 24 Gaussian basis functions

$$\phi_j(x) = e^{-(x-\mu_j)^2}, \quad \mu_j = 0.2 \cdot (j - 12.5), \quad j = 1, \cdots 24$$

by minimizing the regularized error function

$$L^{(l)}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(y_n^{(l)} - \mathbf{w}^\top \boldsymbol{\phi}(x_n))^2 + \frac{\lambda}{2}\mathbf{w}^\top \mathbf{w},$$

where $\mathbf{w} \in \mathbb{R}^{25}$ is the parameter, $\boldsymbol{\phi}(x) = (1, \phi_1(x), \cdots, \phi_{24}(x))^\top$ and λ is the regular coefficient. What's the closed form of the parameter estimator $\hat{\mathbf{w}}^{(l)}$ for the data set $\mathcal{D}^{(l)}$?

2. For $\log_{10}\lambda = -10, -5, -1, 1$, plot the prediction functions $y^{(l)}(x) = f_{\mathcal{D}^{(l)}}(x)$ on $[-1, 1]$ respectively. For clarity, show only the first 25 fits in the figure for each λ.

3. For $\log_{10}\lambda \in [-3, 1]$, calculate the followings:

$$\bar{y}(x) = \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(x)] = \frac{1}{L}\sum_{l=1}^{L} y^{(l)}(x)$$

$$(\text{bias})^2 = \mathbb{E}_X[(\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(X)] - h(X))^2] = \frac{1}{N}\sum_{n=1}^{N}(\bar{y}(x_n) - h(x_n))^2$$

$$\text{variance} = \mathbb{E}_X[\mathbb{E}_{\mathcal{D}}[(f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})])^2]] = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{l=1}^{L}(y^{(l)}(x_n) - \bar{y}(x_n))^2$$

Plot the three quantities, $(\text{bias})^2$, variance and $(\text{bias})^2 +$ variance in one figure, as the functions of $\log_{10}\lambda$. (**Hint:** see [**?**] for an example.)

Solution: 1. Let $\Phi(\mathbf{x}^{(l)}) = \left(\phi(x_1)^{(l)}, \phi(x_2)^{(l)}, \cdots, \phi(x_{25}^{(l)})\right)^\top \in \mathbb{R}^{25\times 25}$ and $\mathbf{y}^{(l)} \in \mathbb{R}^{25}$. We can rewrite the loss function as

$$L^{(l)}(\mathbf{w}) = \frac{1}{2}\|\mathbf{y}^{(l)} - \Phi(\mathbf{x}^{(l)})\mathbf{w}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

$$\frac{\partial L^{(l)}(\mathbf{w})}{\partial \mathbf{w}} = -\Phi^\top(\mathbf{x}^{(l)}\mathbf{w}) + \lambda\mathbf{w} = 0$$

It can get its minimum value at $\hat{\mathbf{w}}^{(l)} = (\Phi^\top(\mathbf{x}^{(l)})\Phi(\mathbf{x}^{(\mathbf{l})}) + \lambda\mathbf{I})^{-1}\Phi^\top(\mathbf{x}^{(l)})\mathbf{y}^{(l)}$.

2. This is python code.

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import os

data_root = "./HW2_DataSet&Ref/Ex5 data"
number = np.linspace(1, 100, 100, dtype=int)
file_name = []
for i in number:
    file_name.append(os.path.join(data_root, "data_{}".format(i)))
datasets = []
columns = ['x', 'y']
for file in file_name:
    content = pd.DataFrame(np.loadtxt(file), columns=columns)
    datasets.append(content)
x = datasets[0]['x'].to_numpy()

def GaussianF(X_scalar):
    J = np.linspace(1, 24, 24, dtype=int)
    miu = 0.2 * (J-12.5)
    return np.insert(np.exp(-(X_scalar-miu)**2), 0, 1)

def GetPhi(X_vec):
    Phi = []
    for x in X_vec:
        phi = GaussianF(x)
        Phi.append(phi)
    return np.array(Phi)

def predict_y(Phi, lamb, Y):
    I = np.eye(25)
    temp = Phi.transpose()@Phi + lamb * I
    temp_inv = np.linalg.inv(temp)
    w = temp_inv @ Phi.transpose() @ (Y.transpose())
    return (Phi @ w.reshape(25,1))

Phi = GetPhi(x)
Lamb = [1e-10, 1e-5, 1e-1, 10]

fig, axes = plt.subplots(2, 2, figsize=(20, 20))
cor = [(0,0), (0,1), (1,0), (1, 1)]
for j, lamb in enumerate(Lamb):
    y_hat = []
    for i in range(25):
        y_hat.append(predict_y(Phi, lamb, datasets[i]['y'].to_numpy()).flatten())
        color = plt.cm.viridis(i / 25)
        axes[cor[j][0], cor[j][1]].plot(x, y_hat[i], label=f'Curve {i}', color=color)

for i, ax in enumerate(axes.flat):
    row, col = divmod(i, 2)
    ax.set_title(f'Subplot {row+1}-{col+1}  lamda = {Lamb[i]}')
```

```
    ax.legend()
plt.tight_layout()

# show image
plt.show()
```

Ex5-2.py

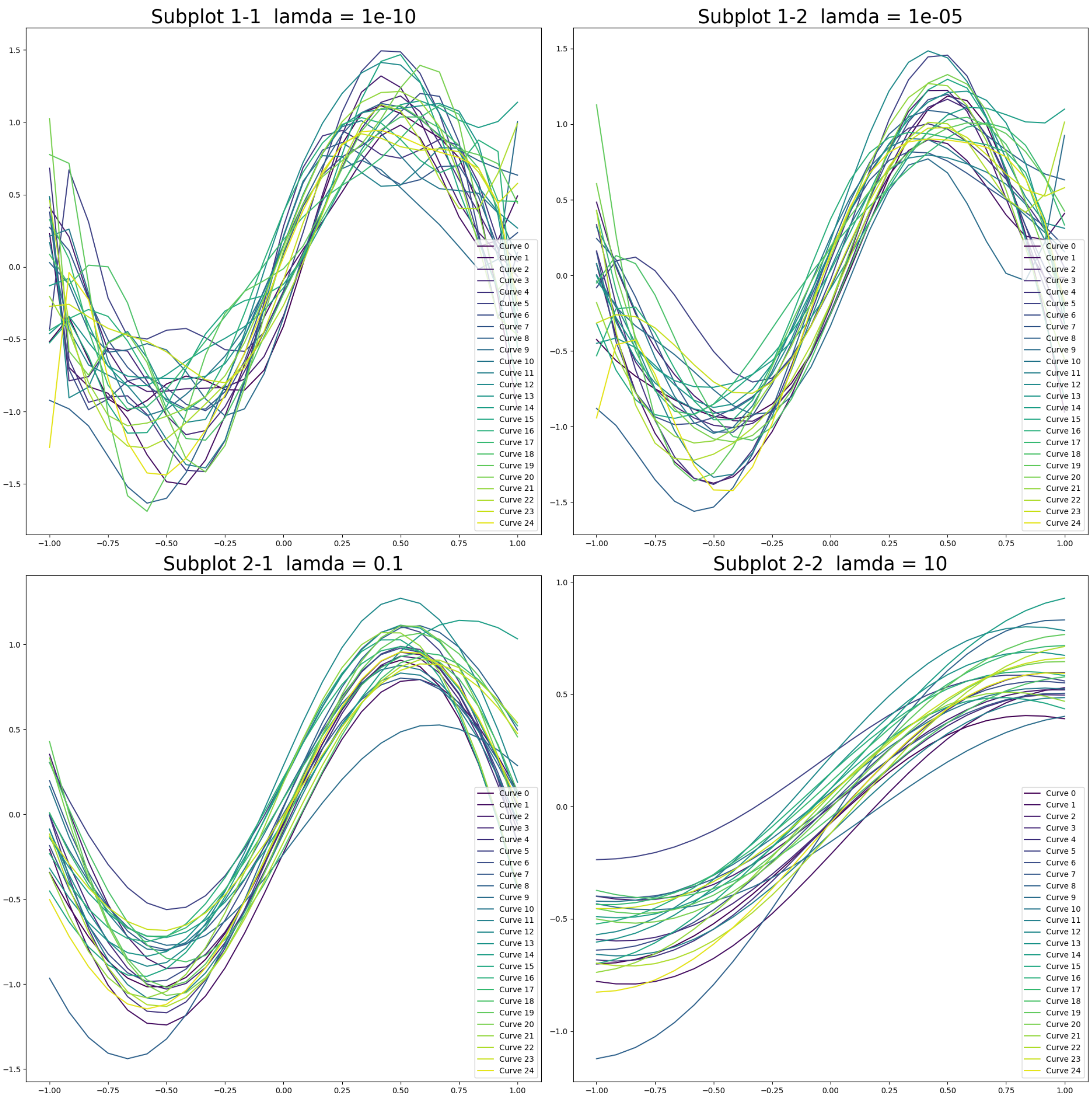

Figure 1: Ex5-2 Different λ

3. This is the python code.

```
def get_y_bar(L, X, datasets, lamb, Phi):
    y_hat = []
```

```python
    for i in range(L):
        y_hat.append(predict_y(Phi, lamb, datasets[i]['y'].to_numpy()).
    flatten())
    y_hat = np.array(y_hat)
    return y_hat.mean(axis=0)
def get_h_x(X):
    return np.sin(3.1415926535 * X)
def get_bias_square(X, datasets, lamb, Phi):
    L = len(datasets)
    y_bar = get_y_bar(L, X, datasets, lamb, Phi)
    h_x = get_h_x(X)
    return (1/25)*np.linalg.norm(y_bar-h_x)**2
def get_variance(X, datasets, lamb, Phi):
    L = len(datasets)
    N = 25
    y_bar = get_y_bar(L, X, datasets, lamb, Phi)
    y_hat = []
    for i in range(L):
        y_hat.append(predict_y(Phi, lamb, datasets[i]['y'].to_numpy()).
    flatten())
    y_hat = np.array(y_hat)
    temp = (y_hat - y_bar)
    ED = (1/L)*np.linalg.norm(temp, ord=2, axis=0)**2
    return ED.mean()

fig, axes = plt.subplots(1, 1)
axes.plot(log_Lamb, bias_square, label='bias_square', color='b')
axes.plot(log_Lamb, variance, label='variance', color='g')
axes.plot(log_Lamb, bias_square_plus_var, label='bias_square_plus_var',
    color='r')
axes.legend()
plt.show()
```

Ex5-3.py

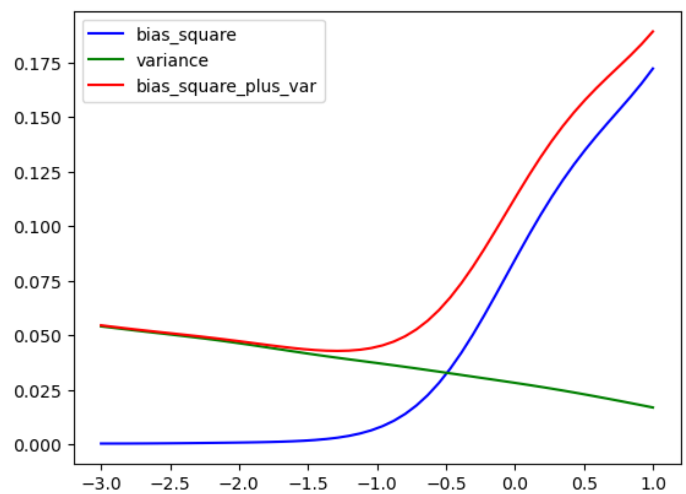

Figure 2: Ex5-3

■

Exercise 6: Linear Regression (Programming Exercise)

Consider a data set $\{(x_i, y_i)\}_{i=1}^n$, where $x_i, y_i \in \mathbb{R}$.

1. If we want to fit the data by a linear model

$$y = w_0 + w_1 x, \tag{1}$$

   please find $\hat{w}_0$ and $\hat{w}_1$ by the least squares approach (you need to find expressions of $\hat{w}_0$ and $\hat{w}_1$ by $\{(x_i, y_i)\}_{i=1}^n$, respectively).

2. We provide you a data set $\{(x_i, y_i)\}_{i=1}^{30}$. Consider the model in (1) and the one as follows:

$$y = w_0 + w_1 x + w_2 x^2. \tag{2}$$

   Which model do you think fits better the data? Please detail your approach first and then implement it by your favorite programming language. The required output includes

   (a) your detailed approach step by step;

   (b) your code with detailed comments according to your planned approach;

   (c) a plot showing the data and the fitting models;

   (d) the model you finally choose [$\hat{w}_0$ and $\hat{w}_1$ if you choose the model in (1), or $\hat{w}_0$, $\hat{w}_1$, and $\hat{w}_2$ if you choose the model in (2)].

Solution: 1. $\bar{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix}^\top$, $\mathbf{w} = \begin{pmatrix} w_0 & w_1 \end{pmatrix}^\top$ and $\mathbf{y} = \begin{pmatrix} y_1 & y_2 & \cdots & y_n \end{pmatrix}^\top$.

This linear model can be written as

$$\mathbf{y} = \bar{X}\mathbf{w}$$

According to the theory of least squares, we know the solution of least squares is $\hat{\mathbf{w}} = (\bar{X}^\top \bar{X})^{-1} \bar{X}^\top \mathbf{y}$.

$$(\bar{X}^\top \bar{X})^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}$$

$$\bar{X}^\top \mathbf{y} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$$

We can derive that

$$\hat{\mathbf{w}} = \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i \\ n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i \end{pmatrix}$$

Thus we can get:

$$\hat{w_0} = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$
$$\hat{w_1} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

2. We have detailed the model in (1), next we will solve the model in (2). $\bar{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ x_1^2 & x_2^2 & \cdots & x_n^2 \end{pmatrix}^\top$, $\mathbf{w} = \begin{pmatrix} w_0 & w_1 & w_2 \end{pmatrix}^\top$ and $\mathbf{y} = \begin{pmatrix} y_1 & y_2 & \cdots & y_n \end{pmatrix}^\top$. This linear model can be written as

$$\mathbf{y} = \bar{X}\mathbf{w}$$

According to the theory of least squares, we know the solution of least squares is $\hat{\mathbf{w}} = (\bar{X}^\top \bar{X})^{-1} \bar{X}^\top \mathbf{y}$.
In order to compare the performance of the two models on this data set, one can solve the parameter $\mathbf{w}$ of the two models and plot their result. Next one can compute the $R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$. After comparing the curves fitted by the two models with the actual curves and the R^2 of the two models, one can choose the better model.

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import os

# Load dataset
data_path = "./HW2_DataSet&Ref/Ex6 data.xls"
data = np.loadtxt(data_path)
x = data[:, 0]
y = data[:, 1]
X_1 = np.stack([np.ones_like(x), x]).T
X_2 = np.stack([np.ones_like(x), x, x**2]).T

# Compute parameters of model1 and model2
w_1 = np.linalg.inv(X_1.T @ X_1) @ X_1.T @ y.T
w_2 = np.linalg.inv(X_2.T @ X_2) @ X_2.T @ y.T
print(f"w1={w_1}")
print(f"w2={w_2}")

# In order to plot smooth curve, we enter 100 data points from -1 to 1
x_range = np.linspace(-1, 1, 100)
X_new_1 = np.stack([np.ones_like(x_range), x_range]).T
X_new_2 = np.stack([np.ones_like(x_range), x_range, x_range**2]).T
y_1 = X_new_1 @ w_1
y_2 = X_new_2 @ w_2

fig, axes = plt.subplots(1, 1)
axes.scatter(x, y, label='Ground Truth', color='C1')
axes.plot(x_range, y_1, label='Model 1', color='C2')
axes.plot(x_range, y_2, label='Model 2', color='C3')
axes.legend()
```

```
# show image
plt.show()

# Compute the coefficient of determination of these 2 models
y_bar = y.mean()
y_hat_1 = X_1 @ w_1
y_hat_2 = X_2 @ w_2
sst = ((y - y_bar)**2).sum()
sse_1 = ((y_hat_1 - y)**2).sum()
sse_2 = ((y_hat_2 - y)**2).sum()

R_sqr_1 = 1. - (sse_1 / sst)
R_sqr_2 = 1. - (sse_2 / sst)
print(f'R_sqr_1 = {R_sqr_1}')
print(f'R_sqr_2 = {R_sqr_2}')
```

Ex6-2.py

We can obtain a plot showing the data and the fitting models.

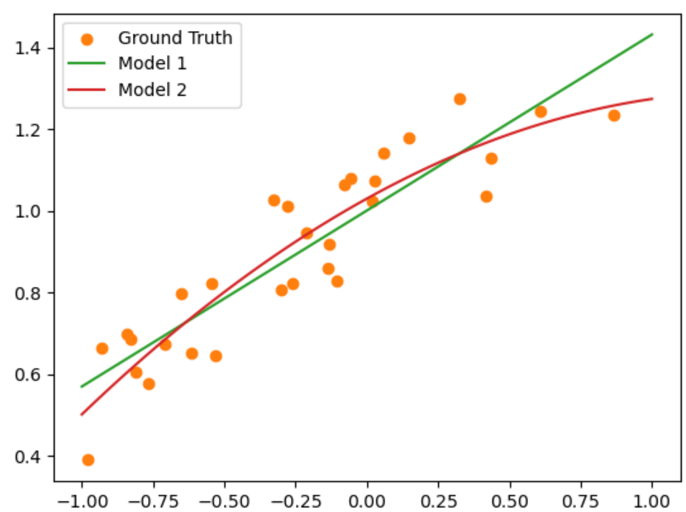

Figure 3: Ex6-2

Also, we can obtain the coefficient of determination of the 2 models:

$$R_1^2 = 0.814894689163735$$
$$R_2^2 = 0.840921711424471$$
$$R_1^2 < R_2^2$$

Therefore, in terms of the 30 data given, Model 2 is better than Model 1. I think model 2 (quadratic model) fits better the given data.

$\hat{w}_0 = 1.02956837, \hat{w}_1 = 0.38614333, \hat{w}_2 = -0.14215111.$

■

Exercise 7: (Optional) Positive Semi-definite Matrices and the Polyhedron

Please show that $\mathbb{S}_+^n$ is not a polyhedron.

Solution: Now, suppose for the purpose of contradiction that $\mathbb{S}_+^n$ is polyhedron. According to the definition of polyhedron, polyhedron is the intersection of a finite number of halfspaces and hyperplanes.
$\mathbf{z} \in \mathbb{R}^n, \mathbf{z} \neq 0$, $\mathbf{z}^\top \mathbf{A}\mathbf{z}$ is a linear function of $\mathbf{A}$. Thus $\{\mathbf{A} \in \mathbb{S}_+^n | \mathbf{z}^\top \mathbf{A}\mathbf{z} \geq 0\}$ is a halfspace of $\mathbb{S}_+^n$. Accroding to the definition of polyhedron, we can take

$$\mathbb{S}_+^n = \bigcap_{k=1}^{N} \{\mathbf{A} \in \mathbb{S}_+^n | \mathbf{x}_k^\top \mathbf{A}\mathbf{x}_k \geq 0\}$$

We let $\mathbf{y}_k = \frac{\mathbf{x}_k}{\|\mathbf{x}_k\|_2}$. $\mathbf{y}_k$ is unit vector. And we can take

$$\mathbb{S}_+^n = \bigcap_{k=1}^{N} \{\mathbf{A} \in \mathbb{S}_+^n | \mathbf{y}_k^\top \mathbf{A}\mathbf{y}_k \geq 0\}$$

$\forall \mathbf{y} \in \mathbb{R}^n$, $\mathbf{y}$ is a unit vector and not a multiple of $\mathbf{x}_k$ for $k = 1, 2, \cdots, N$.

$$\alpha = \max_{k=1,2,\cdots,N} (\mathbf{y}_k^\top \mathbf{y})^2 < 1$$

Note that this maximum necessarily exists since it is the maximum of a finite set. Let $\mathbf{X} = \alpha \mathbf{I} - \mathbf{y}\mathbf{y}^\top$.

$$\mathbf{y}_k^\top \mathbf{X}\mathbf{y}_k = \alpha \mathbf{y}_k^\top \mathbf{y}_k - \mathbf{y}_k^\top \mathbf{y}\mathbf{y}^\top \mathbf{y}_k = \alpha - (\mathbf{y}_k^\top \mathbf{y})^2 \geq 0$$

Therefore, $\mathbf{X} \in \mathbb{S}_+^n$. However,

$$\mathbf{y}^\top \mathbf{X}\mathbf{y} = \alpha - 1 < 0$$

It means $\mathbf{X}$ is not a positive semi-definite matrix, which contradicts our premise. So, $\mathbb{S}_+^n$ is not a polyhedron. ■

You will attain **one extra point** of bonus in your **final rating** if you work out this problem.

References

[Gre93] George D. Greenwade. The Comprehensive Tex Archive Network (CTAN). *TUG-Boat*, 14(3):342–351, 1993.