Solutions to Homework 5
CS-526 Learning Theory

Problem 1. VC dimension of union

1. Let $\mathcal{H} = \bigcup_{i=1}^{r} \mathcal{H}_i$. By definition of the growth function we have $\tau_{\mathcal{H}}(m) \leq \sum_{i=1}^{r} \tau_{\mathcal{H}_i}(m)$ for any set of m points. If $k > d+1$ points are shattered by $\mathcal{H}$ then $2^k = \tau_{\mathcal{H}}(k) \leq \sum_{i=1}^{r} \tau_{\mathcal{H}_i}(k) \leq rk^d$, where the last inequality follows directly from Sauer's lemma. Taking the logarithm on both sides and using the inequality yields
$$k \leq \frac{4d}{\log(2)} \log\left(\frac{2d}{\log(2)}\right) + 2\frac{\log(r)}{\log(2)} \; .$$
Note that this inequality is trivially satisfied if $k \leq d+1$.

2. Assume that $k \geq 2d+2$. It is enough to prove that $\tau_{\mathcal{H}_1 \cup \mathcal{H}_2}(k) < 2^k$.
$$\tau_{\mathcal{H}_1 \cup \mathcal{H}_2}(k) \leq \tau_{\mathcal{H}_1}(k) + \tau_{\mathcal{H}_2}(k) \leq \sum_{i=0}^{d} \binom{k}{i} + \sum_{i=0}^{d} \binom{k}{i} =$$
$$= \sum_{i=0}^{d} \binom{k}{i} + \sum_{i=0}^{d} \binom{k}{k-i} = \sum_{i=0}^{d} \binom{k}{i} + \sum_{i=k-d}^{k} \binom{k}{i} \leq$$
$$\leq \sum_{i=0}^{d} \binom{k}{i} + \sum_{i=d+2}^{k} \binom{k}{i} < \sum_{i=0}^{d} \binom{k}{i} + \sum_{i=d+1}^{k} \binom{k}{i} =$$
$$= \sum_{i=0}^{k} \binom{k}{i} = 2^k$$

Lemma *(Sauer-Shelah-Perles)* Let $\mathcal{H}$ be a hypothesis class with $VCdim(H) \leq d < \infty$ and growth function $\tau_{\mathcal{H}}$. Then, for all m, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$. In particular, if $m > d+1$ and $d > 2$ then $\tau_{\mathcal{H}}(m) < m^d$.

Problem 2. Least squares and regularized least squares

1. We have
$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \mathcal{J}(\beta) := \|y - X\beta\|^2$$
$$\mathcal{J}(\beta) = \beta^T X^T X \beta - 2\beta^T X^T y + y^T y$$
$$\nabla \mathcal{J}(\beta) = 2(X^T X \beta - X^T y)$$
Equating $\nabla \mathcal{J}$ to 0, we get $\hat{\beta} = (X^T X)^{-1} X^T y$.

2. In this case, we have

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \mathcal{J}'(\beta) := \|y - X\beta\|^2 + \lambda\|\beta\|^2$$

$$\mathcal{J}'(\beta) = \beta^T X^T X\beta - 2\beta^T X^T y + y^T y + \lambda\beta^T\lambda$$
$$\nabla\mathcal{J}'(\beta) = 2(X^T X\beta - X^T y + \lambda\beta)$$

Equating $\nabla\mathcal{J}$ to 0, we get $\hat{\beta} = (X^T X + \lambda I_d)^{-1} X^T y$.
Increasing the regularization parameter reduces the variance of the model at the cost of increasing its bias towards solutions with a small l_2-norm.

Problem 3. Linear regression with projections

Refer to the lecture notes.

Problem 4. Bias-variance decomposition

The three contributions are

$$\begin{aligned} \text{Noise} &= \mathbb{E}_{x,y}\left[(\bar{h}(x) - y)^2\right] \\ (\text{Bias})^2 &= \mathbb{E}_x\left[\left(\mathbb{E}_S\left[h_S(x)\right] - \bar{h}(x)\right)^2\right] \\ \text{Variance} &= \mathbb{E}_S\mathbb{E}_{x|S}\left[(h_S(x) - \mathbb{E}_S\left[h_S(x)\right])^2\right]. \end{aligned}$$

First let us compute the optimal estimator $\bar{h}$:

$$\bar{h}(x) = \mathbb{E}\left[y|x\right] = \beta^T x$$

With this we can already compute the noise part:

$$\mathbb{E}_{x,y}\left[(\bar{h}(x) - y)^2\right] \; = \mu^2\mathbb{E}\left[\epsilon^2\right] = \mu^2.$$

Let's now focus on the data-dependent estimator:

$$h_S(x) = \begin{cases} ((X_\mathcal{A}^T X_\mathcal{A})^{-1} X_\mathcal{A}^T y)^T x_\mathcal{A}, & p < n-1 \\ (X_\mathcal{A}^T (X_\mathcal{A}^T X_\mathcal{A})^\dagger y)^T x_\mathcal{A}, & p > n+1. \end{cases}$$

Consider the quantity $\mathbb{E}\left[h_S(x)\right]$ for $p < n-1$:

$$\begin{aligned} \mathbb{E}\left[((X_\mathcal{A}^T X_\mathcal{A})^{-1} X_\mathcal{A}^T y)^T x_\mathcal{A}\right] &= \mathbb{E}\left[(X_\mathcal{A}\beta_\mathcal{A} + X_{\mathcal{A}^C}\beta_{\mathcal{A}^C} + \mu\epsilon)^T X_\mathcal{A}(X_\mathcal{A}^T X_\mathcal{A})^{-1} x_\mathcal{A}\right] \\ &= \mathbb{E}\left[\beta_\mathcal{A}^T X_\mathcal{A}^T X_\mathcal{A}(X_\mathcal{A}^T X_\mathcal{A})^{-1} x_\mathcal{A}\right] \\ &= \beta_\mathcal{A}^T x_\mathcal{A}. \end{aligned}$$

Similarly for $p > n+1$:

$$\begin{aligned} \mathbb{E}\left[(X_\mathcal{A}^T (X_\mathcal{A}^T X_\mathcal{A})^\dagger y)^T x_\mathcal{A}\right] &= \beta_\mathcal{A}^T \mathbb{E}\left[X_\mathcal{A}^T (X_\mathcal{A} X_\mathcal{A}^T)^\dagger X_\mathcal{A}\right] x_\mathcal{A} \\ &= \frac{n}{p}\beta_\mathcal{A} x_\mathcal{A}. \end{aligned}$$

Let's define the following quantity:

$$\psi = \begin{cases} 1; & p < n-1 \\ n/p; & p > n+1. \end{cases}$$

Then we have

$$\mathbb{E}\left[h_S(x)\right] = \psi \beta_{\mathcal{A}} x_{\mathcal{A}}.$$

Computation of the (Bias)^2 contribution:

$$\begin{aligned} \mathbb{E}_x\left[\left(\mathbb{E}_S\left[h_S(x)\right] - \bar{h}(x)\right)^2\right] &= \mathbb{E}_x\left[\left(\psi\beta_{\mathcal{A}}^T x_{\mathcal{A}} - \beta^T x\right)^2\right] \\ &= \mathbb{E}_x\left[\left((\psi-1)\beta_{\mathcal{A}}^T x_{\mathcal{A}} - \beta_{\mathcal{A}^C}^T x_{\mathcal{A}^C}\right)^2\right] \\ &= \mathbb{E}_x\left[(\psi-1)^2\left(\beta_{\mathcal{A}}^T x_{\mathcal{A}}\right)^2\right] + \mathbb{E}_x\left[\left(\beta_{\mathcal{A}^C}^T x_{\mathcal{A}^C}\right)^2\right] \\ &= (\psi-1)^2\|\beta_{\mathcal{A}}\|^2 + \|\beta_{\mathcal{A}^C}\|^2. \end{aligned}$$

We now compute the variance:

$$\begin{aligned} \mathbb{E}_S\mathbb{E}_{x|S}\left[(h_S(x) - \mathbb{E}_S\left[h_S(x)\right])^2\right] &= \mathbb{E}_S\mathbb{E}_{x|S}\left[\left(\hat{\beta}_{\mathcal{A}} x_{\mathcal{A}} - \psi\beta_{\mathcal{A}} x_{\mathcal{A}}\right)^2\right] \\ &= \mathbb{E}_S\left[\|\hat{\beta}_{\mathcal{A}} - \psi\beta_{\mathcal{A}}\|^2\right] = \mathbb{E}_S\left[\|\beta_{\mathcal{A}} - \hat{\beta}_{\mathcal{A}} + (\psi-1)\beta_{\mathcal{A}}\|^2\right] \\ &= \mathbb{E}_S\left[\|\beta_{\mathcal{A}} - \hat{\beta}_{\mathcal{A}}\|^2 + (\psi^2 - 2\psi + 1)\|\beta_{\mathcal{A}}\|^2 + 2(\beta_{\mathcal{A}} - \hat{\beta}_{\mathcal{A}})^T\beta_{\mathcal{A}}(\psi-1)\right] \\ &= \mathbb{E}_S\left[\|\beta_{\mathcal{A}} - \hat{\beta}_{\mathcal{A}}\|^2\right] + (\psi^2-1)\|\beta_{\mathcal{A}}\|^2 - 2(\psi-1)\mathbb{E}_S\left[\beta_{\mathcal{A}}^T\hat{\beta}_{\mathcal{A}}\right]. \end{aligned}$$

Focusing on the last term which exists only when $p > n+1$, we have:

$$\begin{aligned} \mathbb{E}_S\left[\beta_{\mathcal{A}}^T\hat{\beta}_{\mathcal{A}}\right] &= \mathbb{E}_S\left[\beta_{\mathcal{A}}^T X_{\mathcal{A}}^T (X_{\mathcal{A}} X_{\mathcal{A}}^T)^\dagger y\right] \\ &= \mathbb{E}_S\left[\beta_{\mathcal{A}}^T X_{\mathcal{A}}^T (X_{\mathcal{A}} X_{\mathcal{A}}^T)^\dagger (X_{\mathcal{A}}\beta_{\mathcal{A}} + X_{\mathcal{A}^C}\beta_{\mathcal{A}^C} + \mu\epsilon)\right] \\ &= \mathbb{E}_S\left[\beta_{\mathcal{A}}^T X_{\mathcal{A}}^T (X_{\mathcal{A}} X_{\mathcal{A}}^T)^\dagger X_{\mathcal{A}}\beta_{\mathcal{A}}\right] \\ &= \mathbb{E}_S\left[\mathrm{Tr}\{\beta_{\mathcal{A}}\beta_{\mathcal{A}}^T X_{\mathcal{A}}^T (X_{\mathcal{A}} X_{\mathcal{A}}^T)^\dagger X_{\mathcal{A}}\}\right] \\ &= \mathrm{Tr}\{\beta_{\mathcal{A}}\beta_{\mathcal{A}}^T \mathbb{E}_S\left[X_{\mathcal{A}}^T (X_{\mathcal{A}} X_{\mathcal{A}}^T)^\dagger X_{\mathcal{A}}\right]\} \\ &= \mathrm{Tr}\{\beta_{\mathcal{A}}\beta_{\mathcal{A}}^T I_p \frac{n}{p}\} = \frac{n}{p}\|\beta_{\mathcal{A}}\|^2. \end{aligned}$$

Plugging back, we get

$$\mathbb{E}_S\mathbb{E}_{x|S}\left[(h_S(x) - \mathbb{E}_S\left[h_S(x)\right])^2\right] = \mathbb{E}_S\left[\|\beta_{\mathcal{A}} - \hat{\beta}_{\mathcal{A}}\|^2\right] - (1-\psi)^2\|\beta_{\mathcal{A}}\|^2.$$

Therefore,

$$\text{Variance} = \begin{cases} \mathbb{E}_S\left[\|\beta_{\mathcal{A}} - \hat{\beta}_{\mathcal{A}}\|^2\right], & p < n-1 \\ \mathbb{E}_S\left[\|\beta_{\mathcal{A}} - \hat{\beta}_{\mathcal{A}}\|^2\right] - (1-\frac{n}{p})^2\|\beta_{\mathcal{A}}\|^2, & p > n+1. \end{cases}$$

By using the expression obtained for $\mathbb{E}_S\left[\|\beta_{\mathcal{A}} - \hat{\beta}_{\mathcal{A}}\|^2\right]$ in the class, we have the following contributions to the error.

For $p < n - 1$:

$$\text{Error} = \underbrace{\mu^2}_{\text{Noise}} + \underbrace{\|\beta_{\mathcal{A}^C}\|^2}_{\text{Bias}^2} + \underbrace{\frac{p}{n-p-1}(\mu^2 + \|\beta_{\mathcal{A}^C}\|^2)}_{\text{Variance}}.$$

For $p > n + 1$:

$$\begin{aligned}\text{Error} = \underbrace{\mu^2}_{\text{Noise}} &+ \underbrace{\|\beta_{\mathcal{A}^C}\|^2 + (1-n/p)^2\|\beta_{\mathcal{A}}\|^2}_{\text{Bias}^2} \\ &+ \underbrace{(1-n/p)\|\beta_{\mathcal{A}^C}\|^2 + \frac{n}{p-n-1}(\mu^2 + \|\beta_{\mathcal{A}^C}\|^2) - (1-n/p)^2\|\beta_{\mathcal{A}}\|^2}_{\text{Variance}}.\end{aligned}$$

Define $\alpha = p/n$ and $\varphi = n/d$. Assume $\mathcal{A}$ as a uniformly random subset of $1, 2, \cdots, d$ and taking $p, n, d \to \infty$ with α and φ finite. For $\alpha < 1$:

$$\text{Error} = \underbrace{\mu^2}_{\text{Noise}} + \underbrace{(1-\alpha\varphi)\|\beta\|^2}_{\text{Bias}^2} + \underbrace{\mu^2 + (1-\alpha\varphi)\|\beta\|^2 \frac{\alpha}{1-\alpha}}_{\text{Variance}}.$$

For $\alpha > 1$:

$$\begin{aligned}\text{Error} = \underbrace{\mu^2}_{\text{Noise}} &+ \underbrace{(1-\alpha\varphi)\|\beta\|^2 + \frac{(\alpha-1)^2}{\alpha}\varphi\|\beta\|^2}_{\text{Bias}^2} \\ &+ \underbrace{(\alpha-1)\varphi\|\beta\|^2 + \frac{\mu^2 + (1-\alpha\varphi)}{\alpha-1}\|\beta\|^2 - \frac{(\alpha-1)^2}{\alpha}\varphi\|\beta\|^2}_{\text{Variance}}.\end{aligned}$$