Machine Learning Homework 2

Liu Pei PB22061259

Exercise 1: Projection

1.

Assume that there are two $\mathbf{P_A}(\mathbf{x})$,which we can denote them as $\mathbf{P_A}(\mathbf{x})$ and $\mathbf{P_A}(\mathbf{x})'$.
First,we are going to prove that $\mathcal{C}(A)$ is a convex set:

> we can choose any two points in $\mathcal{C}(A)$,which we can call $\mathbf{x}_1, \mathbf{x}_2$,because $\mathcal{C}(A)$ is a linear space,so we can get ,
>
> $$\forall \lambda_1, \lambda_2 \in \mathrm{R},\ \ \lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2 \in \mathcal{C}(A)$$
>
> So,it is easy to assert that
>
> $$\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C}(A) \text{ and } \forall \theta \in [0,1],\text{ we have: } \theta \mathbf{x}_1 + (1-\theta)\mathbf{x}_2 \in \mathcal{C}(A).$$
>
> Therefore, $\mathcal{C}(A)$ is convex.

Next,we are going to prove that two $\mathbf{P_A}(\mathbf{x})$ is ridiculous:

> Because $\mathcal{C}(A)$ is a convex set,so there is a line in $\mathcal{C}(A)$ which is between $\mathbf{P_A}(\mathbf{x})$ and $\mathbf{P_A}(\mathbf{x})'$.We denote the midpoint of this line as $\mathbf{m}$ .So, we have:
>
> $$\mathbf{m} = \frac{\mathbf{P_A}(\mathbf{x}) + \mathbf{P_A}(\mathbf{x})'}{2}$$
>
> Then, we can get:
>
> $$\begin{aligned} \|\mathbf{m} - \mathbf{z}\|_2 = \|\frac{\mathbf{P_A}(\mathbf{x}) + \mathbf{P_A}(\mathbf{x})'}{2} - \mathbf{z}\|_2 = \|\frac{\mathbf{P_A}(\mathbf{x})' - \mathbf{z}}{2} + \frac{\mathbf{P_A}(\mathbf{x}) - \mathbf{z}}{2}\|_2 \\ \leq \|\frac{\mathbf{P_A}(\mathbf{x}) - \mathbf{z}}{2}\|_2 + \|\frac{\mathbf{P_A}(\mathbf{x})' - \mathbf{z}}{2}\|_2 = \|\mathbf{P_A}(\mathbf{x}) - \mathbf{z}\|_2 \end{aligned}$$
>
> That is ridiculous,which is against the definition of projection.

So, we can conclude that $\mathbf{P_A}(\mathbf{x})$ is unique.

2.

(a)

We have:$\mathbf{P}_{\mathbf{v}_1}(\mathbf{w}) = \underset{\mathbf{z}\in\mathbb{R}^n}{\operatorname{argmin}}\{\|\mathbf{w} - \mathbf{z}\|_2 : \mathbf{z} \in \mathcal{C}(\mathbf{v}_1)\}$
Because $\mathbf{v}_i \in \mathbb{R}^n$,so we can get:

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

Let $f(\lambda) = (\mathbf{w} - \lambda \mathbf{v}_1)^T(\mathbf{w} - \lambda \mathbf{v}_1) = \mathbf{w}^T\mathbf{w} + \lambda^2 \mathbf{v}_1^T \mathbf{v}_1 - 2\lambda \mathbf{v}_1^T \mathbf{w}$,
then we let:

$$\frac{\partial f}{\partial \lambda} = 2\lambda \mathbf{v}_1^T \mathbf{v}_1 - 2\mathbf{v}_1^T \mathbf{w} = 0$$

So, we have:

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

Therefore, we have:
So, we have:

$$\mathbf{P}_{\mathbf{v}_1}(\mathbf{w}) = \frac{\mathbf{v}_1^T \mathbf{w}}{\mathbf{v}_1^T \mathbf{v}_1}\mathbf{v}_1$$

(b)

$$\begin{aligned}\mathbf{P}_{\mathbf{v}_1}(\alpha\mathbf{u}+\beta\mathbf{w}) = \underset{\mathbf{w},\mathbf{u}\in\mathbb{R}^n}{\operatorname{argmin}}\{\|\alpha\mathbf{u}+\beta\mathbf{w}-\lambda\mathbf{v}_1\|_2, \lambda\in\mathbb{R}\} = (\alpha\mathbf{u}+\beta\mathbf{w})\cdot\mathbf{v}_1(\mathbf{v}_1^T\mathbf{v}_1)^{-1} \\ = \alpha(\mathbf{u}\mathbf{v_1})\cdot\mathbf{v}_1(\mathbf{v}_1^T\mathbf{v}_1)^{-1} + \beta(\mathbf{w}\mathbf{v_1})\cdot\mathbf{v}_1(\mathbf{v}_1^T\mathbf{v}_1)^{-1} = \alpha\mathbf{P}_{\mathbf{v}_1}(\mathbf{u}) + \beta\mathbf{P}_{\mathbf{v}_1}(\mathbf{w})\end{aligned}$$

(c)

Let $\mathbf{P}_{\mathbf{v}_1}(\mathbf{w}) = \mathbf{H}_1\mathbf{w} = \frac{\mathbf{v}_1^T\mathbf{w}}{\mathbf{v}_1^T\mathbf{v}_1}\mathbf{v}_1$
So we have :$\mathbf{H}_1 = \mathbf{v}_1\mathbf{v_1}^T(\mathbf{v}_1^T\mathbf{v}_1)^{-1}$

(d)

Let $f(\lambda) = (\mathbf{w} - \sum_{i=1}^d \lambda_i\mathbf{v}_i)^T(\mathbf{w} - \sum_{i=1}^d \lambda_i\mathbf{v}_i) = \mathbf{w}^T\mathbf{w} - \sum_{i=1}^d \lambda_i\mathbf{v}_i^T\mathbf{w} - \sum_{i=1}^d \lambda_i\mathbf{w}^T\mathbf{v}_i + (\sum_{i=1}^d \lambda_i\mathbf{v}_i^T)(\sum_{i=1}^d \lambda_i\mathbf{v}_i)$
Then we let:

$$\frac{\partial f}{\partial\lambda_i} = -\mathbf{v}_i^T\mathbf{w} - \mathbf{w}^T\mathbf{v}_i + \mathbf{v}_i^T(\sum_{i=1}^d \lambda_i\mathbf{v}_i) + (\sum_{i=1}^d \lambda_i\mathbf{v}_i^T)\mathbf{v}_i = 2[\mathbf{v}_i^T(\sum_{i=1}^d \lambda_i\mathbf{v}_i) - \mathbf{v}_i^T\mathbf{w}] = 0$$

So, we have:

$$\mathbf{v}_i^T(\sum_{i=1}^d \lambda_i\mathbf{v}_i) = \mathbf{v}_i^T\mathbf{w}$$

> 不会写，但是d个方程确实有d对应d个未知数，但是写不出来解的表达式。

If we have $\mathbf{v}_i^T\mathbf{v}_j = 0, \forall i \neq j$,then we have:

$$\mathbf{v}_i^T(\sum_{i=1}^d \lambda_i\mathbf{v}_i) = \lambda_i\mathbf{v}_i^T\mathbf{v}_i = \mathbf{v}_i^T\mathbf{w}$$

So, we have:

$$\lambda_i = \frac{\mathbf{v}_i^T\mathbf{w}}{\mathbf{v}_i^T\mathbf{v}_i}$$

Therefore, we have:

$$\mathbf{P}_{\mathbf{v}}(\mathbf{w}) = \sum_{i=1}^d \frac{\mathbf{v}_i^T\mathbf{w}}{\mathbf{v}_i^T\mathbf{v}_i}\mathbf{v}_i$$

3.

(a)

The coordinates unique,and they are $[\mathbf{x}\cdot(1,0),\mathbf{x}\cdot(0,1)]$.

(b)

The column vectors in A are not linearly independent,so we cannot get a unique coordinate.

Exercise 2: Projection to a Matrix Space

1.

The first question is easy to prove,because we can get:
$\forall A,B\in\mathbb{R}^{n\times n}$,which are two diagonal matrices ,we have $\alpha A+\beta B$ is a diagonal matrix.And,the other conditions are obviously satisfied.
So, we assert that Show that the set of diagonal matrices in $\mathbb{R}^{n\times n}$
forms a linear space.

Let $\Sigma = diag(\lambda_1,\lambda_2,\cdots,\lambda_n)$,where $\lambda_i\in\mathbb{R}$.

Then we have:

$$\mathbf{P}_{\mathbf{\Sigma}}(\mathbf{A})=\underset{\mathbf{z}\in\Sigma}{\operatorname{argmin}}\{\|\mathbf{A}-\mathbf{z}\|_2:\mathbf{z}\in\Sigma\}$$

Here , A is a matrix in $\mathbb{R}^{n\times n}$,then we can get:

$$\begin{gathered}\|\mathbf{A}-\mathbf{z}\|_2=\|A-\Sigma\|_2=tr[(\mathbf{A}-\Sigma)^T(\mathbf{A}-\Sigma)]\\ =tr(\mathbf{A}^T\mathbf{A}-\mathbf{A}^T\Sigma-\Sigma^T\mathbf{A}+\Sigma^T\Sigma)\\ =tr(\mathbf{A}^T\mathbf{A})-2tr(\mathbf{A}^T\Sigma)+tr(\Sigma^T\Sigma)\text{ denoted as }f(\lambda)\end{gathered}$$

Then, we let:

$$\frac{\partial f}{\partial\lambda_i}=2\lambda_i-2a_{ii}=0$$

So, we have:$\lambda_{\mathbf{i}}=a_{ii}$
Therefore, we have:

$$\mathbf{P}_{\mathbf{\Sigma}}(\mathbf{A})=diag(a_{11},a_{22},\cdots,a_{nn})$$

2.

The second question is also easy to prove,because we can get:

> The set of symmetric matrices ,which is denoted as $\mathcal{S}$,we have any two symmetric matrices,there linear combination is absolutely symmetric.

So , we assert that Show that the set of symmetric matrices in $\mathbb{R}^{n\times n}$ is a linear space.
We can easily to compute the dimension of this linear space:
the diagonal dimension is n,the off-diagonal dimension is the totall number of the matrix elements n^2 minus the diagonal elements n,because the matrix is symmetric.,so we have:

$$dim(\mathcal{S})=n+\frac{n^2-n}{2}=\frac{n^2+n}{2}$$

3.

First ,we are going to prove $\mathrm{tr(AB) = tr(BA)}$:

> $$\mathrm{tr(AB)} = \sum_{\mathrm{i=1}}^{\mathrm{n}} [\mathrm{AB}]_{\mathrm{ii}} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \sum_{\mathrm{j=1}}^{\mathrm{n}} \mathrm{a_{ij} b_{ji}} = \sum_{\mathrm{j=1}}^{\mathrm{n}} \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{b_{ji} a_{ij}} = \sum_{\mathrm{j=1}}^{\mathrm{n}} [\mathrm{BA}]_{\mathrm{jj}} = \mathrm{tr(BA)}$$

Then,we consider a symmetric matrix $\mathbf{A}$and a d skew-symmetric matrix $\mathbf{B}$,we have:

> $$< \mathbf{A}, \mathbf{B} >= \mathrm{tr}(\mathbf{A^T B}) = tr(\mathbf{AB})$$
>
> Because$\mathbf{B}$ is skew-symmetric,so we have:
>
> $$tr(AB) = tr((AB)^T) = tr(B^T A^T) = tr(-BA) = -tr(\mathbf{BA}) = -tr(AB)$$

So. we have:

$$< \mathbf{A}, \mathbf{B} >= 0$$

Next,we are going to prove any matrix can be decomposed as the sum of a symmetric matrix and a skew-symmetric matrix.
Let $\mathbf{A}$ be a matrix,then we let:

$$A_1 = (A + A^T)/2, A_2 = (A - A^T)/2$$

It is easy to see that A_1 is a symmetric matrix and A_2 is a skew-symmetric matrix.

4.

Let $\mathbf{A}$ be a matrix,then we have:

$$\mathbf{A} = \mathbf{Q} + \mathbf{P}$$

In this case, $\mathbf{Q}$ is an symmetric matrix and $\mathbf{P}$ is a skew-symmetric matrix.
Let us denote symmetric matrices space as $\mathcal{S}$,then we have:

$$\mathbf{P}_{\mathcal{S}}(\mathbf{A}) = \underset{\mathbf{z} \in \mathcal{S}}{\operatorname{argmin}} \{ \|\mathbf{A} - \mathbf{z}\|_2 : \mathbf{z} \in \mathcal{S} \}$$

So, we have:

$$\begin{gathered} \|\mathbf{A} - \mathbf{z}\|_2 = \|A - \mathbf{z}\|_2 = \|P + Q - z\|_2 \\ = tr((\mathbf{Q} + \mathbf{P} - \mathbf{z})^T (\mathbf{Q} + \mathbf{P} - \mathbf{z})) \\ = tr(\mathbf{QQ} + \mathbf{QP} - \mathbf{Qz} - \mathbf{PQ} - \mathbf{PP} + \mathbf{Pz} - \mathbf{zQ} - \mathbf{zP} + \mathbf{zz}) \\ = tr(\mathbf{QQ}) - tr(\mathbf{Qz}) - tr(\mathbf{zQ}) - tr(\mathbf{PP}) + tr(\mathbf{zz}) \\ = tr(\mathbf{QQ}) - 2tr(\mathbf{zQ}) + tr(\mathbf{PP}) + tr(\mathbf{zz}) \\ \text{denoted as } f(\mathbf{z}) \end{gathered}$$

In this $f(\mathbf{z})$, the part whice can be influenced by $\mathbf{z}$ is:

$$-2tr(\mathbf{zQ}) + tr(\mathbf{zz}) = tr[\mathbf{z}(\mathbf{z} - 2\mathbf{Q})] = \sum_{i=1}^{n} \sum_{j=1}^{n} z_{ij}(z_{ij} - 2q_{ij})$$

So, we let:

$$\frac{\partial f}{\partial z_{ij}} = 2z_{ij} - 2q_{ij} = 0$$

So, we have:

$$z_{ij} = q_{ij}$$

Therefore, we have:

$$\mathbf{P}_{\mathcal{S}}(\mathbf{A}) = \mathbf{Q} = (\mathbf{A} + \mathbf{A}^T)/2$$

Exercise 3: Projection to a Function Space

1.

(a)

$L^2(\Omega)$ is a linear space,:

> $\forall X, Y \in L^2(\Omega)$, we have:$\mathbb{E}[(\alpha X + \beta Y)^2] < \infty$

And the projection $\mathbf{P}_\Omega(Y)$ is defined as:

$$\mathbf{P}_\Omega(Y) = \underset{\mathbf{z}\in\Omega}{\operatorname{argmin}}\{\|\mathbf{Y} - \mathbf{z}\|_2 : \mathbf{z} \in \Omega\}$$

We know that

$$\begin{aligned} \|\mathbf{Y} - \mathbf{z}\|_2 = \mathbb{E}(\mathbf{Y} - \mathbf{z})^2 = \mathbb{E}(Y^2 - 2Y\mathbf{z} + \mathbf{z}^2) = \mathbb{E}(Y^2) - 2\mathbb{E}(Y\mathbf{z}) + \mathbb{E}(\mathbf{z}^2) \\ = \mathbb{E}(Y^2) - 2\mathbf{z}\mathbb{E}(Y) + \mathbf{z}^2 \text{denoted as } f(z) \end{aligned}$$

So,we let:

$$\frac{\partial f}{\partial z} = 2(\mathbb{E}(Y) - z) = 0$$

So, we have:

$$z = \mathbb{E}(Y)$$

Therefore, we have:

$$\mathbf{P}_\Omega(Y) = \mathbb{E}(Y)$$

(b)

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

$$\begin{aligned} \mathbb{E}[(Y - c)^2] = \mathbb{E}[Y^2 - 2Y\mathbf{c} + \mathbf{c}^2] = \mathbb{E}(Y^2) - 2\mathbb{E}(Y\mathbf{c}) + \mathbb{E}(\mathbf{c}^2) \\ = \mathbb{E}(Y^2) - 2\mathbf{c}\mathbb{E}(Y) + \mathbf{c}^2 \text{denoted as } f(c) \end{aligned}$$

So,we let:

$$\frac{\partial f}{\partial c} = 2(\mathbb{E}(Y) - c) = 0$$

So, we have:

$$c = \mathbb{E}(Y)$$

Therefore, we have:

$$\hat{c} = \mathbb{E}(Y)$$

(c)

$$\begin{aligned}\mathbb{E}[(Y-c)^2] = \mathbb{E}[Y^2 - 2Y\mathbf{c} + \mathbf{c}^2] = \mathbb{E}(Y^2) - 2\mathbb{E}(Y\mathbf{c}) + \mathbb{E}(\mathbf{c}^2) \\ = \mathbb{E}(Y^2) - 2\mathbf{c}\mathbb{E}(Y) + \mathbf{c}^2 \text{denoted as } f(c)\end{aligned}$$

Let $c = \mathbb{E}(Y)$,we have:

$$\frac{\partial f}{\partial c} = 2(\mathbb{E}(Y) - c) = 0$$

So, we have$\mathbb{E}[(Y-c)^2]$ get its minimum value at $\mathbb{E}(Y)$.

$$\mathbb{E}[(Y-c)^2]_{min} = \mathbb{E}(Y^2)$$

.

The necessary and sufficient condition is $c = \mathbb{E}(Y)$.

Let we treat $\mathbf{c}$ and $\mathbf{Y}$ as vectors, we have $(Y-c)^2$ is the distance between them.
And the minimum distance between any two vectors is $\mathbb{E}[(Y-c)^2]_{min} = \mathbb{E}(Y^2)$,if and only if $c = \mathbb{E}(Y)$,that is $\mathbf{c}$ is the projection of $\mathbf{Y}$ onto Ω.

2.

(a)

$$\mathbb{E}[(f(X)-Y)^2] = \iint (y - f(x))^2 p(x,y) dxdy$$

So,the solution of (a) is the same to:

$$\min_f \left\{ J[f] := \iint (y - f(x))^2 p(x,y) dxdy \right\}$$

Let:

$$\frac{\partial J[f]}{\partial f} = \lim_{\epsilon -> 0} \frac{J[f + \epsilon h] - J[f]}{\epsilon} = 0$$

We have:

$$\int -2(y - f^*(x)) p(x,y) dy = 0$$

So, we have:

$$f^*(\mathbf{x}) = \frac{\int yp(\mathbf{x},y)dy}{\int p(\mathbf{x},y)dy} = \int y\frac{p(\mathbf{x},y)}{p(\mathbf{x})}dy = \int yp(y|\mathbf{x})dy = \mathbb{E}[Y|\mathbf{X}]$$

So,the solution is $f(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}]$

(b)

And the projection $\mathbf{P}_{\mathcal{C}}(Y)$ is defined as:

$$\begin{aligned}\mathbf{P}_{\mathcal{C}}(Y) &= \underset{f(\mathbf{X})\in\mathcal{C}}{\operatorname{argmin}}\{\|\mathbf{Y}-f(\mathbf{X})\|_2 : f(\mathbf{X})\in\mathcal{C}\} \\ &= \underset{f(\mathbf{X})\in\mathcal{C}}{\operatorname{argmin}}\{\mathbb{E}[(\mathbf{Y}-f(\mathbf{X}))^2] : f(\mathbf{X})\in\mathcal{C}\}\end{aligned}$$

So,we can assert that the solution of (a) is $\mathbf{P}_{\mathcal{C}}(Y)$,their function is the same.

(c)

The question 1 is the special case of the question 2, where $\mathcal{C}=\Omega$,X is not a random value but a constant value.
The conditional expectation is the projection of Y onto $\mathcal{C}$.

Exercise 4: Multicollinearity

1.

(a)

$\mathbb{E}(\hat{\mathbf{w}}) = \mathbb{E}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}] = \mathbb{E}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\mathbf{w}+\mathbf{e})]$
$= \mathbb{E}[\mathbf{w}] + \mathbb{E}[\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{e}] = \mathbf{w} + \mathbb{E}[\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]\mathbb{E}[\mathbf{e}] = \mathbf{w}$
Here ,we consider that $\mathbf{X}$ is independent of $\mathbf{e}$.
So,we can get $\mathbb{E}(\hat{\mathbf{w}}) = \mathbf{w}$.

(b)

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

2.

$$
\begin{aligned}
\operatorname{MSE}(\hat{\mathbf{w}}) &= \mathbb{E}[\|\hat{\mathbf{w}}-\mathbf{w}\|^2] \\
&= \mathbb{E}[(\hat{\mathbf{w}}-\mathbf{w})^T(\hat{\mathbf{w}}-\mathbf{w})] \\
&= \mathbb{E}[\hat{\mathbf{w}}^T\hat{\mathbf{w}}-\hat{\mathbf{w}}^T\mathbf{w}-\mathbf{w}^T\hat{\mathbf{w}}+\mathbf{w}^T\mathbf{w}] \\
&= \mathbb{E}\{\sum_{i=1}^p \hat{w}_i^2 - 2\sum_{i=1}^p \hat{w}_i w_i + \sum_{i=1}^p w_i^2\} \\
&= \sum_{i=1}^p [\mathbb{E}\hat{w}_i^2 - (\mathbb{E}\hat{w}_i)^2 + (\mathbb{E}\hat{w}_i)^2 + \mathbb{E}w_i^2 - 2\mathbb{E}(\hat{w}_i w_i)] \\
&= \sum_{i=1}^p \operatorname{Var}(\hat{w}_i) + \sum_{i=1}^p (\mathbb{E}\hat{w}_i - \mathbb{E}w_i)^2 \\
&= \sum_{i=1}^p \operatorname{Var}(\hat{w}_i) + \sum_{i=1}^p (\mathbb{E}\hat{w}_i - w_i)^2
\end{aligned}
$$

3.

$$
\begin{aligned}
\operatorname{MSE}(\hat{\mathbf{w}}) &= \mathbb{E}[\|\hat{\mathbf{w}}-\mathbf{w}\|^2] \\
&= \mathbb{E}[(\hat{\mathbf{w}}-\mathbf{w})^T(\hat{\mathbf{w}}-\mathbf{w})] \\
&= \mathbb{E}[[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\mathbf{w}+\mathbf{e})-\mathbf{w}]^T[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\mathbf{w}+\mathbf{e})-\mathbf{w}]] \\
&= \mathbb{E}[[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{e}]^T[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{e}]] \\
&= \mathbb{E}[\mathbf{e}^T\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-2}\mathbf{X}^\top\mathbf{e}] \\
&= \mathbb{E}[\operatorname{tr}(\mathbf{e}^T\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-2}\mathbf{X}^\top\mathbf{e})] \\
&= \operatorname{tr}(\mathrm{E}[\mathbf{e}^T\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-2}\mathbf{X}^\top\mathbf{e}]) \\
&= \operatorname{tr}(\mathrm{E}[\mathbf{e}\mathbf{e}^T\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-2}\mathbf{X}^\top]) \\
&= \operatorname{tr}(\mathrm{E}[\mathbf{e}\mathbf{e}^T]\mathrm{E}[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-2}\mathbf{X}^\top]) \\
&= \operatorname{tr}(\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-2}\mathbf{X}^\top) \\
&= \sigma^2\operatorname{tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-2}\mathbf{X}^\top) \\
&= \sigma^2\operatorname{tr}(\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-2}) \\
&= \sigma^2\operatorname{tr}((\mathbf{X}^\top\mathbf{X})^{-1}) \\
&= \sigma^2\sum_{i=1}^p \frac{1}{\lambda_i}, \text{where } \lambda_1, \lambda_2, \ldots, \lambda_p \text{ are the eigenvalues of } \mathbf{X}^\top\mathbf{X}.
\end{aligned}
$$

> 这里用到的公式或者性质有：二次型$x^TAx = tr(Axx^T)$,以及迹的循环性质

4.

It means that MSE will be infinity，which means that the model is not suitable for the given data.

Exercise 5: Regularized least squares

1.

Let v be any d-dimensional vector.
We have $v^T\mathbf{X}^\top\mathbf{X}v = (\mathbf{X}v)^T(\mathbf{X}v)$ is a 2-norm of vector $\mathbf{X}v$,so it is absolutely ≥ 0.
So is $\mathbf{X}^\top\mathbf{X}$ always definite semi-definite.
If we have $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d$ is linear independent,then $(\mathbf{X}v) = \mathbf{0}$ if and only if $v = \mathbf{0}$.
So,we can say that $\mathbf{X}^T\mathbf{X}$ is always positive definite.

2.

We have $\mathbf{X}^T\mathbf{X}$ is semi-definite,and $\lambda\mathbf{I}$ is positive .
So,Let v be any d-dimensional vector but not a zero vector,we can get :

$$v^T(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})v = v^T\mathbf{X}^T\mathbf{X}v+\lambda v^Tv > 0$$

So,$\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}$ is positive definite,so its eigenvalues are all positive,so it is always invertible.

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

1.

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

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

Let $\mathbf{M}=[\mathbf{A},\mathbf{b},\mathbf{W}]\in\mathbb{R}^{n\times(n+1+N)}, \mathbf{Y}=[\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N]\in\mathbb{R}^{n\times N}$,
And

$$\mathbf{Z_i}=\begin{bmatrix}\mathbf{x}_i^T\\ 1\\ \phi^T(\mathbf{x}_i)\end{bmatrix}\in\mathbb{R}^{n+1+N}, \mathbf{X}=[\mathbf{Z}_1,\mathbf{Z}_2,\ldots,\mathbf{Z}_N]\in\mathbb{R}^{(n+1+N)\times N}$$

So we can write the problem as:

$$\min_{\mathbf{M}} l := \|\mathbf{MX}-\mathbf{Y}\|_f^2+\left\|(\mathbf{M}-[\mathbf{I_n},0]_{n\times(n+1+N)})diag(\sqrt{\lambda_1^{(n)}},\sqrt{\lambda_2^{(1)}},\sqrt{\lambda_3^{(N)}})\right\|_f^2, \text{Let's denote } arg(\sqrt{\lambda_1^{(n)}},\sqrt{\lambda_2^{(1)}},\sqrt{\lambda_3^{(N)}})\text{as } \mathbf{E}$$

Then,we can write the problem as:

$$\min_{\mathbf{M}} l := \|\mathbf{MX}-\mathbf{Y}\|_f^2+\|(\mathbf{M}-[\mathbf{I_n},0])\mathbf{E}\|_f^2$$

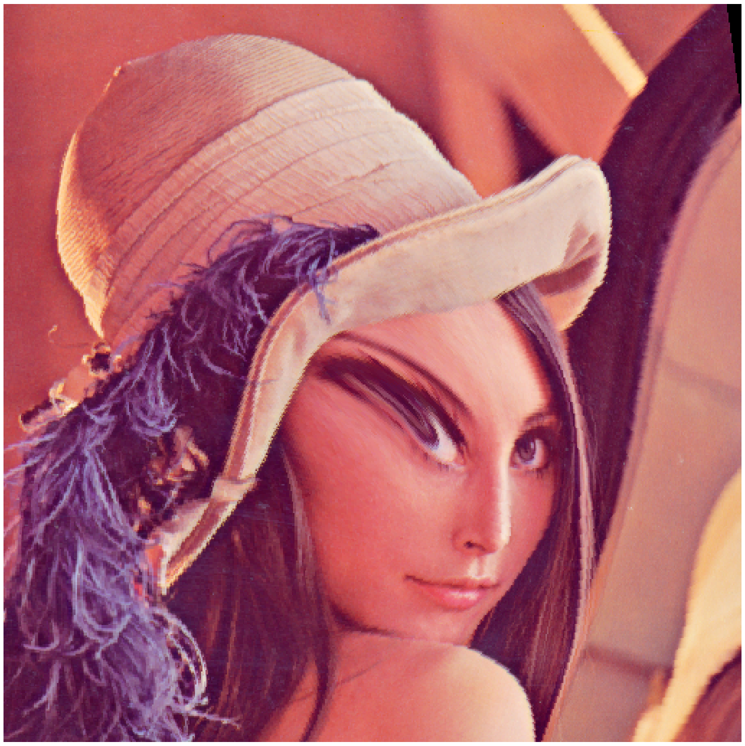
Let:

$$\frac{\partial l}{\partial \mathbf{M}} = 2(\mathbf{MX}-\mathbf{Y})\mathbf{X}^T+2\mathbf{ME^2}-[\mathbf{I_n},0]\mathbf{E^2}=0$$

We can get:

$$\hat{\mathbf{M}} = (\mathbf{YX^T}+[\mathbf{I_n},0]\mathbf{E^2})(\mathbf{XX^T}+\mathbf{E}^2)^{-1}$$

2.



This is a sample image after warping.

Exercise 7: Bias-Variance Trade-off (Programming Exercise)

1.

The design matrix $\Phi(\mathbf{X})^{(l)}$ is an N×25 matrix, where N represents the number of data points and 25 denotes the dimension of $\boldsymbol{\phi}(x) = (1, \phi_1(x), \cdots, \phi_{24}(x))^\top$.
And the column vector $\mathbf{Y}^{(l)}$ is an N-dimensional vector representing the $y_n^{(l)}$
So,the loss function can be written as:

$$L^{(l)}(w) = \frac{1}{2}(Y - \Phi^{(l)}w)^\top (Y - \Phi^{(l)}w) + \frac{\lambda}{2} w^\top w$$

Then,we let:

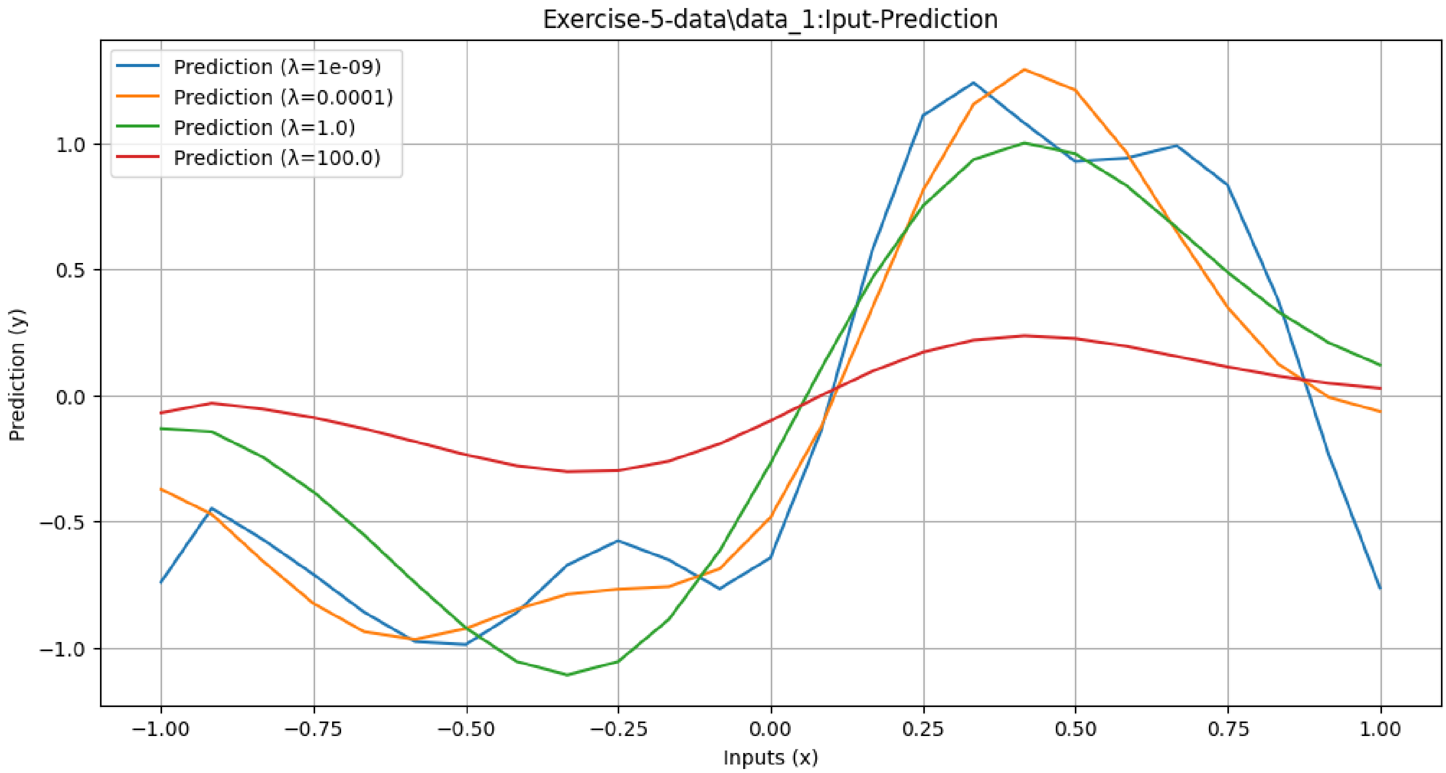
$$\frac{\partial L^{(l)}}{\partial w} = -(\Phi^{(l)})^\top (Y - \Phi^{(l)}w) + \lambda w = 0$$

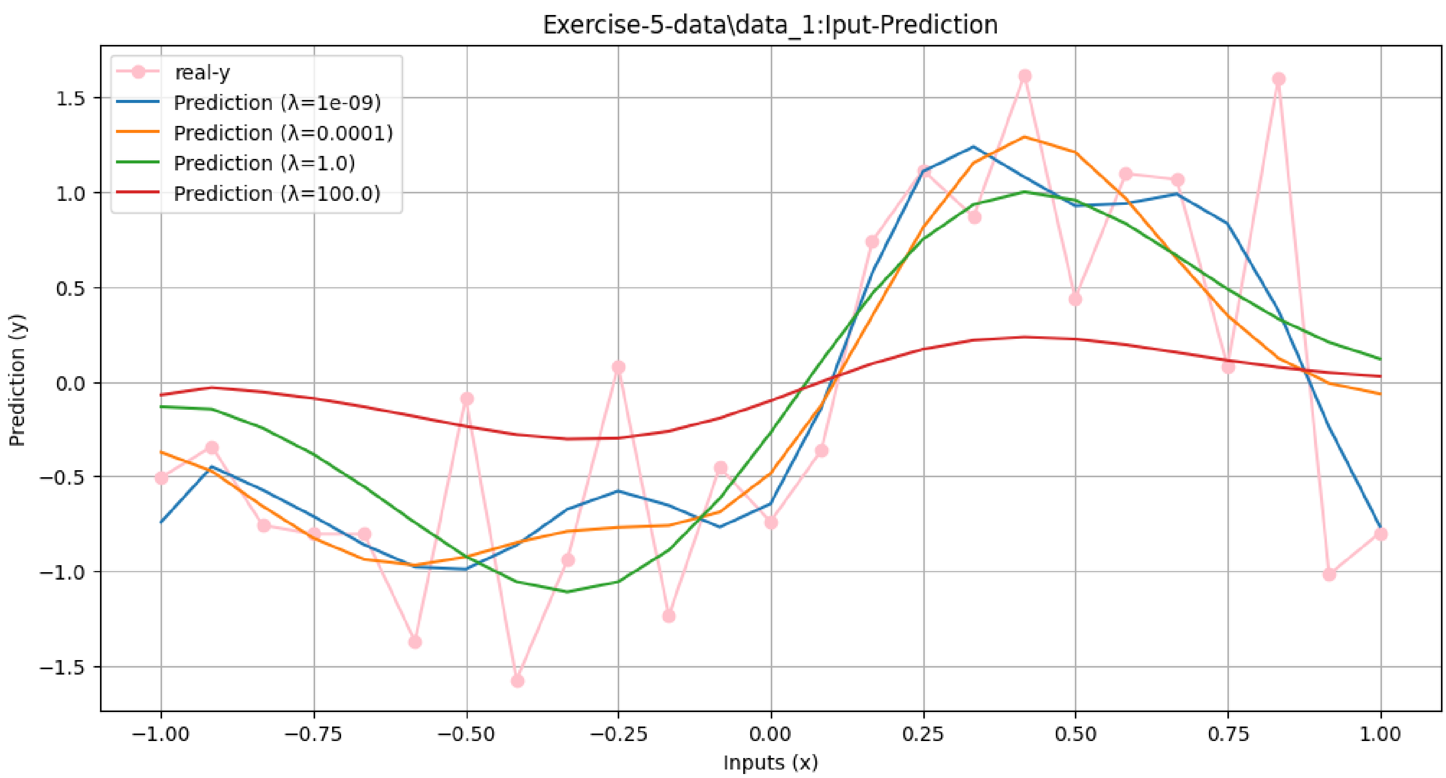
We can get:

$$\hat{w}^{(l)} = ((\Phi^{(l)})^\top \Phi^{(l)} + \lambda I)^{-1} (\Phi^{(l)})^\top Y$$

2.

We can write $f_{\mathcal{D}^{(l)}}(x)$ as $(w^{(l)})^T \boldsymbol{\phi}(x)$
So,$y^{(l)}(x) = f_{\mathcal{D}^{(l)}}(x) = (w^{(l)})^T \boldsymbol{\phi}(x)$
So,$\mathbf{Y}^{(l)}(\mathbf{x}^{(l)}) = f_{\mathcal{D}^{(l)}}(x) = \boldsymbol{\phi}(\mathbf{x})^{(l)}(w^{(l)})$





3.

We have the column vector $\hat{\mathbf{y}}^{(l)}$ is an N-dimensional vector representing the $\hat{y_n}^{(l)}$, so we can let $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}^{(1)}, \hat{\mathbf{y}}^{(1)}, \ldots, \hat{\mathbf{y}}^{(L)}]$
Let $\overline{\mathbf{Y}} = \hat{\mathbf{Y}}$'s column mean,let$\mathbf{H} = [h(x_1), h(x_2), \ldots, h(x_N)]^T$
Then, we can write:

$$(bias)^2 = (\overline{\mathbf{Y}} - \mathbf{H})^T(\overline{\mathbf{Y}} - \mathbf{H})/N$$

$$variance = \|\hat{\mathbf{Y}} - \overline{\mathbf{Y}}\|_f / NL$$

Here, $\overline{\mathbf{Y}}$ has been performed broadcasting along the columns.