Homework 5
CS-526 Learning Theory

Problem 1. VC dimension of union
Let $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_r$ be hypothesis classes over some fixed domain set $\mathcal{X}$. Let $d = \max_i \mathrm{VCdim}(\mathcal{H}_i)$ and assume that $d > 2$.
Prove that:

1. $\mathrm{VCdim}(\bigcup_{i=1}^r \mathcal{H}_i) \leq \frac{4d}{\log(2)} \log\left(\frac{2d}{\log(2)}\right) + \frac{2\log(r)}{\log(2)}$.
   Hint: Use Sauer's lemma for bounding the growth function and the inequality
   $$\text{“Let } a \geq 1 \text{ and } b > 0. \text{ If } x \leq a\log(x) + b \text{ then } x \leq 4a\log(2a) + 2b.\text{”}$$
2. For $r = 2$, the bound can be strengthened to $\mathrm{VCdim}(\mathcal{H}_1 \cup \mathcal{H}_2) \leq 2d + 1$.
   Hint: $\sum_{i=0}^k \binom{k}{i} = 2^k$

Problem 2. Least squares and regularized least squares
Consider the linear least squares minimization problem:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \|y - X\beta\|^2,$$

where $y \in \mathbb{R}^n, \beta \in \mathbb{R}^d, X \in \mathbb{R}^{n \times d}$.

1. Assuming that n and d are such that the inverse of $X^T X$ is well defined, show that:
   $$\hat{\beta} = \left(X^T X\right)^{-1} X^T y.$$
2. Consider now the modified problem
   $$\hat{\beta}_\lambda = \arg\min_{\beta \in \mathbb{R}^d} \|y - X\beta\|^2 + \lambda\|\beta\|^2,$$
   where $\lambda > 0$ is a regularization parameter. Show that
   $$\hat{\beta}_\lambda = \left(X^T X + \lambda I_d\right)^{-1} X^T y$$
   and discuss the role of λ in the light of the bias-variance trade-off.

Problem 3. Linear regression with projections
Consider the linear regression model with projections defined in class. Adapt the calculation made in class for the case $p < n - 1$ and obtain

$$\mathcal{R}_\mathcal{A}(\beta) = \left(\|\beta_{\mathcal{A}^C}\|^2 + \mu^2\right)\left(1 + \frac{p}{n - p - 1}\right)$$

for a given $\mathcal{A}$, and for $\mathcal{A}$ being a uniformly random subset of $[d]$ of cardinality p :

$$\mathcal{R}(\beta) = \left(\left(1 - \frac{p}{d}\right)\|\beta\|^2 + \mu^2\right)\left(1 + \frac{p}{n-p-1}\right)$$

These expressions correspond to the left side of the double descent curve. (Hint: this computation is done in the lecture notes)

Problem 4. Bias-variance decomposition
Consider the bias-variance-noise decomposition derived in class:

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

where $S = \left\{\left(x^k, y^k\right)\right\}_{k=1}^n$ is the training set with $x^k \in \mathbb{R}^d, y^k \in \mathbb{R}$ for $k = 1, \ldots, n$, (x, y) a new data sample outside S, h_S is an estimator, and $\bar{h}$ the optimal estimator.
Consider again the linear regression with projections defined in class.

Starting from the noise decomposition and using results derived in the class derive and plot the noise-bias-variance tradeoff curves as a function of $\alpha = p/n$ in a regime $p, n, d \to +\infty$ at the same rate. You will set $\phi = n/d$ and distinguish two cases $\phi < 1$ and $\phi > 1$ (notice that in this model $\alpha < 1/\phi$ so for $\phi > 1$ there is no overparametrized regime).