1 Hoeffding's Inequality

a. Use Chernoff bounds and Hoeffding's lemma to prove Hoeffding's inequality:

$$P_r(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_{X_i})\geq t)=P_r(\sum_{i=1}^{n}(X_i-\mu_{X_i})\geq nt)=\ P_r(e^{\lambda\sum_{i=1}^{n}(X_i-\mu_{X_i})}\geq e^{\lambda nt})$$

$$\leq \min_{\lambda\geq 0} e^{-\lambda nt}\mathbb{E}[e^{\lambda\sum_{i=1}^{n}(X_i-\mu_{X_i})}]=\min_{\lambda\geq 0} e^{-\lambda nt}\prod_{i=1}^{n}\mathbb{E}[e^{\lambda(X_i-\mu_{X_i})}]$$

$$\leq \min_{\lambda\geq 0} e^{-\lambda nt}\prod_{i=1}^{n}e^{\frac{\lambda^2(b-a)^2}{8}}=\lim_{\lambda\geq 0} e^{-\lambda nt+\frac{n\lambda^2(b-a)^2}{8}}$$

Let $\mathrm{g}(\lambda)=-\lambda nt+\frac{n\lambda^2(b-a)^2}{8}$. It's a quadratic function and achieves its minimum at $\frac{4t}{(b-a)^2}$.

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

b. Give a simple distribution of X_i where the bound can be much sharper then Hoeffding's bound.

$X_1,\ldots,X_n$ are i.i.d from a distribution with mean zero, bounded support [a, b], with variance $\mathbb{E}\,[X^2]=\sigma^2$. Then,

$$P_r(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_{X_i})\geq t)\leq exp(-\frac{nt^2}{2(\sigma^2+(b-a)t)}).$$

This equality is typically known as Bernstein's inequality. Since $\sigma\leq b-a$, when σ is small, this bound can be much sharper than Hoeffding's bound.

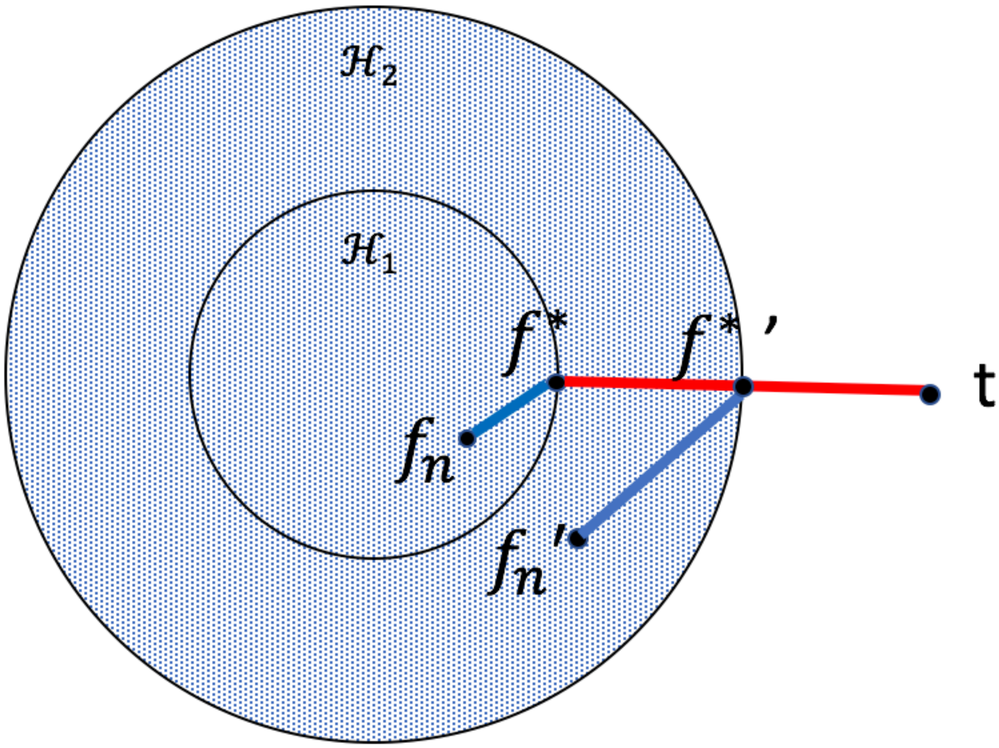
2 VC Dimension

a. What is the VC-dimension of $\mathcal{H}_1$and $\mathcal{H}_2$.

The VC-dimension of $\mathcal{H}_1$ is p. Because the VC-dimension of half-spaces in p dimensions is p+1.

The VC-dimension of $\mathcal{H}_2$ is $\binom{p+2}{2}$. Because the feature space $\boldsymbol{\phi}(\mathbf{x})$ of the second polynomial kernel is of degree $\binom{p+2}{2} = \frac{(p+2)(p+1)}{2}$.

b. Draw a picture for the approximation and estimation error for $\mathcal{H}_1$, $\mathcal{H}_2$ and $\hat{f}_1$, $\hat{f}_2$ and write them down. Explain how the two errors change as n increases.



The blue lines are estimation error and the red lines are approximation error.

Approximation error for $\mathcal{H}_1$ and $\hat{f}_1$ is:

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

Estimation error for $\mathcal{H}_1$ and $\hat{f}_1$ is:

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

Approximation error for $\mathcal{H}_2$ and $\hat{f}_2$ is:

$$R^{true}(f^{*\prime}) - R^*$$

Estimation error for $\mathcal{H}_2$ and $\hat{f}_2$ is:

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

The estimation error decreases and the approximation error doesn't change as n increases.

c. The VC-dimension of the set of sin functions with arbitrarily large or small frequency is infinite. But the number of parameters of sin(ax) is only one.

3 Ridge Regression

a. Derive the closed form solution of $\hat{\beta}^{ridge}$

$$\mathrm{F}(\vec{\lambda}) = \left\|\vec{y} - \bar{\bar{X}}\vec{\beta}\right\|_2^2 + \lambda\left\|\vec{\beta}\right\|_2^2$$

$$\mathrm{F}(\vec{\lambda}) = -2\bar{\bar{X}}^{\mathrm{T}}(\vec{y} - \bar{\bar{X}}\vec{\beta}) + 2\lambda\vec{\beta} = 2(-2\bar{\bar{X}}^{\mathrm{T}}\vec{y} + \bar{\bar{X}}^{\mathrm{T}}\bar{\bar{X}}\vec{\beta} + \lambda\vec{\beta}) = 0$$

$$\bar{\bar{X}}^{\mathrm{T}}\vec{y} = (\bar{\bar{X}}^{\mathrm{T}}\bar{\bar{X}} + \lambda\bar{\bar{\mathrm{I}}})\vec{\beta}^{*}$$

$$\vec{\beta}^{*} = (\bar{\bar{X}}^{\mathrm{T}}\bar{\bar{X}} + \lambda\bar{\bar{I}})^{-1}\bar{\bar{X}}^{\mathrm{T}}\vec{y}$$

b.

Since we know

$$\begin{pmatrix} \mathbf{0} \\ \boldsymbol{L\gamma} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_n{}^T \\ \boldsymbol{U}_p{}^T \\ \boldsymbol{U}_{n-p-1}{}^T \end{pmatrix} \mathrm{U}_p\boldsymbol{LV}^T\boldsymbol{\beta}$$

we have

$$\boldsymbol{L\gamma} = \boldsymbol{U}_p{}^T\mathrm{U}_p\boldsymbol{LV}^T\boldsymbol{\beta}$$

So,

$$\boldsymbol{\gamma} = \boldsymbol{V}^T\boldsymbol{\beta}$$

In the original ridge regression, we have

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

If we plug this equation into the equation above, we have

$$\hat{\boldsymbol{\gamma}} = \boldsymbol{V}^T(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

$$= \boldsymbol{V}^T\left(\left(\mathbf{U}_p\boldsymbol{L}\boldsymbol{V}^T\right)^T\mathbf{U}_p\boldsymbol{L}\boldsymbol{V}^T + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

$$= \boldsymbol{V}^T\left(\boldsymbol{V}\boldsymbol{L}\boldsymbol{U}_p{}^T\mathbf{U}_p\boldsymbol{L}\boldsymbol{V}^T + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

$$= \boldsymbol{V}^T(\boldsymbol{V}\boldsymbol{L}^2\boldsymbol{V}^T + \lambda\boldsymbol{I})^{-1}\left(\mathbf{U}_p\boldsymbol{L}\boldsymbol{V}^T\right)^T\boldsymbol{Y}$$

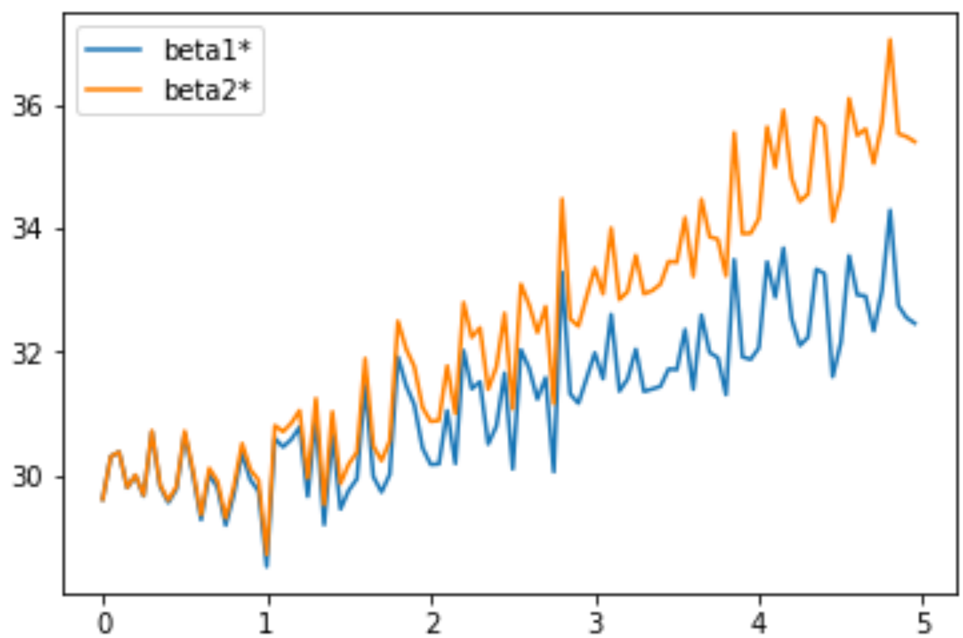
$$= \boldsymbol{V}^T(\boldsymbol{V}\boldsymbol{L}^2\boldsymbol{V}^T + \lambda\boldsymbol{V}\boldsymbol{V}^T)^{-1}\boldsymbol{V}\boldsymbol{L}\boldsymbol{U}_p{}^T\boldsymbol{Y}$$

$$= \boldsymbol{V}^T(\boldsymbol{V}\boldsymbol{L}^2\boldsymbol{V}^T + \boldsymbol{V}\lambda\boldsymbol{V}^T)^{-1}\boldsymbol{V}\boldsymbol{L}\boldsymbol{U}_p{}^T\boldsymbol{Y}$$

$$= \boldsymbol{V}^T\boldsymbol{V}(\boldsymbol{L}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{L}\boldsymbol{U}_p{}^T\boldsymbol{Y}$$

$$= (\boldsymbol{L}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{L}\boldsymbol{U}_p{}^T\boldsymbol{Y}$$

So, the closed form of $\hat{\boldsymbol{\gamma}}$ is $(\boldsymbol{L}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{L}\boldsymbol{U}_p{}^T\boldsymbol{Y}$**.**



For large λ, penalty dominates the loss function. If λ is big, the sum of squares of the coefficients must be small. So, the MSE of both β_1 and β_2 increases as λ increases.

The sum of squares of β_1 is 5.64 and the sum of squares of β_2 is 10.72. So, ridge regression forces β_2 to change more than β_2, and that's why β_2 becomes more different from the original $\beta_2{}^*$. So, the difference between the MSE of β_2 and β_1

increases as λ increases.

c.