Homework 2

Problem 1

We define the feature matrix as:

$$\mathbf{X} = [\cos x \quad \sin x \quad 1] \in \mathbb{R}^{n\times 3},$$

and the parameter vector as:

$$\theta = [\alpha \quad \beta \quad \gamma] \in \mathbb{R}^3.$$

The function can be then rewritten as:

$$\mathbf{y} = \mathbf{X}\theta^\top.$$

To find the optimal estimate of θ (represented as $\hat{\theta}$) as a least square problem, it can be formulated as:

$$\hat{\theta} = \underset{\theta}{\mathbf{argmin}} \left\| \mathbf{y} - \mathbf{X}\theta^\top \right\|_2^2,$$

and according to the lecture note, the closed form of solution is given by

$$\hat{\theta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

Problem 2

1.

Given $\mathbf{z} \sim \mathcal{N}(0, \sigma^2)$, we have $\mathbb{E}(\mathbf{z}) = 0$.

The closed form of solution of $\hat{\theta}$ is given by

$$
\begin{aligned}
\hat{\theta} &= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y} \\
&= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top (\mathbf{y}^* + \mathbf{z}) \\
&= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}^* + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{z} \\
&= \theta^* + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{z},
\end{aligned}
$$

therefore, we can calculate the expectation of $\hat{\theta}$ as:

$$
\begin{aligned}
\mathbb{E}(\hat{\theta}) &= \mathbb{E}(\theta^* + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{z}) \\
&= \theta^* + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbb{E}(\mathbf{z}),
\end{aligned}
$$

recall that we have $\mathbb{E}(\mathbf{z}) = 0$, which leads to

$$
\mathbb{E}(\hat{\theta}) = \theta^*.
$$

To further prove the next equation, firstly utilize the given hint to reformulate the left side of the equation

$$
\begin{aligned}
\mathbb{E}(\left\|\mathbf{y}^* - \mathbf{X}\hat{\theta}\right\|_2^2) &= \mathbb{E}(\left\|\mathbf{y}^* - \mathbf{X}\hat{\theta} + \mathbf{X}\theta^* - \mathbf{X}\theta^*\right\|_2^2) \\
&= \mathbb{E}(\left\|(\mathbf{y}^* - \mathbf{X}\theta^*) + (\mathbf{X}\theta^* - \mathbf{X}\hat{\theta})\right\|_2^2) \\
&= \mathbb{E}(\|\mathbf{y}^* - \mathbf{X}\theta^*\|_2^2 + \left\|\mathbf{X}\theta^* - \mathbf{X}\hat{\theta}\right\|_2^2 - 2 < \mathbf{y}^* - \mathbf{X}\theta^*, \mathbf{X}\theta^* - \mathbf{X}\hat{\theta} >) \\
&= \mathbb{E}(\|\mathbf{y}^* - \mathbf{X}\theta^*\|_2^2) + \mathbb{E}(\left\|\mathbf{X}\theta^* - \mathbf{X}\hat{\theta}\right\|_2^2) - 2\mathbb{E}((\mathbf{y}^* - \mathbf{X}\theta^*)^\top \cdot (\mathbf{X}\theta^* - \mathbf{X}\hat{\theta})) \\
&= \mathbb{E}(\|\mathbf{y}^* - \mathbf{X}\theta^*\|_2^2) + \mathbb{E}(\left\|\mathbf{X}\theta^* - \mathbf{X}\hat{\theta}\right\|_2^2) - 2\mathbb{E}(\mathbf{y}^{*\top}\mathbf{X}\theta^* - \mathbf{y}^{*\top}\mathbf{X}\hat{\theta} - \theta^{*\top}\mathbf{X}^\top\mathbf{X}\theta^* + \theta^{*\top}\mathbf{X}^\top\mathbf{X}\hat{\theta}) \\
&= \mathbb{E}(\|\mathbf{y}^* - \mathbf{X}\theta^*\|_2^2) + \mathbb{E}(\left\|\mathbf{X}\theta^* - \mathbf{X}\hat{\theta}\right\|_2^2) - 2(\mathbf{y}^{*\top}\mathbf{X}\theta^* - \mathbf{y}^{*\top}\mathbf{X}\mathbb{E}(\hat{\theta}) - \theta^{*\top}\mathbf{X}^\top\mathbf{X}\theta^* + \theta^{*\top}\mathbf{X}^\top\mathbf{X}\mathbb{E}(\hat{\theta})).
\end{aligned}
$$

Recall that we have already prooven that $\mathbb{E}(\hat{\theta}) = \theta^*$, which makes that all the terms in the 3.rd bracket cancell each other, i.e.,

$$
\begin{aligned}
\mathbb{E}(\left\|\mathbf{y}^* - \mathbf{X}\hat{\theta}\right\|_2^2) &= \mathbb{E}(\|\mathbf{y}^* - \mathbf{X}\theta^*\|_2^2) + \mathbb{E}(\left\|\mathbf{X}\theta^* - \mathbf{X}\hat{\theta}\right\|_2^2) \\
&= \|\mathbf{y}^* - \mathbb{E}(\mathbf{X}\theta^*)\|_2^2 + \mathbb{E}(\left\|\mathbf{X}\hat{\theta} - \mathbf{X}\theta^*\right\|_2^2)
\end{aligned}
$$

2.

Recall that we have derived the closed form of solution of $\hat{\theta}$ as $\hat{\theta} = \theta^* + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{z}$, we can see that as $\mathbf{z} \sim \mathcal{N}(0, \sigma^2)$, therefore $\hat{\theta}$ is also gaussian distributed, where the mean is by θ^* shifted.

Then we only have to focus on the second term in $\hat{\theta}$, i.e., $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{z}$. According to the task, given $\mathbf{v} \sim \mathcal{N}(0, \mathbf{\Sigma})$, then $\mathbf{A}\mathbf{v} \sim (0, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top)$. Let $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ and $\mathbf{\Sigma} = \sigma^2$, then we have for the variance

$$\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)^\top = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

Therefore, $\hat{\theta} \sim \mathcal{N}(\theta^*, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$

3.

For the given equation:

$$\begin{aligned}\frac{1}{n}\mathbb{E}\left[\left\|\mathbf{X}\hat{\theta} - \mathbf{X}\theta^*\right\|_2^2\right] &= \sigma^2\frac{d}{n}\\ \mathbb{E}\left[\left\|\mathbf{X}\hat{\theta} - \mathbf{X}\theta^*\right\|_2^2\right] &= \sigma^2 d.\end{aligned}$$

Now, let's focus on the left side of the equation

$$\begin{aligned}\text{left side} &= \mathbb{E}\left[\left\langle\mathbf{X}\hat{\theta} - \mathbf{X}\theta^*, \mathbf{X}\hat{\theta} - \mathbf{X}\theta^*\right\rangle\right]\\ &= \mathbb{E}\left[\left(\mathbf{X}\left(\hat{\theta} - \theta^*\right)\right)^\top\left(\mathbf{X}\left(\hat{\theta} - \theta^*\right)\right)\right]\\ &= \mathbb{E}\left[\left(\hat{\theta} - \theta^*\right)^\top\mathbf{X}^\top\mathbf{X}\left(\hat{\theta} - \theta^*\right)\right].\end{aligned}$$

We know that the result of a inner product has to be a scalar, thus we have for the terms inside the expectation operator

$$\left(\hat{\theta} - \theta^*\right)^\top\mathbf{X}^\top\mathbf{X}\left(\hat{\theta} - \theta^*\right) = \operatorname{trace}\left(\left(\hat{\theta} - \theta^*\right)^\top\mathbf{X}^\top\mathbf{X}\left(\hat{\theta} - \theta^*\right)\right),$$

and therefore the expectation value can be calculated with the help of trace

$$\mathbb{E}\left[\left(\hat{\theta}-\theta^*\right)^\top \mathbf{X}^\top\mathbf{X}\left(\hat{\theta}-\theta^*\right)\right]=\mathbb{E}\left[\operatorname{trace}\left(\left(\hat{\theta}-\theta^*\right)^\top \mathbf{X}^\top\mathbf{X}\left(\hat{\theta}-\theta^*\right)\right)\right].$$

Given $\mathbb{E}\left[\operatorname{trace}\left(\mathbf{ABC}\right)\right]=\mathbb{E}\left[\operatorname{trace}\left(\mathbf{CAB}\right)\right]$, let $\mathbf{A}=\left(\hat{\theta}-\theta^*\right)^\top$, $\mathbf{B}=\mathbf{X}^\top\mathbf{X}$, and $\mathbf{C}=\left(\hat{\theta}-\theta^*\right)$, we can re-formulate the right side of equation above as

$$\mathbb{E}\left[\operatorname{trace}\left(\left(\hat{\theta}-\theta^*\right)^\top \mathbf{X}^\top\mathbf{X}\left(\hat{\theta}-\theta^*\right)\right)\right]=\mathbb{E}\left[\operatorname{trace}\left(\left(\hat{\theta}-\theta^*\right)\left(\hat{\theta}-\theta^*\right)^\top \mathbf{X}^\top\mathbf{X}\right)\right]$$

and according to $\mathbb{E}\left[\operatorname{trace}\left(\mathbf{A}\right)\right]=\operatorname{trace}\left(\mathbb{E}\left[\mathbf{A}\right]\right)$, the equation can be further rewritten as:

$$\begin{aligned}\mathbb{E}\left[\operatorname{trace}\left(\left(\hat{\theta}-\theta^*\right)\left(\hat{\theta}-\theta^*\right)^\top \mathbf{X}^\top\mathbf{X}\right)\right]&=\operatorname{trace}\left(\mathbb{E}\left[\left(\hat{\theta}-\theta^*\right)\left(\hat{\theta}-\theta^*\right)^\top \mathbf{X}^\top\mathbf{X}\right]\right)\\&=\operatorname{trace}\left(\mathbb{E}\left[\left(\hat{\theta}-\theta^*\right)\left(\hat{\theta}-\theta^*\right)^\top\right]\mathbf{X}^\top\mathbf{X}\right)\end{aligned}$$

we can see that the term $\mathbb{E}\left[\left(\hat{\theta}-\theta^*\right)\left(\hat{\theta}-\theta^*\right)^\top\right]$ is the covariance matrix of $\hat{\theta}$ since that $\mathbb{E}\left[\hat{\theta}\right]=\theta^*$, and we also had $\hat{\theta}\sim\mathcal{N}\left(\theta^*,\sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right)$, therefore, $\mathbb{E}\left[\left(\hat{\theta}-\theta^*\right)\left(\hat{\theta}-\theta^*\right)^\top\right]=\sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$. So for the right side of the equation above we have:

$$\begin{aligned}\operatorname{trace}\left(\mathbb{E}\left[\left(\hat{\theta}-\theta^*\right)\left(\hat{\theta}-\theta^*\right)^\top\right]\mathbf{X}^\top\mathbf{X}\right)&=\operatorname{trace}\left(\sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\right)\\&=\sigma^2\operatorname{trace}\left(\mathbf{I}\right)\\&=\sigma^2 d.\end{aligned}$$

It is proven that

$$\mathbb{E}\left[\left\|\mathbf{X}\hat{\theta}-\mathbf{X}\theta^*\right\|_2^2\right]=\mathbb{E}\left[\left\langle\mathbf{X}\hat{\theta}-\mathbf{X}\theta^*,\mathbf{X}\hat{\theta}-\mathbf{X}\theta^*\right\rangle\right]=\sigma^2 d,$$

therefore

$$\frac{1}{n}\mathbb{E}\left[\left\|\mathbf{X}\hat{\theta}-\mathbf{X}\theta^*\right\|_2^2\right]=\sigma^2\frac{d}{n}.$$

4.

Given assumption that the underlying function is linear, i.e., all the terms with order higher than 1 in the feature vector should be multiplied by 0 in the parameter vector, which gives us the parameter vector θ as $\theta = [\theta_0 \quad \theta_1 \quad \mathbf{0}_{D-1}] \in \mathbb{R}^{D+1}$.

And $\|\mathbf{y}^* - \mathbf{X}\theta^*\|_2 = 0$.

For this subtask we have $d = D + 1$, and according to previous task, where we proven that $\frac{1}{n}\mathbb{E}\left[\left\|\mathbf{X}\hat{\theta} - \mathbf{X}\theta^*\right\|_2^2\right] = \sigma^2\frac{d}{n}$, we have

$$\frac{1}{n}\mathbb{E}\left[\left\|\mathbf{y}^* - \mathbf{X}\hat{\theta}\right\|_2^2\right] = \sigma^2\frac{D+1}{n}.$$

$\sigma^2\frac{D+1}{n}$ should be upper bounded by ϵ, which means

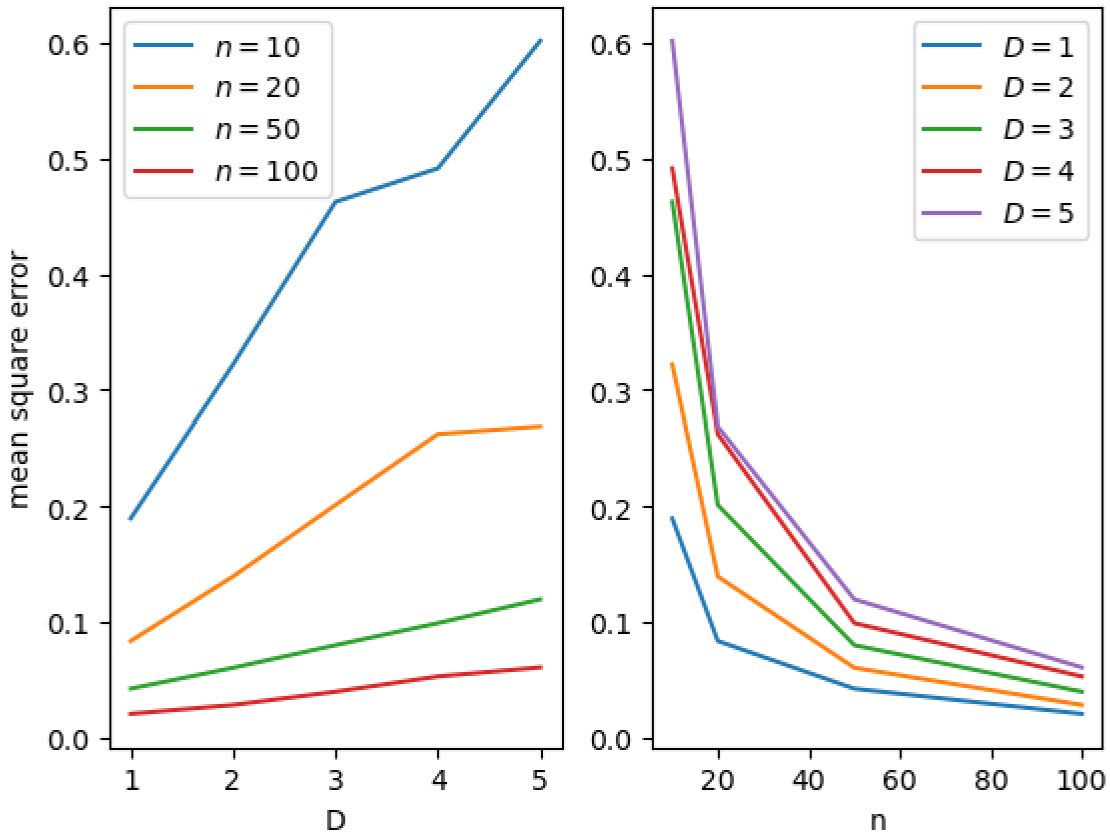
$$\sigma^2\frac{D+1}{n} \leq \epsilon \qquad \Rightarrow \qquad n \geq \sigma^2\frac{D+1}{\epsilon}$$

5.

In [ ]:
```
import numpy as np
from matplotlib import pyplot as plt

def pred_error(n, D):
  alpha = np.random.uniform(-1, 1, n)
  y_star = alpha + 1

  z = np.random.normal(0, 1, n)
  y = y_star + z

  # use numpy.polyfit to get the coefficients of the fitted polynomial
  params = np.polyfit(alpha, y, D)
  # use numpy.poly1d to generate the fitted polynomial
  poly = np.poly1d(params)

  pred_error = ((y_star - poly(alpha))**2).mean()

  return pred_error
```

```python
ns = np.array([10, 20, 50, 100])
Ds = np.array([1, 2, 3, 4, 5])

error = np.zeros((ns.size, Ds.size))

# iterate over all combinations of n and D, for each combination calculate 10 times and average the error
for i in range(0, ns.size):
  for j in range(0, Ds.size):
    for k in range (100):
      error[i, j] += pred_error(ns[i], Ds[j])/100

plt.subplot(1, 2, 1)
for i in range(0, ns.size):
  plt.plot(Ds, error[i, :], label='$n=%d$' % ns[i])
plt.title("")
plt.xlabel("D")
plt.ylabel("mean square error")
plt.legend()

plt.subplot(1, 2, 2)
for i in range(0, Ds.size):
  plt.plot(ns, error[:, i], label='$D=%d$' % Ds[i])
plt.title("")
plt.xlabel("n")
plt.legend()
```

Out[ ]: <matplotlib.legend.Legend at 0xffff87721870>



mean square error
D
n
$n = 10$
$n = 20$
$n = 50$
$n = 100$
$D = 1$
$D = 2$
$D = 3$
$D = 4$
$D = 5$