Homework 5

Xixiang Chen

1 Hoeffding's Inequality

(a)

Let $X = X_1 + X_2 + ... + X_n$, and $\mu_X = \mu_{X_1} + \mu_{X_2} + ... + \mu_{X_n}$.

$$\begin{aligned}
Pr[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_{X_i}) \geq t] &= Pr[\sum_{i=1}^{n}(X_i - \mu_{X_i}) \geq nt] \\
&= Pr[X - \mu_X \geq nt] \\
&\leq \min_{\lambda \geq 0} M_{X-\mu_X}(\lambda)e^{-\lambda nt} \\
&\leq e^{\lambda nt}E[e^{\lambda(X-\mu_X)}] \\
&\leq e^{\lambda nt}\prod_{i=1}^{n}E[e^{\lambda(X_i-\mu_{X_i})}] \\
&\leq e^{\lambda nt}\exp(\frac{\lambda^2 n(b-a)^2}{8}) \\
&= \exp(-\lambda nt + \frac{\lambda^2 n(b-a)^2}{8})
\end{aligned}$$

Since the exponent is a quadratic function of λ, it has global minimum when its derivative is zero. Take derivative of the function $f(\lambda) = -\lambda nt + \frac{\lambda^2 n(b-a)^2}{8}$ to get:

$$\lambda = \frac{4t}{(b-a)^2}$$

And substitute back:

$$\begin{aligned}
\exp(-\lambda nt + \frac{\lambda^2 n(b-a)^2}{8}) &= \exp(-\frac{4nt^2}{(b-a)^2} + \frac{2nt^2}{(b-a)^2}) \\
&= \exp(-\frac{2nt^2}{(b-a)^2})
\end{aligned}$$

Therefore:

$$Pr[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_{X_i}) \geq t] \leq \exp(-\frac{2nt^2}{(b-a)^2})$$

(b)

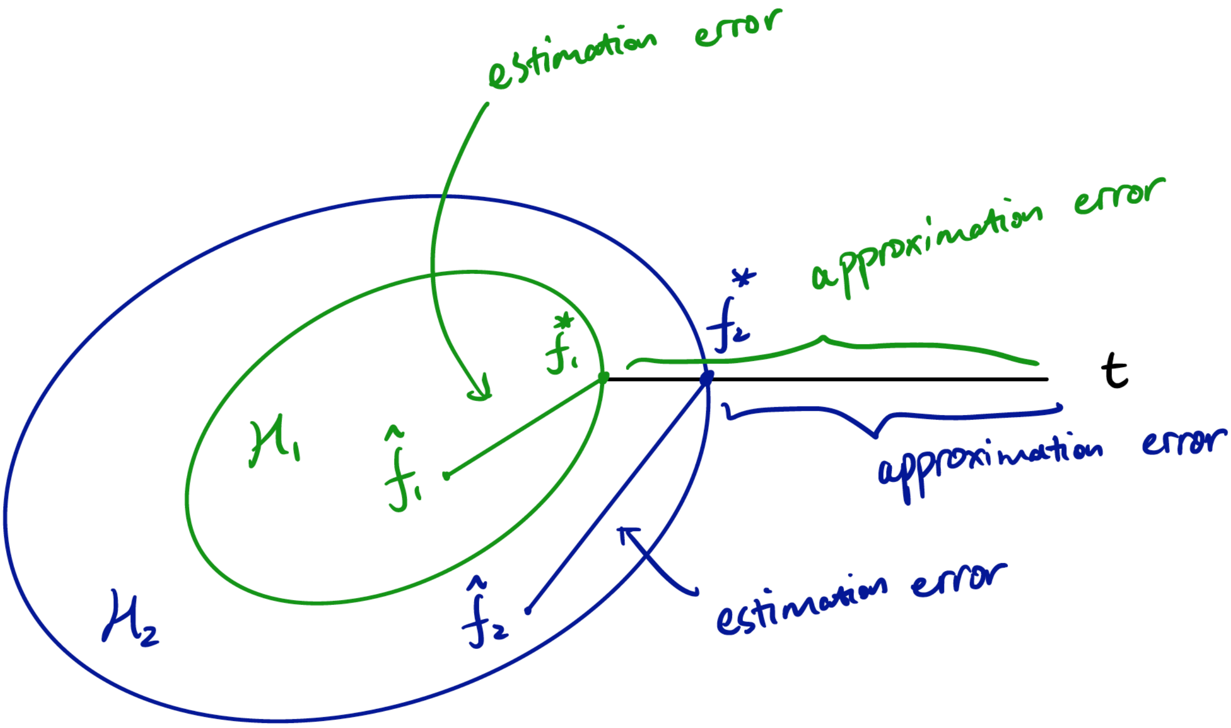
Heoffding's Bound depends on interval [a,b]. When (b-a) is large and the random variable X_i is concentrated around $\frac{b-a}{2}$, say $Pr[X_i = \frac{b-a}{2}] = 0.99$ and $Pr[X_i = b \ or \ X_i = a] = 0.01$, the bound can be much sharper than the equation above.

2 VC Dimension

(a)

VC-dimension for $\mathcal{H}_1$ is p.
VC-dimension for $\mathcal{H}_2$ is $\frac{(p+1)(p+2)}{2}$, which is equal to the number of parameters.

(b)



$$AR = R^{true}(f^*) - R^*$$

$$ER = R^{true}(\hat{f}) - R^{true}(f^*) \leq 2\sqrt{\frac{\log(N) + \log(\frac{2}{\delta})}{2n}}$$

As n increases, the estimation error decreases. Approximation error stays the same since it doesn't depend on the dataset.

(c)

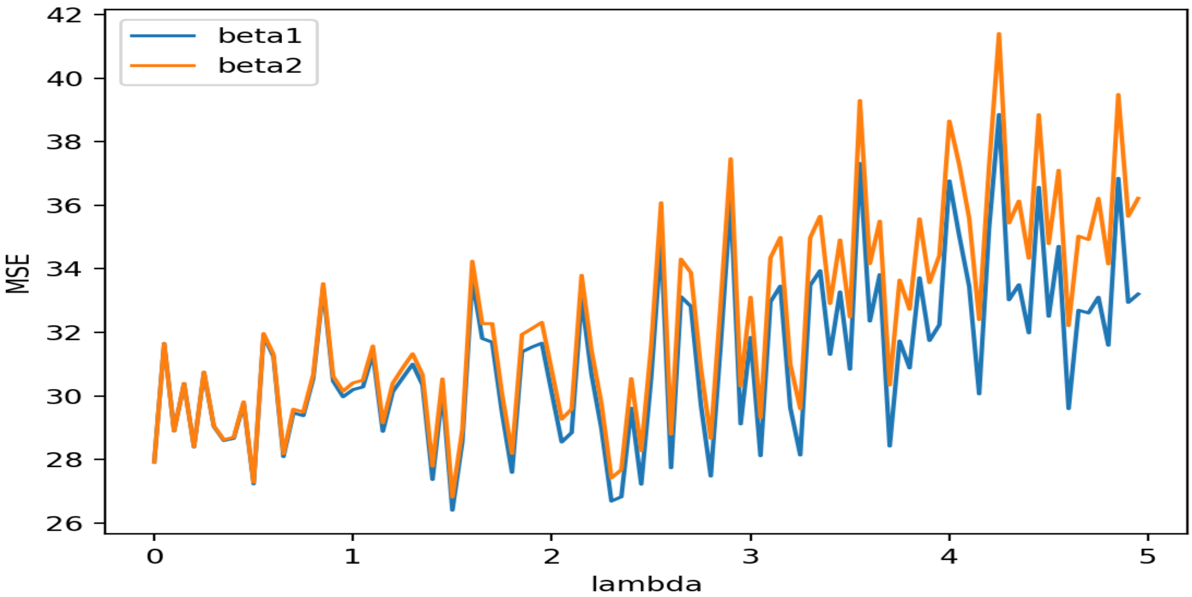
$\mathcal{F} = (f : f(x) = Sign(\sin\theta x))$. Number of parameter is 1. VC-dimension is infinity.

3 Ridge Regression

(a)

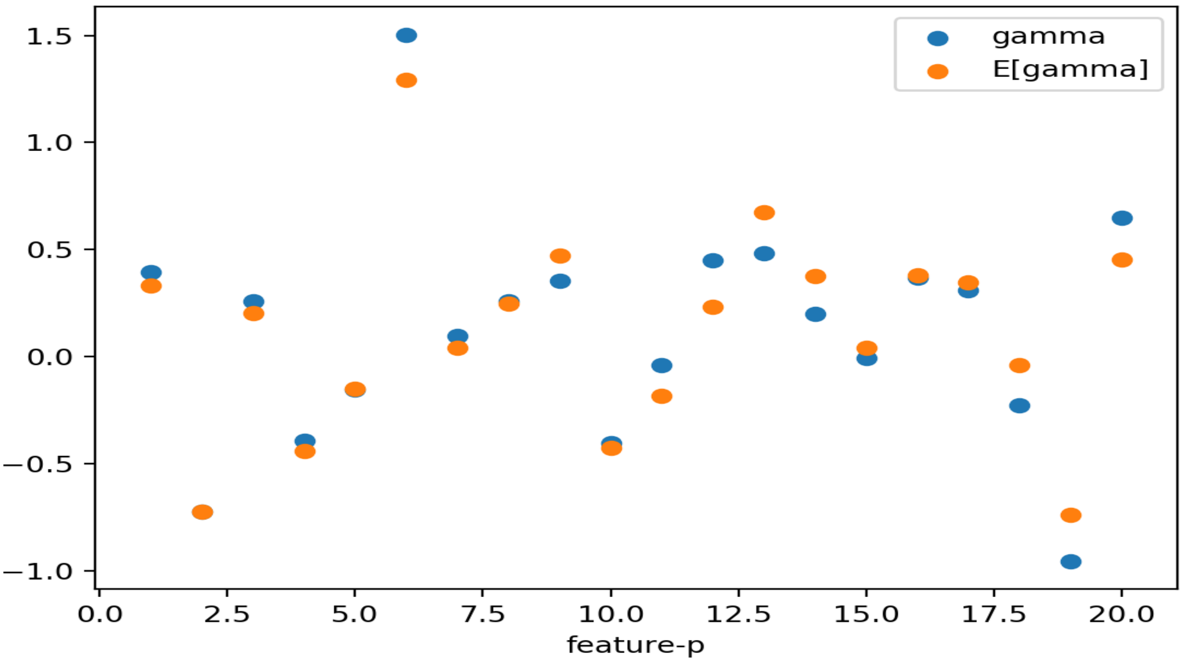
$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T Y$

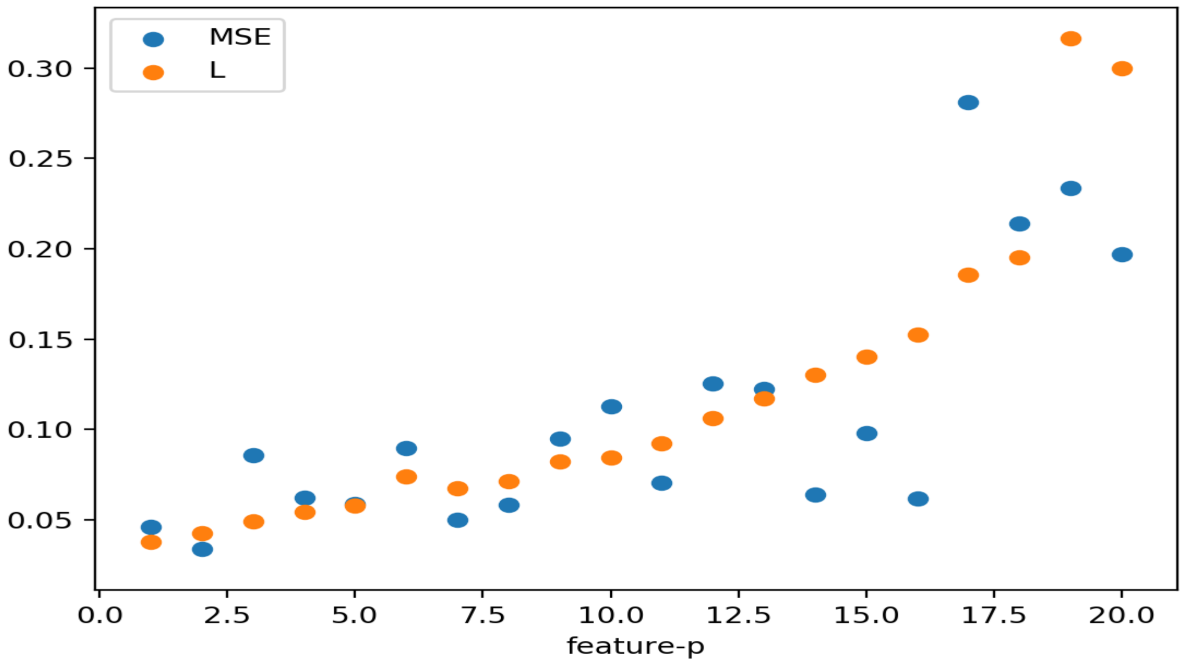
(b)



(c)

$$
\begin{aligned}
\hat{\gamma} &= V^T(X^TX + \lambda I)^{-1}X^TY \\
&= V^T(U_pLV^T)^TU_pLV^T + \lambda VV^T)^{-1}(U_pLV^T)^TY \\
&= V^T(VL^2V^T + \lambda VV^T)^{-1}VLU_p^TY \\
&= V^T(V(L^2 + \lambda I)V^T)^{-1}VLU_p^TY \\
&= V^TV(L^2 + \lambda I)^{-1}V^TVLU_p^TY \\
&= (L^2 + \lambda I)^{-1}LU_p^TY
\end{aligned}
$$


gamma
E[gamma]
1.5
1.0
0.5
0.0
−0.5
−1.0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
feature-p


MSE
L
0.30
0.25
0.20
0.15
0.10
0.05
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
feature-p