

# ECE 595: Homework 6

Yi Qiao, Class ID 187

(Spring 2019)

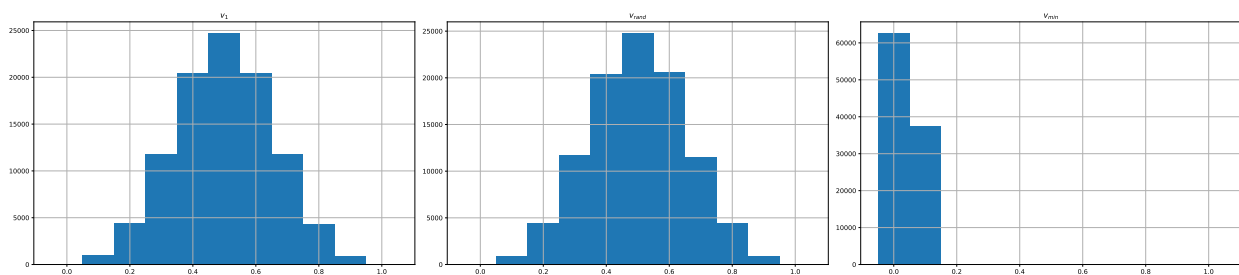
## Exercise 1: Hoeffding Inequality

(a) probability of getting a head for coins  $c_1$ ,  $c_{rand}$  and  $c_{min}$

Since they are all fair coins,

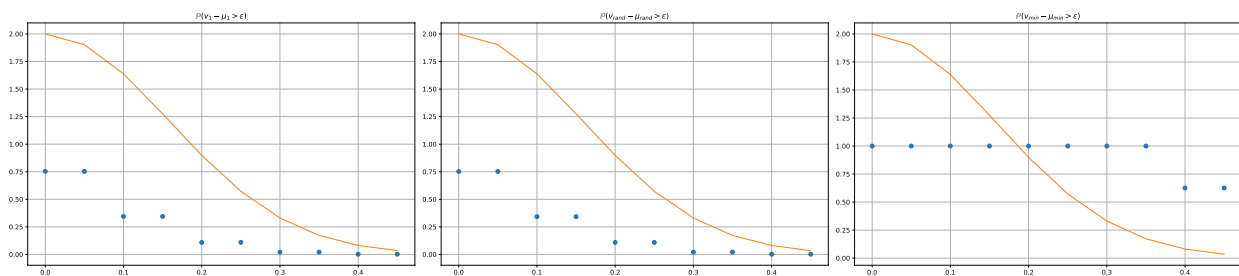
$$\mathbb{P}(c_1 = \text{head}) = \mathbb{P}(c_{rand} = \text{head}) = \mathbb{P}(c_{min} = \text{head}) = 0.5$$

(b) python experiment



The first two are binominal, while the third one is not.

(c) plots



(d)

By observation, we can see that the first two obviously obey the Hoeffding's bound while the third one does not.

(e)

something something...

## Exercise 2: VC Dimension

### (a) Compute the VC dimension

(i)

$$\mathcal{H} = \{h : \mathbb{R} \rightarrow \{-1, +1\} | h(x) = +1, \forall x \in [a, \infty), a \in \mathbb{R}\} \cup \{h : \mathbb{R} \rightarrow \{-1, +1\} | h(x) = +1, \forall x \in (-\infty, a], a \in \mathbb{R}\} \quad (1)$$

By inspection, the VC dimension of the above hypothesis set is 2.

(ii)

$$\mathcal{H} = \{h : \mathbb{R} \rightarrow \{-1, +1\} | h(x) = +1, \forall x \in [a, b], a, b \in \mathbb{R}\} \cup \{h : \mathbb{R} \rightarrow \{-1, +1\} | h(x) = -1, \forall x \in [a, b], a, b \in \mathbb{R}\} \quad (2)$$

By inspection, the VC dimension of the above hypothesis set is 3.

(iii)

$$\mathcal{H} = \left\{ h : \mathbb{R}^d \rightarrow \{-1, +1\} | h(x) = +1, \forall x \text{ where } \sqrt{\sum_{j=1}^d x_j^2} \leq b, b \in \mathbb{R} \right\} \quad (3)$$

By inspection, hypothesis function is a hyper ball. Thus, the VC dimension of the above hypothesis set is 1.

(b)

$$\mathcal{H} = \left\{ h_\alpha : \mathbb{R} \rightarrow \mathbb{R} | h_\alpha(x) = (-1)^{\lfloor \alpha x \rfloor}, \alpha \in \mathbb{R} \right\} \quad (4)$$

Even though the above hypothesis set has only one parameter, it is periodic, thus by tuning the period/frequency, you can match any number of data points you want by finding their GCD. This hypothesis set has simply too large VC dimension, which is far beyond the model complexity. Thus, this will perform far worse than perceptron due to over-fitting.

### Exercise 3: Bias-Variance Trade-off

(a)

$$\begin{aligned}
\boldsymbol{\theta}_{\mathcal{D}} &= \underset{\boldsymbol{\theta}_h}{\operatorname{argmin}} E_{\text{aug}}(h) \\
&= \underset{\boldsymbol{\theta}_h}{\operatorname{argmin}} E_{\text{in}}(h) + \frac{\lambda}{N} \boldsymbol{\theta}_h^T \boldsymbol{\theta}_h \\
&= \underset{\boldsymbol{\theta}_h}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^N (\boldsymbol{\theta}_h^T \mathbf{x}_n - y_n)^2 + \frac{\lambda}{N} \boldsymbol{\theta}_h^T \boldsymbol{\theta}_h \\
&= \underset{\boldsymbol{\theta}_h}{\operatorname{argmin}} \left\| \begin{bmatrix} \mathbf{x}_1^T & 1 \\ \mathbf{x}_2^T & 1 \\ \vdots & \vdots \\ \mathbf{x}_N^T & 1 \end{bmatrix} \boldsymbol{\theta}_h - \mathbf{y} \right\|_2^2 + \lambda \boldsymbol{\theta}_h^T \boldsymbol{\theta}_h
\end{aligned} \tag{5}$$

substitute  $\begin{bmatrix} \mathbf{x}_1^T & 1 \\ \mathbf{x}_2^T & 1 \\ \vdots & \vdots \\ \mathbf{x}_n^T & 1 \end{bmatrix}$  with  $\mathbf{A}$ , we get

$$\boldsymbol{\theta}_{\mathcal{D}} = \underset{\boldsymbol{\theta}_h}{\operatorname{argmin}} \|\mathbf{A}\boldsymbol{\theta}_h - \mathbf{y}\|_2^2 + \lambda \boldsymbol{\theta}_h^T \boldsymbol{\theta}_h \tag{6}$$

Taking the derivative, we can see

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}_h} &= 2\mathbf{A}^T(\mathbf{A}\boldsymbol{\theta}_h - \mathbf{y}) + 2\lambda\boldsymbol{\theta}_h = 0 \\
(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})\boldsymbol{\theta}_h &= \mathbf{A}^T\mathbf{y} \\
\boldsymbol{\theta}_{\mathcal{D}} = \boldsymbol{\theta}_h^* &= (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T(\mathbf{A}\boldsymbol{\theta}_f + \boldsymbol{\epsilon})
\end{aligned} \tag{7}$$

(b)

Continue from the last problem, expand what we already got,

$$\begin{aligned}
\boldsymbol{\theta}_{\mathcal{D}} &= (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T(\mathbf{A}\boldsymbol{\theta}_f + \boldsymbol{\epsilon}) \\
&= (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\mathbf{A}\boldsymbol{\theta}_f + (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\boldsymbol{\epsilon} \\
&= (\mathbf{I} - \lambda(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1})\boldsymbol{\theta}_f + (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\boldsymbol{\epsilon} \\
&= \boldsymbol{\theta}_f - \lambda(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\boldsymbol{\theta}_f + (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\boldsymbol{\epsilon}
\end{aligned} \tag{8}$$

(c)

(i)

$$\begin{aligned}
\bar{g}(\mathbf{x}) &= \mathbb{E}_{\mathcal{D}} [h^{(\mathcal{D})}(\mathbf{x})] = \mathbb{E}_{\mathcal{D}} [\boldsymbol{\theta}_{\mathcal{D}}^T \mathbf{x}] \\
&= \mathbb{E}_{\mathcal{D}} \left[ (\boldsymbol{\theta}_f - \lambda(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\boldsymbol{\theta}_f + (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\boldsymbol{\epsilon})^T \mathbf{x} \right] \\
&= \boldsymbol{\theta}_f^T \mathbf{x} - \lambda \mathbf{x}^T \mathbb{E}_{\mathbf{A}} [(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}] \boldsymbol{\theta}_f + \mathbf{x}^T \mathbb{E}_{\mathbf{A}} [(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1} \mathbf{A}^T] \mathbb{E}[\boldsymbol{\epsilon}] \\
&= \boldsymbol{\theta}_f^T \mathbf{x} - \lambda \mathbf{x}^T \mathbb{E}_{\mathbf{A}} [(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}] \boldsymbol{\theta}_f
\end{aligned} \tag{9}$$

(ii)

$$\begin{aligned}
(\bar{g}(\mathbf{x}) - f(\mathbf{x}))^2 &= (\boldsymbol{\theta}_f^T \mathbf{x} - \lambda \mathbf{x}^T \mathbb{E}_{\mathbf{A}}[(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1}] \boldsymbol{\theta}_f - \boldsymbol{\theta}_f^T \mathbf{x})^2 \\
&= \lambda^2 (\mathbf{x}^T \mathbb{E}_{\mathbf{A}}[(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1}] \boldsymbol{\theta}_f)^T (\mathbf{x}^T \mathbb{E}_{\mathbf{A}}[(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1}] \boldsymbol{\theta}_f) \\
&= \lambda^2 (\boldsymbol{\theta}_f^T \mathbb{E}_{\mathbf{A}}[(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1}]^T \mathbf{x} \mathbf{x}^T \mathbb{E}_{\mathbf{A}}[(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1}] \boldsymbol{\theta}_f) \\
&= \lambda^2 \text{trace}(\mathbf{x} \mathbf{x}^T (\mathbb{E}_{\mathbf{A}}[(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1}] \boldsymbol{\theta}_f \boldsymbol{\theta}_f^T \mathbb{E}_{\mathbf{A}}[(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1}]))
\end{aligned} \tag{10}$$

(iii)

Plug in  $\mathbf{A}^T \mathbf{A} \approx N \mathbf{I}$ , we got

$$\begin{aligned}
\text{bias} &= \mathbb{E}_{\mathcal{X}}[(\bar{g}(\mathbf{x}) - f(\mathbf{x}))^2] \approx \lambda^2 \text{trace}(\mathbb{E}_{\mathcal{X}}[\mathbf{x} \mathbf{x}^T] ((N + \lambda) \mathbf{I})^{-1} \boldsymbol{\theta}_f \boldsymbol{\theta}_f^T ((N + \lambda) \mathbf{I})^{-1}) \\
&= \frac{\lambda^2}{(N + \lambda)^2} \text{trace}(\boldsymbol{\theta}_f \boldsymbol{\theta}_f^T) \\
&= \frac{\lambda^2}{(N + \lambda)^2} \boldsymbol{\theta}_f^T \boldsymbol{\theta}_f \\
&= \frac{\lambda^2}{(N + \lambda)^2} \|\boldsymbol{\theta}_f\|_2^2
\end{aligned} \tag{11}$$

(iv)

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[(h^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2] &= \mathbb{E}_{\mathcal{D}}[(\boldsymbol{\theta}_{\mathcal{D}}^T \mathbf{x} - \boldsymbol{\theta}_f^T \mathbf{x} + \lambda \mathbf{x}^T (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \boldsymbol{\theta}_f)^2] \\
&= \mathbb{E}_{\mathcal{D}}[(\boldsymbol{\theta}_f - \lambda (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \boldsymbol{\theta}_f + (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \boldsymbol{\epsilon})^T \mathbf{x} - \boldsymbol{\theta}_f^T \mathbf{x} + \lambda \mathbf{x}^T (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \boldsymbol{\theta}_f]^2] \\
&= \mathbb{E}_{\mathcal{D}}[(\mathbf{x}^T (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \boldsymbol{\epsilon})^2] \\
&= \mathbb{E}_{\mathbf{A}}[(\mathbf{x}^T (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T)^2] \mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^2] \\
&= \sigma^2 \mathbb{E}_{\mathbf{A}}[(\mathbf{x}^T (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T)^T (\mathbf{x}^T (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T)] \\
&= \sigma^2 \mathbb{E}_{\mathbf{A}}[\text{trace}(\mathbf{A} (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{x} \mathbf{x}^T (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T)] \\
&= \sigma^2 \mathbb{E}_{\mathbf{A}}[\text{trace}(\mathbf{x} \mathbf{x}^T (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{A} (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1})]
\end{aligned} \tag{12}$$

(v)

$$\begin{aligned}
\text{var} &= \mathbb{E}_{\mathcal{X}}[\mathbb{E}_{\mathcal{D}}[(h^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2]] \\
&= \sigma^2 \mathbb{E}_{\mathbf{A}}[\text{trace}(\mathbb{E}_{\mathcal{X}}[\mathbf{x} \mathbf{x}^T] (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{A} (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1})] \\
&= \sigma^2 \mathbb{E}_{\mathbf{A}}[\text{trace}(\mathbf{I} (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{A} (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1})] \\
&= \sigma^2 \mathbb{E}_{\mathbf{A}}[\text{trace}((\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{A} (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{A} (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1})] \\
&= \frac{\sigma^2}{N} \mathbb{E}_{\mathbf{A}}[\text{trace}(\mathbf{A} (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{A} (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T)] \\
&= \frac{\sigma^2}{N} \mathbb{E}_{\mathbf{A}}[\text{trace}(H^2(\lambda))]
\end{aligned} \tag{13}$$