

# Statistical Learning Theory (S&DS 669)

## Problem Set 1

Instructor: Omar Montasser

TA: Herlock Rahimi

Due Date: Friday, September 26, 2025 (11:59 PM ET).

### Instructions

- You may collaborate with your classmates, but each student must write solutions on their own. You also need to write down who you worked with.
- If you use a language model, then you need to provide a transcript of the interaction as a supplement. Refer to the policy on the Canvas page for more guidelines.
- If you use any sources other than the class material and references available on Canvas, then mention that. It's fine to look up a complicated sum or inequality, but don't look up an entire solution.
- Submit your solutions in  $\text{\LaTeX}$  by uploading a PDF to Canvas.

### Problems

**Problem 1** (VC Dimension of Halfspaces). Let  $\mathcal{X} = \mathbb{R}^d$  and let  $\mathcal{H}$  denote the class of homogenous halfspaces defined as

$$\mathcal{H} = \left\{ x \mapsto \text{sign}(\langle w, x \rangle) \mid w \in \mathbb{R}^d \right\}.$$

- (a) Show that  $\text{vc}(\mathcal{H}) \geq d$ . That is, show that there exists  $d$  points  $x_1, \dots, x_d \in \mathbb{R}^d$  that can be shattered by  $\mathcal{H}$ .
- (b) Show that  $\text{vc}(\mathcal{H}) \leq d$ . That is, show that *any* set of  $d + 1$  points *can not* be shattered by  $\mathcal{H}$ . **Hint:** Recall that any set of  $d + 1$  points  $x_1, x_2, \dots, x_{d+1} \in \mathbb{R}^d$  must be linearly dependent, i.e., there exists scalar coefficients  $a_1, \dots, a_{d+1}$  (not all zero) such that  $a_1 x_1 + \dots + a_{d+1} x_{d+1} = 0$ .

**Problem 2** (VC Dimension of Composition). Let  $\mathcal{X}$  be an arbitrary instance space and let  $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$  be an arbitrary binary hypothesis class where  $\text{vc}(\mathcal{H}) = d$ . Let  $k \in \mathbb{N}$

and fix a Boolean function  $f : \{\pm 1\}^k \rightarrow \{\pm 1\}$ . Consider the binary class resulting from composing  $k$  functions from  $\mathcal{H}$  with  $f$ ,

$$\mathcal{H}_f^k = \{x \mapsto f(h_1(x), \dots, h_k(x)) \mid h_1, \dots, h_k \in \mathcal{H}\}.$$

Show that  $\text{vc}(\mathcal{H}_f^k) \leq O(dk \log(k))$ .

**Hint:** Use the **Sauer-Shelah-Perles Lemma**. Suppose that we have a set  $S$  of  $m$  points. The number of ways of labeling  $S$  using functions in  $\mathcal{H}$  is at most  $\left(\frac{em}{d}\right)^d$ . Use this to bound the number of ways of labeling  $S$  using functions in  $\mathcal{H}_f^k$ . Now choose  $m$  such that this is less than  $2^m$ , which implies that  $\text{vc}(\mathcal{H}_f^k)$  must be less than the chosen  $m$ .

**Problem 3** (VC Dimension and Number of Parameters). Let  $\mathcal{X} = \mathbb{R}$  and consider the hypothesis class

$$\mathcal{H} = \{x \mapsto \text{sign}(\sin(\theta x)) \mid \theta \in \mathbb{R}\}.$$

Show that  $\text{vc}(\mathcal{H})$  is infinite. This example highlights that the VC dimension is not always proportional to the number of parameters defining a hypothesis.

**Hint:** You may find the identity  $\sin(\pi(n+a)) = (-1)^n \sin(\pi a)$  helpful.

**Problem 4** (Optimistic Bounds, Bernstein Inequality). Let  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$  be a finite hypothesis class and  $D$  be an arbitrary distribution over  $\mathcal{X} \times \mathcal{Y}$ . Let  $\delta \in (0, 1)$ , and  $m \in \mathbb{N}$ . Recall that in lecture, we proved uniform convergence bounds of the following forms:

- (Realizable Case). If  $\exists h^* \in \mathcal{H}$  such that  $L_D(h) = 0$ , then with probability at least  $1 - \delta$  over  $S \sim D^m$ :

$$\forall h \in \mathcal{H} : L_S(h) = 0 \implies L_D(h) \leq \frac{\log |\mathcal{H}| + \log(1/\delta)}{m}. \quad (1)$$

- (Agnostic Case). With probability at least  $1 - \delta$  over  $S \sim D^m$ :

$$\forall h \in \mathcal{H} : |L_S(h) - L_D(h)| \leq \sqrt{\frac{\log |\mathcal{H}| + \log(2/\delta)}{2m}}. \quad (2)$$

In this problem, we will prove a more general uniform convergence bound that interpolates the realizable and agnostic cases, using Bernstein's inequality. Specifically, our goal is to show that with probability at least  $1 - \delta$  over  $S \sim D^m$ ,

$$\forall h \in \mathcal{H} : |L_S(h) - L_D(h)| \leq c_1 \sqrt{\frac{L_D(h) \log(|\mathcal{H}|/\delta)}{m}} + c_2 \frac{\log(|\mathcal{H}|/\delta)}{m}, \quad (3)$$

for some absolute constants  $c_1, c_2$ . Observe that the bound in Equation 3 can be much better than the agnostic bound in Equation 2, when  $L_D(h)$  is small. For example, when  $L_D(h) = 0$ , Equation 3 recovers the bound in Equation 1 (up to constants).

- To prove the bound in Equation 3, we will use Bernstein's inequality, which is stated next.

**Theorem 1** (Bernstein's Inequality). Let  $X_1, \dots, X_m$  be  $m$  iid real-valued random variables with mean zero ( $\mathbb{E}[X_i] = 0$ ), and bounded  $|X_i| \leq M$ . Then, for all  $t > 0$

$$\mathbb{P} \left[ \sum_{i=1}^m X_i \geq t \right] \leq \exp \left( - \frac{\frac{t^2}{2}}{\sum_{i=1}^m \mathbb{E}[X_i^2] + M \frac{t}{3}} \right).$$

Use the inequality above to show that with probability at least  $1 - \delta$ ,

$$\left| \frac{1}{m} \sum_{i=1}^m X_i \right| \leq \sqrt{\frac{2\mathbb{E}[X_1^2] \log(2/\delta)}{m}} + \frac{2M \log(2/\delta)}{3m}. \quad (4)$$

- (b) Use Equation 4 and the union bound over  $\mathcal{H}$  to show Equation 3. **Hint:** this requires carefully choosing/defining an appropriate random variable  $X_i$ .
- (c) Let  $\hat{h} = \text{ERM}_{\mathcal{H}}(S) = \text{argmin}_{h \in \mathcal{H}} L_S(h)$  be the output of  $\text{ERM}_{\mathcal{H}}$  when receiving a sample  $S$  as input. Using Equation 3, show that with probability at least  $1 - \delta$  over  $S \sim D^m$ ,

$$L_D(\hat{h}) \leq L_D(h^*) + c_3 \sqrt{\frac{L_D(h^*) \log(|\mathcal{H}|/\delta)}{m}} + c_4 \frac{\log(|\mathcal{H}|/\delta)}{m}, \quad (5)$$

where  $h^* = \text{argmin}_{h \in \mathcal{H}} L_D(h)$ , and  $c_3, c_4$  are some absolute constants.