Stat 6390.001: High Dimensional Probability and Statistics, Fall 2026 **Topic 12**

Rademacher Complexity, Polynomial Discrimination, and VC Dimension

Lecturer: Yun Wei *Scribe: Md Maidul Husain*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

1 Recall: Rademacher Complexity

We first recall the empirical Rademacher complexity and the Rademacher complexity of a function class. Let $X_1, \ldots, X_n$ be i.i.d. observations from P, let $\mathcal{F}$ be a class of real-valued functions, and let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables, independent of the observations. For a fixed sample $x_1^n = (x_1, \ldots, x_n)$, the empirical Rademacher complexity is

$$R\big(\mathcal{F}(x_1^n)/n\big) := \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right]. \tag{1}$$

The Rademacher complexity of the class $\mathcal{F}$ is

$$R_n(\mathcal{F}) := \mathbb{E}_X \left[R\big(\mathcal{F}(X_1^n)/n\big) \right]. \tag{2}$$

We also recall the symmetrization inequality

$$\mathbb{E}\big[\|P_n - P\|_{\mathcal{F}}\big] \le 2R_n(\mathcal{F}), \tag{3}$$

where $P_n f = n^{-1} \sum_{i=1}^n f(X_i)$, $Pf = \mathbb{E}[f(X)]$, and $\|P_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |P_n f - Pf|$.

Theorem 1.1. *For any b-uniformly bounded class of functions $\mathcal{F}$, any $n \ge 1$, and any scalar $\delta > 0$, we have*

$$\|P_n - P\|_{\mathcal{F}} \le 2R_n(\mathcal{F}) + \delta \tag{4}$$

with P-probability at least $1 - \exp\big(-n\delta^2/(2b^2)\big)$. Here b-uniformly bounded means $|f(x)| \le b$ for all $f \in \mathcal{F}$ and x.

Proof. For each $f \in \mathcal{F}$, write

$$\overline{f}(x) := f(x) - \mathbb{E}[f(X)].$$

Define

$$G(X_1, \ldots, X_n) := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \overline{f}(X_i) \right|.$$

Thus, $\|P_n - P\|_{\mathcal{F}} = G(X_1, \ldots, X_n)$.

Let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ satisfy $y_i = x_i$ for all $i \ne k$, where $k \in \{1, \ldots, n\}$. Then

$$\begin{aligned} |G(x) - G(y)| &\le \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \overline{f}(x_i) - \frac{1}{n} \sum_{i=1}^n \overline{f}(y_i) \right| \\ &= \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \overline{f}(x_k) - \overline{f}(y_k) \right| \\ &= \sup_{f \in \mathcal{F}} \frac{1}{n} |f(x_k) - f(y_k)| \\ &\le \frac{2b}{n}. \end{aligned}$$

Consequently, G satisfies the bounded differences property with $L_k = 2b/n$ for $k = 1, \dots, n$. Applying the bounded differences inequality yields, for every $\delta > 0$,

$$\begin{aligned} &P\left(G(X_1, \dots, X_n) - \mathbb{E}[G(X_1, \dots, X_n)] > \delta\right) \\ &\quad \le \exp\left(-\frac{2\delta^2}{\sum_{k=1}^n L_k^2}\right) = \exp\left(-\frac{2\delta^2}{n(2b/n)^2}\right) = \exp\left(-\frac{n\delta^2}{2b^2}\right). \end{aligned}$$

Equivalently,

$$P\left(\|P_n - P\|_{\mathcal{F}} > \mathbb{E}[\|P_n - P\|_{\mathcal{F}}] + \delta\right) \le \exp\left(-\frac{n\delta^2}{2b^2}\right).$$

Finally, using (3),

$$\begin{aligned} &P\left(\|P_n - P\|_{\mathcal{F}} > 2R_n(\mathcal{F}) + \delta\right) \\ &\quad \le \exp\left(-\frac{n\delta^2}{2b^2}\right), \qquad \delta > 0. \end{aligned}$$

This is the claimed probability bound. □

Our next goal is to obtain a bound on $R_n(\mathcal{F})$ that tends to zero as n increases.

2 Classes with Polynomial Discrimination (Section 4.3.1)

We follow Section 4.3.1. If $\mathcal{F}$ is a class of binary-valued functions, the number of distinct vectors

$$\mathcal{F}(x_1^n) := \{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}\}$$

is at most 2^n, because there are two choices at each of the n coordinates. For example, n sample points allow at most 2^n binary combinations. Polynomial discrimination requires this number to grow no faster than a polynomial in n.

Definition 2.1 (Polynomial discrimination)**.** A class $\mathcal{F}$ of functions with domain $\mathcal{X}$ has *polynomial discrimination* of order $\nu \ge 1$ with parameter n_0 if, for any $n \ge n_0$ and any $x_1^n = (x_1, \dots, x_n)$ of n points in $\mathcal{X}$,

$$\operatorname{card}(\mathcal{F}(x_1^n)) \le (n+1)^\nu. \tag{5}$$

Lemma 2.2. *Suppose that $\mathcal{F}$ has polynomial discrimination of order $\nu \ge 1$ and parameter n_0. Then, for every $n \ge n_0$ and for any $x_1^n = (x_1, \dots, x_n)$,*

$$R\big(\mathcal{F}(x_1^n)/n\big) \le 2D(x_1^n)\sqrt{\frac{\nu \log(n+1)}{n}}, \tag{6}$$

where

$$D(x_1^n) := \sup_{f \in \mathcal{F}} \sqrt{\frac{1}{n}\sum_{i=1}^n f(x_i)^2} \tag{7}$$

is the ℓ_2-radius of the set $\mathcal{F}(x_1^n)/\sqrt{n}$.

Homework 3. The proof is assigned as Homework 3. □

Lemma 2.3. *If $\mathcal{F}$ is b-uniformly bounded and has polynomial discrimination of order $\nu \ge 1$ with parameter n_0, then*

$$R_n(\mathcal{F}) \le 2b\sqrt{\frac{\nu \log(n+1)}{n}}, \qquad \forall n \ge n_0. \tag{8}$$

Proof. By the definition of Rademacher complexity and Lemma 2.2,

$$\begin{aligned} R_n(\mathcal{F}) &= \mathbb{E}_X\left[R\big(\mathcal{F}(X_1^n)/n\big)\right] \\ &\le \mathbb{E}_X\left[2D(X_1^n)\sqrt{\frac{\nu\log(n+1)}{n}}\right] \\ &= 2\mathbb{E}_X[D(X_1^n)]\sqrt{\frac{\nu\log(n+1)}{n}}. \end{aligned}$$

Since $|f(x)| \le b$ for every $f \in \mathcal{F}$ and x,

$$\mathbb{E}_X[D(X_1^n)] = \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\sqrt{\frac{1}{n}\sum_{i=1}^n f(X_i)^2}\right] \le b.$$

Therefore,

$$R_n(\mathcal{F}) \le 2b\sqrt{\frac{\nu\log(n+1)}{n}}.$$

□

3 The Classical Glivenko–Cantelli Theorem

Corollary 3.1 (Classical Glivenko–Cantelli)**.** *Let $F(t) = P(X \le t)$ be the cumulative distribution function of X, and let $\widehat{F}_n(t) = n^{-1}\sum_{i=1}^n I_{(-\infty,t]}[X_i]$ be the empirical cumulative distribution function. Then, for all $\delta > 0$,*

$$P\left(\|\widehat{F}_n - F\|_\infty \ge 4\sqrt{\frac{\log(n+1)}{n}} + \delta\right) \le \exp\left(-\frac{n\delta^2}{2}\right) \quad \forall\delta \ge 0 \tag{9}$$

Consequently,

$$\|\widehat{F}_n - F\|_\infty \longrightarrow 0 \qquad \textit{almost surely.} \tag{10}$$

Proof. Consider the class of indicator functions

$$\mathcal{F} := \{f_t : t \in \mathbb{R}\}, \qquad f_t(x) = I_{(-\infty,t]}[x].$$

For this class,

$$\|\widehat{F}_n - F\|_\infty = \|P_n - P\|_{\mathcal{F}}.$$

Since each f_t takes values in $\{0,1\}$, the class is 1-uniformly bounded. Theorem 1.1 therefore gives

$$P\left(\|\widehat{F}_n - F\|_\infty \ge 2R_n(\mathcal{F}) + \delta\right) \le \exp\left(-\frac{n\delta^2}{2}\right), \qquad \delta \ge 0. \tag{11}$$

To verify polynomial discrimination, take any sample points and order them as $x_{(1)} \le \cdots \le x_{(n)}$. As t moves from $-\infty$ to $+\infty$, the vector $(f_t(x_1), \ldots, f_t(x_n))$ can change only when t passes one of the sample points. Thus,

$$\operatorname{card}(\mathcal{F}(x_1^n)) \le n + 1.$$

In particular, $\mathcal{F}$ has polynomial discrimination of order $\nu = 1$ and parameter $n_0 = 1$. Lemma 2.3 then gives

$$R_n(\mathcal{F}) \le 2\sqrt{\frac{\log(n+1)}{n}}. \tag{12}$$

Substituting (12) into (11) proves (9). The almost-sure convergence follows from the preceding bound, as can be seen by Borel-Cantelli lemma. □

4 Vapnik–Chervonenkis Dimension (Section 4.3.2)

We now turn to Section 4.3.2 of the textbook: Vapnik–Chervonenkis (VC) dimension.

Definition 4.1 (Shattering and VC dimension)**.** Given a class $\mathcal{F}$ of binary-valued functions, we say that a set of points $x_1^n = (x_1, \ldots, x_n)$ is *shattered* by $\mathcal{F}$ if

$$\operatorname{card}(\mathcal{F}(x_1^n)) = 2^n. \tag{13}$$

The *VC dimension* $\nu(\mathcal{F})$ is the largest integer n for which there exists some x_1^n that is shattered by $\mathcal{F}$. If $\nu(\mathcal{F}) < \infty$, then $\mathcal{F}$ is called a *VC class*.

For example, a binary-valued function class indexed by a collection $\mathcal{S}$ of subsets can be written as

$$\mathcal{F} = \{\mathbb{I}_S[\cdot] : S \in \mathcal{S}\}.$$

For any $x \in \mathcal{X}$, the coordinate $\mathbb{I}_S[x]$ equals 1 when $x \in S$, and equals 0 otherwise. Thus, selecting a set $S \in \mathcal{S}$ selects which sample points receive label 1. A sample is shattered precisely when all 2^n such labelings can be realized.

References

[1] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge University Press, 2019. Sections 4.2 and 4.3