Introduction to Machine Learning

Fall 2024

University of Science and Technology of China

Lecturer: Jie Wang | Homework 2

Posted: Oct. 15, 2024 | Due: Oct. 24, 2024

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 $d \leq n$, 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: Projection

1. ① Strict Convexity:

   Define
   $$f(\mathbf{z}) = \|\mathbf{x}-\mathbf{z}\|_2^2 = \|\mathbf{z}\|_2^2 + 2\langle \mathbf{x}, \mathbf{z}\rangle + \|\mathbf{x}\|_2^2$$
   Then $\nabla^2 f(\mathbf{z}) = 2\mathbf{I_m} > 0 \Longrightarrow f$ is strictly convex.

   The column space $\mathcal{C}(\mathbf{A})$ is convex $\Longrightarrow \forall\ \mathbf{a}, \mathbf{b} \in \mathcal{C}(\mathbf{A}), \forall\ t \in (0,1),\ t\mathbf{a}+(1-t)\mathbf{b} \in \mathcal{C}(\mathbf{A})$.

   Suppose $\mathbf{z_1}, \mathbf{z_2} \in \mathcal{C}(\mathbf{A})$ are both minimizers of f over $\mathcal{C}(\mathbf{A})$. Then according to the convexity of f,
   $$f(t\mathbf{z_1} + (1-t)\mathbf{z_2}) < tf(\mathbf{z_1}) + (1-t)f(\mathbf{z_2}) = f(\mathbf{z_1}) = f(\mathbf{z_2})$$
   contradicting the minimality of $\mathbf{z_1}$ and $\mathbf{z_2}$. Therefore, the minimizer $\mathbf{P_A}(\mathbf{x})$ is unique.

   ② Orthogonality:

   Suppose $\mathbf{z_0}$ is a minimizer of f over $\mathcal{C}(\mathbf{A})$. For any $\mathbf{y} \in \mathcal{C}(\mathbf{A})$ and $t \in \mathbb{R}$, define
   $$\Phi(t) = \|\mathbf{x} - (\mathbf{z_0} + t\mathbf{y})\|_2^2 = \|\mathbf{x}-\mathbf{z_0}\|_2^2 - 2t\langle \mathbf{x}-\mathbf{z_0}, \mathbf{y}\rangle + t^2\|\mathbf{y}\|_2^2$$
   Notice that $\mathbf{z_0} + t\mathbf{y} \in \mathcal{C}(\mathbf{A})$.

   Since $\mathbf{z_0}$ minimizes f over $\mathcal{C}(\mathbf{A})$, $\Phi(t)$ achieves its minimum at $t=0 \Longrightarrow \Phi'(0) = -2\langle \mathbf{x}-\mathbf{z_0}, \mathbf{y}\rangle = 0 \Longrightarrow \mathbf{x}-\mathbf{z_0} \perp \mathcal{C}(\mathbf{A})$, i.e., $\mathbf{x}-\mathbf{z_0} \in \mathcal{C}(\mathbf{A})^\perp$.

   Furthermore, if $\mathbf{x}-\mathbf{z_0} \perp \mathcal{C}(\mathbf{A})$, then for any $\mathbf{y} \in \mathcal{C}(\mathbf{A})$,
   $$\|\mathbf{x}-\mathbf{y}\|_2^2 = \|\mathbf{x}-\mathbf{z_0}+\mathbf{z_0}-\mathbf{y}\|_2^2 = \|\mathbf{x}-\mathbf{z_0}\|_2^2 + \|\mathbf{z_0}-\mathbf{y}\|_2^2 \geq \|\mathbf{x}-\mathbf{z_0}\|_2^2$$
   $\Longrightarrow \mathbf{z_0}$ is a minimizer of f over $\mathcal{C}(\mathbf{A})$.

   If $\mathbf{z_1}$, $\mathbf{z_2}$ both satisfy $\mathbf{x}-\mathbf{z_i} \perp \mathcal{C}(\mathbf{A})$ $(i=1,2)$, then $\mathbf{z_1}-\mathbf{z_2} \in \mathcal{C}(\mathbf{A})$ and $\mathbf{z_1}-\mathbf{z_2} = (\mathbf{x}-\mathbf{z_2}) - (\mathbf{x}-\mathbf{z_1}) \perp \mathcal{C}(\mathbf{A}) \Longrightarrow \mathbf{z_1}-\mathbf{z_2} = \mathbf{0} \Longrightarrow \mathbf{z_1} = \mathbf{z_2}$, i.e. the minimizer $\mathbf{P_A}(\mathbf{x})$ is unique.

2. (a)
   $$\mathbf{P}_{\mathbf{v}_1}(\mathbf{w}) = \underset{\substack{\alpha\mathbf{v_1} \\ \alpha\in\mathbb{R}, \mathbf{v}_1\in\mathbb{R}^n}}{\mathbf{argmin}}\ \|\mathbf{w}-\alpha\mathbf{v}_1\|_2$$
   $$\frac{\mathrm{d}}{\mathrm{d}\alpha}\|\mathbf{w}-\alpha\mathbf{v}_1\|_2^2 = \frac{\mathrm{d}}{\mathrm{d}\alpha}(\mathbf{w}-\alpha\mathbf{v}_1)^\top(\mathbf{w}-\alpha\mathbf{v}_1) = -2\mathbf{v}_1^\top(\mathbf{w}-\alpha\mathbf{v}_1) = 0$$
   $$\Longrightarrow \alpha^* = \frac{\mathbf{v_1^T w}}{\mathbf{v_1^T v_1}} \Longrightarrow \mathbf{P}_{\mathbf{v}_1}(\mathbf{w}) = \frac{\mathbf{v_1^T w}}{\mathbf{v_1^T v_1}}\mathbf{v}_1$$

   (b) According to the result in (a),
   $$\begin{aligned}\mathbf{P}_{\mathbf{v}_1}(\alpha\mathbf{u}+\beta\mathbf{w}) = \frac{\mathbf{v_1^T}(\alpha\mathbf{u}+\beta\mathbf{w})}{\mathbf{v_1^T v_1}}\mathbf{v}_1 &= \alpha\frac{\mathbf{v_1^T u}}{\mathbf{v_1^T v_1}}\mathbf{v}_1 + \beta\frac{\mathbf{v_1^T w}}{\mathbf{v_1^T v_1}}\mathbf{v}_1 \\ &= \alpha\mathbf{P}_{\mathbf{v}_1}(\mathbf{u}) + \beta\mathbf{P}_{\mathbf{v}_1}(\mathbf{w})\end{aligned}$$

(c)

$$\mathbf{P}_{\mathbf{v}_1}(\mathbf{w}) = \frac{\mathbf{v_1^T w}}{\mathbf{v_1^T v_1}}\mathbf{v}_1 = \left(\frac{\mathbf{v_1 v_1^T}}{\mathbf{v_1^T v_1}}\right)\mathbf{w} := \mathbf{H_1 w}$$

(d) i.

$$\mathbf{P_V}(\mathbf{w}) = \underset{\substack{\mathbf{Vy} \\ \mathbf{y}\in\mathbb{R}^d, \mathbf{V}\in\mathbb{R}^{n\times d}}}{\mathbf{argmin}} \|\mathbf{w} - \mathbf{Vy}\|_2$$

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{y}}\|\mathbf{w} - \mathbf{Vy}\|_2^2 = \frac{\mathrm{d}}{\mathrm{d}\mathbf{y}}(\mathbf{w} - \mathbf{Vy})^\top(\mathbf{w} - \mathbf{Vy}) = -2\mathbf{V}^\top(\mathbf{w} - \mathbf{Vy}) = 0$$

$\mathbf{v_i}$, $i = 1, \ldots, d$ are linearly independent $\Longrightarrow$ $\mathbf{V}$ has full column rank $\Longrightarrow$ $\mathbf{V}^\top\mathbf{V}$ is invertible.

$$\Longrightarrow \mathbf{y}^* = (\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}^\top\mathbf{w} \Longrightarrow \mathbf{P_V}(\mathbf{w}) = \mathbf{V}(\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}^\top\mathbf{w}$$

$$\Longrightarrow \mathbf{H} = \mathbf{V}(\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}^\top$$

ii.

$$\mathbf{v_i^T v_j} = 0, \forall i \neq j \Longrightarrow \mathbf{V}^\top\mathbf{V} = \begin{bmatrix} \mathbf{v_1^T v_1} & 0 & \cdots & 0 \\ 0 & \mathbf{v_2^T v_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{v_d^T v_d} \end{bmatrix}$$

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

$$\Longrightarrow \mathbf{H} = \sum_{i=1}^{d} \frac{\mathbf{v_i v_i^T}}{\mathbf{v_i^T v_i}}$$

3. (a)

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \Longrightarrow \mathbf{x} \in \mathcal{C}(\mathbf{A}) = \mathbb{R}^2 \Longrightarrow \mathbf{P_A}(\mathbf{x}) = \mathbf{x} = \mathbf{A}\cdot\mathbf{x}$$

The coordinates of $\mathbf{P_A}(\mathbf{x})$ with respect to the column vectors in $\mathbf{A}$ are unique and equal to $\mathbf{x}$ itself.

(b)

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix} \Longrightarrow \mathcal{C}(\mathbf{A}) = \{\alpha(1,1)^T : \alpha \in \mathbb{R}\}$$

$$\mathbf{P_A}(\mathbf{x}) = \underset{\substack{\alpha(1,1)^T \\ \alpha\in\mathbb{R}}}{\mathbf{argmin}} \|\mathbf{x} - \alpha(1,1)^T\|_2$$

$$\xrightarrow{2.(a)} \mathbf{P_A}(\mathbf{x}) = \frac{(1,1)\mathbf{x}}{(1,1)(1,1)^T}(1,1)^T = \frac{x_1 + x_2}{2}(1,1)^T = \mathbf{A}\cdot(c_1, c_2)^T$$

$\Longrightarrow$ The set of all coordinates of $\mathbf{P_A}(\mathbf{x})$ with respect to the column vectors in $\mathbf{A}$ is the affine line:

$$\left\{(c_1, c_2)^T \in \mathbb{R}^2 : c_1 + 2c_2 = \frac{x_1 + x_2}{2}\right\}$$

Thus the coordinates are not unique.

4. (a) Let λ be an eigenvalue of $\mathbf{P}$ and $\mathbf{v}$ be the corresponding eigenvector. Then $\mathbf{Pv} = \lambda\mathbf{v}$. Since $\mathbf{P}$ is a projection matrix, $\mathbf{Pv}$ is the projection of $\mathbf{v}$ onto $\mathcal{C}(\mathbf{P})$ $\Longrightarrow \mathbf{v} = \mathbf{Pv} + (\mathbf{v} - \mathbf{Pv})$, where $\mathbf{Pv} \in \mathcal{C}(\mathbf{P})$ and $\mathbf{v} - \mathbf{Pv} \in \mathcal{C}(\mathbf{P})^\perp$.

$$\Longrightarrow \|\mathbf{v}\|_2^2 = \|\mathbf{Pv}\|_2^2 + \|\mathbf{v} - \mathbf{Pv}\|_2^2 \xRightarrow{\mathbf{Pv}=\lambda\mathbf{v}} \left(\lambda^2 - \lambda\right)\|\mathbf{v}\|_2^2 = 0 \Longrightarrow \lambda \in \{0, 1\}$$

(b) ① Necessity:
$\forall\ \mathbf{x} \in \mathbb{R}^m$, $\mathbf{Px} \in \mathcal{C}(\mathbf{P})$. We need to prove that $\mathbf{x} - \mathbf{Px} \in \mathcal{C}(\mathbf{P})^\perp$.
$\forall\ \mathbf{y} \in \mathcal{C}(\mathbf{P})$, $\exists\ \mathbf{z} \in \mathbb{R}^m$, s.t. $\mathbf{y} = \mathbf{Pz}$. Then

$$\begin{aligned}\langle \mathbf{x} - \mathbf{Px}, \mathbf{y}\rangle = \langle \mathbf{x} - \mathbf{Px}, \mathbf{Pz}\rangle &= \langle \mathbf{P^T}(\mathbf{x} - \mathbf{Px}), \mathbf{z}\rangle \\ &\xlongequal{\mathbf{P^T}=\mathbf{P}} \langle \mathbf{P}(\mathbf{x} - \mathbf{Px}), \mathbf{z}\rangle \\ &\xlongequal{\mathbf{P^2}=\mathbf{P}} \langle \mathbf{Px} - \mathbf{P}^2\mathbf{x}, \mathbf{z}\rangle = 0\end{aligned}$$

$\Longrightarrow \mathbf{x} - \mathbf{Px} \in \mathcal{C}(\mathbf{P})^\perp \Longrightarrow \mathbf{P}$ is a projection matrix.
② Sufficiency:
$\mathbf{P}$ is a projection matrix $\Longrightarrow \forall\ \mathbf{x} \in \mathbb{R}^m$, $\mathbf{Px} \in \mathcal{C}(\mathbf{P})$
$\Longrightarrow \mathbf{P}(\mathbf{Px}) = \mathbf{argmin}_{\mathbf{z}\in\mathcal{C}(\mathbf{P})}\|\mathbf{Px} - \mathbf{z}\|_2 = \mathbf{Px} \Longrightarrow \mathbf{P^2} = \mathbf{P}$.
$\mathbf{P}$ is a projection matrix $\Longrightarrow \forall\ \mathbf{x}, \mathbf{y} \in \mathbb{R}^m$, $\mathbf{x} = \mathbf{Px} + (\mathbf{x} - \mathbf{Px})$, where $\mathbf{Px} \in \mathcal{C}(\mathbf{P})$ and $\mathbf{x} - \mathbf{Px} \in \mathcal{C}(\mathbf{P})^\perp$. Since $\mathbf{Py} \in \mathcal{C}(\mathbf{P})$, we have

$$\langle \mathbf{Px}, \mathbf{y} - \mathbf{Py}\rangle = 0 \Longrightarrow \langle \mathbf{Px}, \mathbf{y}\rangle = \langle \mathbf{Px}, \mathbf{Py}\rangle = \langle \mathbf{x}, \mathbf{Py}\rangle = \langle \mathbf{P^T}\mathbf{x}, \mathbf{y}\rangle$$

$\Longrightarrow \mathbf{P^T} = \mathbf{P}$.

5. $\forall\ \mathbf{x} \in \mathbb{R}^m$,

$\mathcal{C}(\mathbf{B})$ is a proper subspace of $\mathcal{C}(\mathbf{A}) \Longrightarrow \forall \mathbf{z} \in \mathcal{C}(\mathbf{P}), z \in \mathcal{C}(\mathbf{A}) \Longrightarrow \mathbf{P_A}(\mathbf{x}) - \mathbf{z} \in \mathcal{C}(\mathbf{A})$

$$\begin{aligned}\mathbf{x} - \mathbf{P_A}(\mathbf{x}) \in \mathcal{C}(\mathbf{A})^\perp \Longrightarrow \|\mathbf{x} - z\|_2^2 &= \|(\mathbf{P_A}(\mathbf{x}) - \mathbf{z}) + (\mathbf{x} - \mathbf{P_A}(\mathbf{x}))\|_2^2 \\ &= \|\mathbf{P_A}(\mathbf{x}) - \mathbf{z}\|_2^2 + \|\mathbf{x} - \mathbf{P_A}(\mathbf{x})\|_2^2\end{aligned}$$

$$\Longrightarrow \underset{\mathbf{z}\in\mathcal{C}(\mathbf{B})}{\mathbf{argmin}}\|\mathbf{x} - \mathbf{z}\|_2 = \underset{\mathbf{z}\in\mathcal{C}(\mathbf{B})}{\mathbf{argmin}}\|\mathbf{P_A}(\mathbf{x}) - \mathbf{z}\|_2 \Longrightarrow \mathbf{P_B}(\mathbf{x}) = \mathbf{P_B}(\mathbf{P_A}(\mathbf{x}))$$

■

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 2: Projection to a Matrix Space

1. Set $\mathcal{D} = \{D \in \mathbb{R}^{n\times n} \mid D \text{ is diagonal}\}$ as the set of diagonal matrices.

   $\forall\ D_1,\ D_2 \in \mathcal{D}$ and $\alpha,\ \beta \in \mathbb{R}$, we have $\alpha D_1 + \beta D_2$ is diagonal $\Longrightarrow \alpha D_1 + \beta D_2 \in \mathcal{D}$. Besides, $\mathbf{0} \in \mathcal{D}$.

   Hence $\mathcal{D}$ is a linear space.

   $\forall\ A = (a_{ij})_{n\times n} \in \mathbb{R}^{n\times n}$, the projection of A onto $\mathcal{D}$ is given by:

$$\begin{aligned}
\mathbf{P}_{\mathcal{D}}(A) &= \underset{D=\mathrm{diag}(x_1,\cdots,x_n)\in\mathcal{D}}{\mathbf{argmin}} \|A-D\|_F = \underset{D=\mathrm{diag}(x_1,\cdots,x_n)\in\mathcal{D}}{\mathbf{argmin}} \|A-D\|_F^2 \\
&= \underset{D=\mathrm{diag}(x_1,\cdots,x_n)\in\mathcal{D}}{\mathbf{argmin}} \langle A-D, A-D\rangle_F \\
&= \underset{D=\mathrm{diag}(x_1,\cdots,x_n)\in\mathcal{D}}{\mathbf{argmin}} \sum_{i,j}(a_{ij}-x_i\delta_{ij})^2 = \underset{D=\mathrm{diag}(x_1,\cdots,x_n)\in\mathcal{D}}{\mathbf{argmin}} \sum_{i}(a_{ii}-x_i)^2 \\
&= \mathrm{diag}(a_{11},\cdots,a_{nn}) = \mathrm{diag}(A)
\end{aligned}$$

2. $\forall\ S_1,\ S_2 \in S^n$ and $\alpha,\ \beta \in \mathbb{R}$, $(\alpha S_1 + \beta S_2)^\top = \alpha S_1^\top + \beta S_2^\top = \alpha S_1 + \beta S_2 \Longrightarrow \alpha S_1 + \beta S_2 \in S^n$. Besides, $\mathbf{0} \in S^n$.

   Hence S^n is a linear space.

   It is clear that there is a set of linearly independent basis that can represent all the matrices in the space:

$$\{E_{ii}\}_{i=1}^n \cup \{E_{ij}+E_{ji}\}_{1\le i<j\le n},$$

   where E_{ij} is the matrix with 1 in the (i,j) position and 0 elsewhere.

   Therefore,

$$\dim S^n = n + \binom{n}{2} = \frac{n(n+1)}{2}.$$

3. Denote K^n as the set of skew-symmetric matrices.

   Then $\forall\ A \in S^n$ and $B \in K^n$, we have

$$\begin{aligned}
\langle A, B\rangle = \mathrm{tr}(A^\top B) &= \mathrm{tr}(B^\top A) \\
&\overset{B\in K^n}{=\!=\!=} -\mathrm{tr}(BA) \\
&\overset{A\in S^n}{=\!=\!=} -\mathrm{tr}(BA^\top) = -\mathrm{tr}(A^\top B) = -\langle A, B\rangle \Longrightarrow \langle A, B\rangle = 0.
\end{aligned}$$

   Decomposition:

$$A = \underbrace{\frac{A+A^\top}{2}}_{\overset{\text{def}}{=} S(A)\in S^n} + \underbrace{\frac{A-A^\top}{2}}_{\overset{\text{def}}{=} K(A)\in K^n}.$$

4. According to (3), $A = S(A) + K(A)$, where $S(A) \in S^n$ and $K(A) \perp S^n$.

   Therefore, $\mathbf{P}_{S^n}(A) = S(A) = \frac{A+A^\top}{2}$.

■

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

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

   [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 3: Projection to a Function Space

1. (a) First, $L^2(\Omega)$ is a linear space: $\forall\ Z_1,\ Z_2 \in L^2(\Omega)$ and $\alpha, \beta \in \mathbb{R}$,

$$(\alpha Z_1+\beta Z_2)^2 \leq 2\alpha^2 Z_1^2+2\beta^2 Z_2^2 \Longrightarrow \mathbb{E}[(\alpha Z_1+\beta Z_2)^2] \leq 2\alpha^2\mathbb{E}[Z_1^2]+2\beta^2\mathbb{E}[Z_2^2] < \infty.$$

Thus $\alpha Z_1 + \beta Z_2 \in L^2(\Omega)$.
Second, $\langle X, Y\rangle := \mathbb{E}[XY]$ defines an inner product in $L^2(\Omega)$: $\forall\ X, Y, X_1, X_2 \in L^2(\Omega)$ and $\alpha, \beta \in \mathbb{R}$,

- Positivity:

$$\langle X, X\rangle = \mathbb{E}[X^2] \geq 0, \text{ and } \langle X, X\rangle = 0 \Longleftrightarrow X = 0\ a.s.$$

- Symmetry:

$$\langle X, Y\rangle = \mathbb{E}[XY] = \mathbb{E}[YX] = \langle Y, X\rangle.$$

- Linearity:

$$\langle \alpha X_1+\beta X_2, Y\rangle = \mathbb{E}[(\alpha X_1+\beta X_2)Y] = \alpha\mathbb{E}[X_1Y]+\beta\mathbb{E}[X_2Y] = \alpha\langle X_1, Y\rangle+\beta\langle X_2, Y\rangle.$$

$$\langle X, \alpha Y_1 + \beta Y_2\rangle = \mathbb{E}[X(\alpha Y_1 + \beta Y_2)] = \alpha\mathbb{E}[XY_1] + \beta\mathbb{E}[XY_2] = \alpha\langle X, Y_1\rangle + \beta\langle X, Y_2\rangle.$$

Third, set $\mathcal{C} = \{c \cdot \mathbf{1} : c \in \mathbb{R}^2\} \subset L^2(\Omega)$, where $\mathbf{1}$ is the constant function with value 1. The projection of Y on $\mathcal{C}$ is

$$\mathbf{P}_{\mathcal{C}}(Y) = \frac{\langle Y, \mathbf{1}\rangle}{\langle \mathbf{1}, \mathbf{1}\rangle}\mathbf{1} = \mathbb{E}[Y] \cdot \mathbf{1}.$$

(b)

$$\begin{aligned}\mathbb{E}[(Y - c)^2] = \mathbb{E}[Y^2] - 2c\mathbb{E}[Y] + c^2 &= (c - \mathbb{E}[Y])^2 + \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 \\ &= (c - \mathbb{E}[Y])^2 + \mathrm{Var}(Y).\end{aligned}$$

$$\Longrightarrow \hat{c} = \underset{c\in\mathbb{R}}{\mathbf{argmin}}\, \mathbb{E}[(Y - c)^2] = \mathbb{E}[Y]$$

(c)

$$\min_{c\in\mathbb{R}} \mathbb{E}[(Y - c)^2] = \mathrm{Var}(Y) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = \mathbb{E}[Y^2] \Longleftrightarrow \mathbb{E}[Y] = 0.$$

According to part (a), the geometric interpretation is that the projection of Y on $\mathcal{C}$ is $\mathbb{E}(Y) = 0$, i.e., Y is orthogonal to all constant functions $\Longleftrightarrow$ Y lies in the orthogonal complement, $\mathcal{C}^{\perp}$.

2. (a)

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

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

(b)

$$\begin{aligned}
\mathbf{P}_{\mathcal{C}(X)}(Y) &= \underset{f:\mathbb{R}\to\mathbb{R}}{\mathbf{argmin}}\{\|f(X)-Y\| : f(X)\in\mathcal{C}(X)\} \\
&= \underset{f:\mathbb{R}\to\mathbb{R}}{\mathbf{argmin}}\{\|f(X)-Y\| : \mathbb{E}[f(X)^2]<\infty\} \\
&= \underset{f:\mathbb{R}\to\mathbb{R}}{\mathbf{argmin}}\{\|f(X)-Y\|^2 : \mathbb{E}[f(X)^2]<\infty\} \\
&= \underset{f:\mathbb{R}\to\mathbb{R}}{\mathbf{argmin}}\{\langle f(X)-Y, f(X)-Y\rangle : \mathbb{E}[f(X)^2]<\infty\} \\
&= \underset{f:\mathbb{R}\to\mathbb{R}}{\mathbf{argmin}}\{\mathbb{E}[(f(X)-Y)^2] : \mathbb{E}[f(X)^2]<\infty\} \\
&\overset{(a)}{=\!=} \mathbb{E}[Y|X]
\end{aligned}$$

where $\mathbb{E}[(\mathbb{E}[Y|X])^2] \overset{\phi(t)=t^2}{=\!=\!=} \mathbb{E}[\phi(\mathbb{E}[Y|X])] \overset{Jansen}{\leq} \mathbb{E}[\mathbb{E}[\phi(Y)|X]] = \mathbb{E}[Y^2|X] < \infty$.

(c) When $f(X)$ is confined with constant subspace, question 1 is reduced to question 2, i.e., $\mathbb{E}[Y|X \equiv c] = \mathbb{E}[Y]$. Geometrically, the conditional expectation $\mathbb{E}[Y|X]$ is the projection of Y onto the subspace $\mathcal{C}(X)$ of $L^2(\Omega)$ consisting of all square-integrable functions of X. When X is a constant (so that the σ-algebra generated by X is trivial), then $\mathcal{C}(X)$ is the subspace of constants. Therefore, the projection in Question 2 reduces to the projection in Question 1.

■

Exercise 4: Multicollinearity

Consider the linear regression problem formulated as below:

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}, \mathbb{E}(\mathbf{e}) = \mathbf{0}, \mathrm{Cov}(\mathbf{e}) = \sigma^2\mathbf{I_n},$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$ and $\mathbf{X} \in \mathbb{R}^{n\times p}$. Suppose that $\mathbf{X}^\top\mathbf{X}$ is invertible, then $\hat{\mathbf{w}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$ is the least squares estimator of $\mathbf{w}$.

1. Recall that the covariance matrix of p-dimensional random vectors is defined as

   $$\mathrm{Cov}(\hat{\mathbf{w}}) = \mathbb{E}[(\hat{\mathbf{w}} - \mathbb{E}(\hat{\mathbf{w}}))(\hat{\mathbf{w}} - \mathbb{E}(\hat{\mathbf{w}}))^\top].$$

   Please show that

   (a) $\mathbb{E}(\hat{\mathbf{w}}) = \mathbf{w}$;

   (b) $\mathrm{Cov}(\hat{\mathbf{w}}) = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$.

2. We usually measure the quality of an estimator by mean squared error (MSE). The mean squared error (MSE) of estimator $\hat{\mathbf{w}}$ is defined as

   $$\mathrm{MSE}(\hat{\mathbf{w}}) = \mathbb{E}[\|\hat{\mathbf{w}} - \mathbf{w}\|^2].$$

   Please derive that MSE can be decomposed into the variance of the estimator and the squared bias of the estimator, i.e.,

   $$\begin{aligned}\mathrm{MSE}(\hat{\mathbf{w}}) &= \mathrm{tr}\left(\mathrm{Cov}(\hat{\mathbf{w}})\right) + \|\mathbb{E}[\hat{\mathbf{w}}] - \mathbf{w}\|^2 \\ &= \sum_{i=1}^{p}\mathrm{Var}(\hat{w_i}) + \sum_{i=1}^{p}(\mathbb{E}(\hat{w_i}) - w_i)^2.\end{aligned}$$

3. Please show that

   $$\mathrm{MSE}(\hat{\mathbf{w}}) = \sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i},$$

   where $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the eigenvalues of $\mathbf{X}^\top\mathbf{X}$.

4. What would happen if there exists an eigenvalue $\lambda_k \approx 0$?

Solution 4: Multicollinearity

1. (a)

$$\begin{aligned}\mathbb{E}(\hat{\mathbf{w}}) = \mathbb{E}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}\right] &= \mathbb{E}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top(\mathbf{X}\mathbf{w}+\mathbf{e})\right] \\ &= \mathbb{E}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\mathbf{X}^\top\mathbf{X}\right)\mathbf{w} + \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{e}\right] \\ &= \mathbf{w} + \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbb{E}(\mathbf{e}) = \mathbf{w}\end{aligned}$$

   (b) According to (a),

$$\begin{aligned}\mathrm{Cov}(\hat{\mathbf{w}}) &= \mathbb{E}[(\hat{\mathbf{w}}-\mathbf{w})(\hat{\mathbf{w}}-\mathbf{w})^\top] \\ &= \mathbb{E}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{e}\mathbf{e}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right] \\ &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbb{E}[\mathbf{e}\mathbf{e}^\top]\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \\ &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\left(\sigma^2\mathbf{I}_n + \mathbb{E}[\mathbf{e}]\mathbf{E}[\mathbf{e}]^\top\right)\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\end{aligned}$$

2.

$$\begin{aligned}\|\hat{\mathbf{w}}-\mathbf{w}\|^2 &= \|\hat{\mathbf{w}}-\mathbb{E}[\hat{\mathbf{w}}]+\mathbb{E}[\hat{\mathbf{w}}]-\mathbf{w}\|^2 \\ &= \|\hat{\mathbf{w}}-\mathbb{E}[\hat{\mathbf{w}}]\|^2 + \|\mathbb{E}[\hat{\mathbf{w}}]-\mathbf{w}\|^2 + 2\left(\hat{\mathbf{w}}-\mathbb{E}[\hat{\mathbf{w}}]\right)^\top\left(\mathbb{E}[\hat{\mathbf{w}}]-\mathbf{w}\right)\end{aligned}$$

$$\begin{aligned}\implies \mathrm{MSE}(\hat{\mathbf{w}}) = \mathbb{E}[\|\hat{\mathbf{w}}-\mathbf{w}\|^2] &= \mathbb{E}[\|\hat{\mathbf{w}}-\mathbb{E}[\hat{\mathbf{w}}]\|^2] + \|\mathbb{E}[\hat{\mathbf{w}}]-\mathbf{w}\|^2 \\ &= \mathbb{E}\left[\mathrm{tr}\left((\hat{\mathbf{w}}-\mathbb{E}[\hat{\mathbf{w}}])(\hat{\mathbf{w}}-\mathbb{E}[\hat{\mathbf{w}}])^\top\right)\right] + \|\mathbb{E}[\hat{\mathbf{w}}]-\mathbf{w}\|^2 \\ &= \mathrm{tr}\left(\mathbb{E}\left[(\hat{\mathbf{w}}-\mathbb{E}[\hat{\mathbf{w}}])(\hat{\mathbf{w}}-\mathbb{E}[\hat{\mathbf{w}}])^\top\right]\right) + \|\mathbb{E}[\hat{\mathbf{w}}]-\mathbf{w}\|^2 \\ &= \mathrm{tr}\left(\mathrm{Cov}(\hat{\mathbf{w}})\right) + \|\mathbb{E}[\hat{\mathbf{w}}]-\mathbf{w}\|^2 \\ &= \sum_{i=1}^{p}\mathrm{Var}(\hat{w_i}) + \sum_{i=1}^{p}(\mathbb{E}(\hat{w_i})-w_i)^2.\end{aligned}$$

3.

$$\mathrm{MSE}(\hat{\mathbf{w}}) \overset{(2)}{=\!=} \mathrm{tr}\left(\mathrm{Cov}(\hat{\mathbf{w}})\right) + \|\mathbb{E}[\hat{\mathbf{w}}]-\mathbf{w}\|^2 \overset{(1)}{=\!=} \sigma^2\,\mathrm{tr}\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right)$$

$\mathbf{X}^\top\mathbf{X}$ is invertible $\implies$ $\mathbf{X}$ is full column rank $\implies$ $\mathbf{X}^\top\mathbf{X}$ is symmetric positive definite $\implies$ there exists an orthogonal matrix $\mathbf{Q}$ and a diagonal matrix $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \cdot, \lambda_p)$, such that

$$\mathbf{X}^\top\mathbf{X} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top = \mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^\top$$

$$\implies \mathrm{MSE}(\hat{\mathbf{w}}) = \sigma^2\,\mathrm{tr}\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right) = \sigma^2\,\mathrm{tr}\left(\mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^\top\right) = \sigma^2\,\mathrm{tr}(\Lambda^{-1}) = \sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i}$$

4. According to (3), if there exists an eigenvalue $\lambda_k \approx 0$, then $\text{MSE}(\hat{\mathbf{w}}) \to +\infty$. This indicates that the estimator $\hat{\mathbf{w}}$ is very unstable, i.e., a small change in $\mathbf{y}$ may result in a large change in $\hat{\mathbf{w}}$. Therefore, multicollinearity (i.e., some features are highly linearly correlated, leading to $\lambda_k \approx 0$) should be avoided in linear regression.

■

Exercise 5: 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

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

   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 5: Regularized least squares

1. $\forall \mathbf{z} \in \mathbb{R}^d$,
$$\mathbf{z}^\top \mathbf{X}^\top \mathbf{X}\mathbf{z} = (\mathbf{X}\mathbf{z})^\top (\mathbf{X}\mathbf{z}) = \|\mathbf{X}\mathbf{z}\|_2^2 \geq 0.$$
Therefore, $\mathbf{X}^\top \mathbf{X}$ is positive semi-definite.
Moreover, $\mathbf{X}^\top \mathbf{X}$ is positive definite $\iff \forall \mathbf{z} \in \mathbb{R}^d \backslash \{\mathbf{0}\}$, $\mathbf{z}^\top \mathbf{X}^\top \mathbf{X}\mathbf{z} = \|\mathbf{X}\mathbf{z}\|_2^2 > 0$
$$\iff \mathbf{X}\mathbf{z} \neq \mathbf{0} \iff \text{the columns of } \mathbf{X} \text{ are linearly independent.}$$

2. $\forall \mathbf{z} \in \mathbb{R}^d \backslash \{\mathbf{0}\}$,
$$\mathbf{z}^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})\mathbf{z} = \mathbf{z}^\top \mathbf{X}^\top \mathbf{X}\mathbf{z} + \lambda \mathbf{z}^\top \mathbf{I}\mathbf{z} = \|\mathbf{X}\mathbf{z}\|_2^2 + \lambda \|\mathbf{z}\|_2^2 > 0.$$
Therefore, $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}$ is positive definite $\implies \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}$ is invertible.

3. Let $\mathbf{w}_1^* = \mathbf{w}^*(\lambda_1)$ and $\mathbf{w}_2^* = \mathbf{w}^*(\lambda_2)$.
Since $\mathbf{w}_1^*$ is the minimizer of $L(\mathbf{w}) + \lambda_1 \Omega(\mathbf{w})$,
$$L(\mathbf{w}_1^*) + \lambda_1 \Omega(\mathbf{w}_1^*) \leq L(\mathbf{w}_2^*) + \lambda_1 \Omega(\mathbf{w}_2^*). \tag{1}$$
Similarly, since $\mathbf{w}_2^*$ is the minimizer of $L(\mathbf{w}) + \lambda_2 \Omega(\mathbf{w})$,
$$L(\mathbf{w}_2^*) + \lambda_2 \Omega(\mathbf{w}_2^*) \leq L(\mathbf{w}_1^*) + \lambda_2 \Omega(\mathbf{w}_1^*). \tag{2}$$
Adding the above two inequalities gives
$$(\lambda_1 - \lambda_2)(\Omega(\mathbf{w}_1^*) - \Omega(\mathbf{w}_2^*)) \leq 0. \xrightarrow{\lambda_1 < \lambda_2} \Omega(\mathbf{w}_1^*) \geq \Omega(\mathbf{w}_2^*).$$
$$\implies L(\mathbf{w}_2^*) - L(\mathbf{w}_1^*) \geq \lambda_1 (\Omega(\mathbf{w}_1^*) - \Omega(\mathbf{w}_2^*)) \geq 0. \implies L(\mathbf{w}_1^*) \leq L(\mathbf{w}_2^*).$$
It is clear that
$$\Omega(\mathbf{w}_1^*) = \Omega(\mathbf{w}_2^*) \underset{(1)}{\overset{(2)}{\rightleftharpoons}} L(\mathbf{w}_1^*) = L(\mathbf{w}_2^*)$$
Thus, if one of the above two equalities holds, the other one also holds. However, this contradicts the Strict convexity of $\mathbf{w}^*(\lambda)$, i.e. $\nabla^2[L(\mathbf{w}) + \lambda \Omega(\mathbf{w})] = \frac{2}{n}\mathbf{X}^\top \mathbf{X} + 2\lambda \mathbf{I} > 0$ when $\mathbf{w} \neq 0$. Therefore, the two inequalities in (1) and (2) are strict, i.e.,
$$L(\mathbf{w}^*(\lambda_1)) < L(\mathbf{w}^*(\lambda_2)), \qquad \Omega(\mathbf{w}^*(\lambda_1)) > \Omega(\mathbf{w}^*(\lambda_2)).$$
Intuitively, a larger regularization parameter λ places more emphasis on minimizing the regularization term $\Omega(\mathbf{w})$, which encourages smaller norm solutions and simultaneously diminishes the regularization term. As a result, the model may fit the training data less closely, leading to a higher loss $L(\mathbf{w})$. Conversely, a smaller λ allows the model to focus more on minimizing the loss, potentially resulting in a better fit to the training data but with a larger norm for $\mathbf{w}$.

■

Exercise 6: High-Dimensional Linear Regression for Image Warping (Programming Exercise)

Consider a data set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ where $\mathbf{x}_i \in \mathbb{R}^n, \mathbf{y}_i \in \mathbb{R}^m,\ i = 1,\ 2,\ \cdots N$, we want to find a map $\phi : \mathbb{R}^n \to \mathbb{R}^m$ such that $\phi(\mathbf{x}_i) = \mathbf{y}_i,\ i = 1,\ 2,\ \cdots N$. Now given a set of basis functions $\phi_i : \mathbb{R}^n \to \mathbb{R}$

$$\phi_i(\mathbf{x}) = \left(\|\mathbf{x} - \mathbf{x}_i\|_2^2 + r^2\right)^{\frac{\mu}{2}}$$

where μ, r are costume constants. We can approximate function ϕ by a $\hat{\phi}(\mathbf{x})$:

$$\hat{\phi}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b} + \mathbf{W}\boldsymbol{\phi}(\mathbf{x})$$

which is a linear combination of basis functions, where $\boldsymbol{\phi}(\mathbf{x}) := \begin{pmatrix}\phi_1(\mathbf{x}) & \phi_2(\mathbf{x}) & \dots & \phi_N(\mathbf{x})\end{pmatrix}^T \in \mathbb{R}^N$, and parameters $\mathbf{W} \in \mathbb{R}^{m\times N}, \mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m$.

1. Find $\mathbf{W}, \mathbf{A}, \mathbf{b}$ such that

$$\min_{\mathbf{W},\mathbf{A},\mathbf{b}} l := \sum_{i=1}^{N} \left\| \hat{\phi}(\mathbf{x}_i) - \mathbf{y}_i \right\|_2^2 + \lambda_1 \|\mathbf{A} - \mathbf{I}\|_f^2 + \lambda_2 \|\mathbf{b}\|_2^2 + \lambda_3 \|\mathbf{W}\|_f^2 \tag{3}$$

   where $\|\cdot\|_f$ is the Frobenius norm and $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}_+$.

   Hint (1) Take the partial derivatives with respect to $\mathbf{W}, \mathbf{A}, \mathbf{b}$, and set them to zero. (2) $\partial_{\mathbf{X}} \|\mathbf{X} - \mathbf{C}\|_f^2 = 2(\mathbf{X} - \mathbf{C})$.

2. Image warping is a technique used to smoothly distort or reshape an image based on specified transformation rules (shown in Figure 1). The user defines these rules by selecting points on the image, with each transformation mapping an initial position $\mathbf{x}_i$ to a target position $\mathbf{y}_i$. These point pairs form the training set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, where $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^2$. A linear function $\hat{\phi}$ is then learned by minimizing the loss l over the training set. This function is applied to every pixel $\hat{\mathbf{x}}_\mathbf{i}$ in the test image, mapping it to an output coordinate $\hat{\mathbf{y}}_\mathbf{i} = \hat{\phi}(\hat{\mathbf{x}}_\mathbf{i})$. Finally, the pixel value at $\hat{\mathbf{y}}_\mathbf{i}$ replaces that at $\hat{\mathbf{x}}_\mathbf{i}$, generating the warped image.

   Now let $r = 0.5, \mu = 1, \lambda_1 = \lambda_2 = \lambda_3 = 1e-2$. Please implement the image warping method in the provided framework. You can submit any image you like.

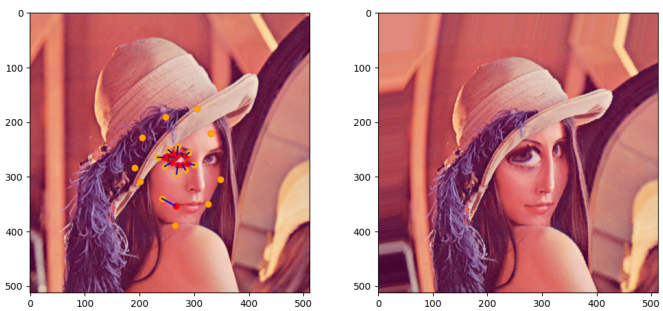

Figure 1: Image warping example

Solution 6: High-Dimensional Linear Regression for Image Warping

1. Define

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{n\times N}, \qquad \mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N] \in \mathbb{R}^{m\times N}$$

$$\mathbf{\Phi} \in \mathbb{R}^{N\times N}, (\mathbf{\Phi})_{ij} = \phi_j(\mathbf{x}_i) \qquad \mathbf{1}_N = (1, 1, \ldots, 1)^T \in \mathbb{R}^N$$

$$\begin{aligned}
\Longrightarrow l &= \sum_{i=1}^{N} \left\| \hat{\phi}(\mathbf{x}_i) - \mathbf{y}_i \right\|_2^2 + \lambda_1 \|\mathbf{A} - \mathbf{I}\|_f^2 + \lambda_2 \|\mathbf{b}\|_2^2 + \lambda_3 \|\mathbf{W}\|_f^2 \\
&= \left\| \mathbf{A}\mathbf{X} + \mathbf{b}\mathbf{1}_N^\top + \mathbf{W}\mathbf{\Phi} - \mathbf{Y} \right\|_f^2 + \lambda_1 \|\mathbf{A} - \mathbf{I}\|_f^2 + \lambda_2 \|\mathbf{b}\|_2^2 + \lambda_3 \|\mathbf{W}\|_f^2
\end{aligned}$$

$$\Longrightarrow \begin{cases}
\partial_{\mathbf{W}} l = 2\left(\mathbf{A}\mathbf{X} + \mathbf{b}\mathbf{1}_N^\top + \mathbf{W}\mathbf{\Phi} - \mathbf{Y}\right)\mathbf{\Phi}^\top + 2\lambda_3 \mathbf{W} = 0 \\
\partial_{\mathbf{A}} l = 2\left(\mathbf{A}\mathbf{X} + \mathbf{b}\mathbf{1}_N^\top + \mathbf{W}\mathbf{\Phi} - \mathbf{Y}\right)\mathbf{X}^\top + 2\lambda_1(\mathbf{A} - \mathbf{I}) = 0 \\
\partial_{\mathbf{b}} l = 2\left(\mathbf{A}\mathbf{X} + \mathbf{b}\mathbf{1}_N^\top + \mathbf{W}\mathbf{\Phi} - \mathbf{Y}\right)\mathbf{1}_N + 2\lambda_2 \mathbf{b} = 0
\end{cases}$$

$$\Longrightarrow \begin{cases}
\mathbf{W}^* = \left(\mathbf{Y} - \mathbf{A}^*\mathbf{X} - \mathbf{b}^*\mathbf{1}_N^\top\right)\mathbf{\Phi}^\top \left(\mathbf{\Phi}\mathbf{\Phi}^\top + \lambda_3 \mathbf{I}\right)^{-1} \\
\mathbf{A}^* = \left(\left(\mathbf{Y} - \mathbf{b}^*\mathbf{1}_N^\top - \mathbf{W}^*\mathbf{\Phi}\right)\mathbf{X}^\top + \lambda_1 \mathbf{I}\right)\left(\mathbf{X}\mathbf{X}^\top + \lambda_1 \mathbf{I}\right)^{-1} \\
\mathbf{b}^* = (N + \lambda_2)^{-1}\left(\mathbf{Y} - \mathbf{A}^*\mathbf{X} - \mathbf{W}^*\mathbf{\Phi}\right)\mathbf{1}_N
\end{cases}$$

2. **Principles and Thought Process**:

   Define augmented design matrix:

$$M = \begin{bmatrix} X \\ \mathbf{1}_N \\ \Phi^\top \end{bmatrix} \in \mathbb{R}^{(2+1+N)\times N}$$

   Define augmented parameter matrix:

$$\Theta \;=\; \begin{bmatrix} A \\ b^\top \\ W^\top \end{bmatrix} \;\in\; \mathbb{R}^{(2+1+N)\times 2}.$$

   Then the model's prediction on all samples is:

$$\hat{\phi}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b} + \mathbf{W}\boldsymbol{\phi}(\mathbf{x}) = \Theta^\top \mathbf{M}$$

   Furthermore, our target is:

$$\min_{\mathbf{W},\mathbf{A},\mathbf{b}} l = \left\| \mathbf{A}\mathbf{X} + \mathbf{b}\mathbf{1}_N^\top + \mathbf{W}\mathbf{\Phi} - \mathbf{Y} \right\|_f^2 + \lambda_1 \|\mathbf{A} - \mathbf{I}\|_f^2 + \lambda_2 \|\mathbf{b}\|_2^2 + \lambda_3 \|\mathbf{W}\|_f^2$$

Define

$$\mathbf{\Lambda} \;=\; \mathrm{diag}\big(\lambda_1 I_2,\; \lambda_2,\; \lambda_3 I_N\big) \;\in\; \mathbb{R}^{(2+1+N)\times(2+1+N)}.$$

$$\mathbf{C} \;=\; \begin{bmatrix} \lambda_1 I_2 \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \;\in\; \mathbb{R}^{(2+1+N)\times 2}.$$

Then the original target is equivalent to:

$$\min_{\Theta} l = \|\mathbf{Y} - \Theta^\top \mathbf{M}\|_f^2 + \mathrm{tr}\left(\Theta^\top \Lambda \Theta\right) - 2\,\mathrm{tr}\left(\mathbf{C}^\top \Theta\right) + \lambda_1\,\mathrm{tr}(\mathbf{I}_2)$$

because of $\|\mathbf{X}\|_f^2 = \mathrm{tr}\left(\mathbf{X}^\top \mathbf{X}\right)$.

Then we have:

$$\frac{\partial l}{\partial \Theta} = -2\mathbf{M}\mathbf{Y}^\top + 2\mathbf{M}\mathbf{M}^\top\Theta + 2\mathbf{\Lambda}\Theta - 2\mathbf{C}$$

$$\Longrightarrow \left(\mathbf{M}\mathbf{M}^\top + \mathbf{\Lambda}\right)\Theta = \mathbf{M}\mathbf{Y}^\top + \mathbf{C}$$

Finally, we can get Θ through solving this linear equation.

In summary, all of the "to do" codes in my .py files are from these principles.

Here is my trial result:

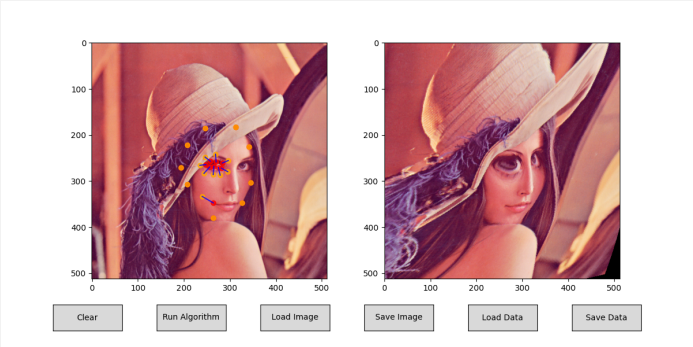

Figure 2: Image warping results

■

Exercise 7: 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 $(\text{bias})^2 + \text{variance}$ in one figure, as the functions of $\log_{10}\lambda$. (**Hint:** see [1] for an example.)

Solution 7: Bias-Variance Trade-off

1. For each data set $\mathcal{D}^{(l)}$, define the design matrix

$$\mathbf{\Phi}^{(l)} \in \mathbb{R}^{N\times 25}, \quad (\mathbf{\Phi}^{(l)})_{n,:} = (1, \phi_1(x_n), \phi_2(x_n), \ldots, \phi_{24}(x_n))$$

and the target vector

$$\mathbf{y}^{(l)} = (y_1^{(l)}, y_2^{(l)}, \ldots, y_N^{(l)})^\top \in \mathbb{R}^N.$$

Then the regularized error function can be rewritten as

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

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

$$\Longrightarrow \hat{\mathbf{w}}^{(l)} = \left((\mathbf{\Phi}^{(l)})^\top\mathbf{\Phi}^{(l)} + \lambda\mathbf{I}\right)^{-1}(\mathbf{\Phi}^{(l)})^\top\mathbf{y}^{(l)}$$

2. Here is prediction plot learned from "data_1":

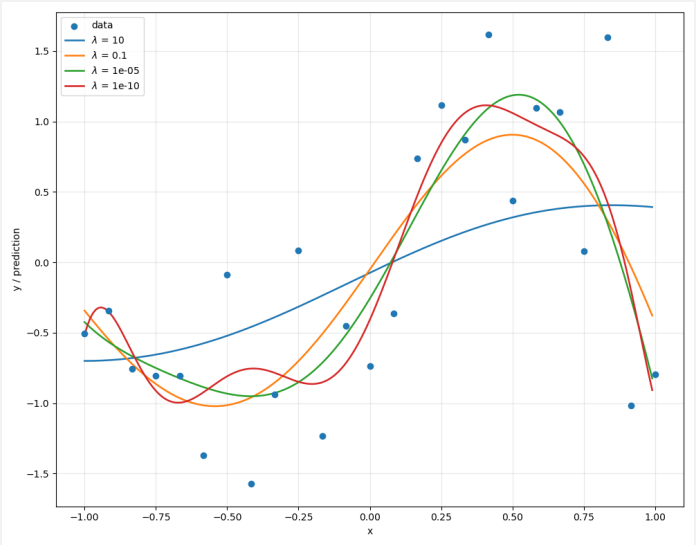

Figure 3: Prediction Plots

3. Here are Bias, Variance and total error plots:

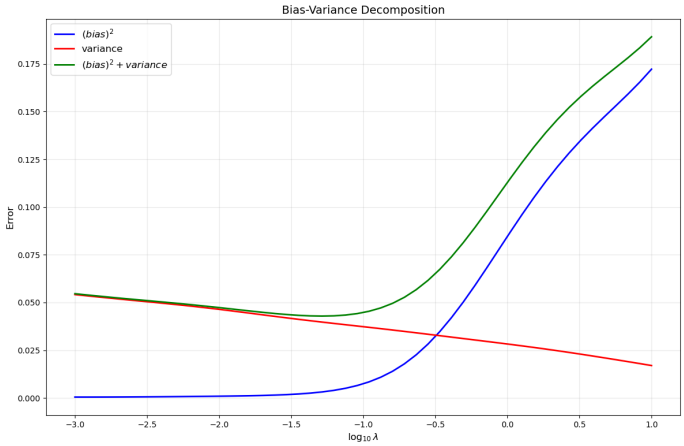

Figure 4: Error Plots

■

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.