STA3000F: Advanced Theory of Statistics **Fall 2026**

Lecture 7: October 23, 2026

Lecturer: Wenlong Mou

1 General empirical process bound continued

Example of the general theorem: consider $\Theta \subseteq \mathbb{B}(0, R)$ in d-dimensional Euclidean space, and consider the function class parametrized by $\theta \in \Theta$:

$$\mathcal{F} = \{f_\theta : \theta \in \Theta\}$$

(e.g. generalized linear model)

We assume that f_θ is Lipschitz in θ for every x, i.e., there exists $M(x)$ such that

$$|f_\theta(x) - f_{\theta'}(x)| \le M(x)\|\theta - \theta'\|_2, \quad \forall \theta, \theta' \in \Theta.$$

(We allow the Lipschitz constant to depend on x.) We also assume that $\mathbb{E}[M^2(X)] < +\infty$.

In such a case, we can apply the main theorem to bound $\mathbb{E}\big[\sup_{f\in\mathcal{F}} |(P_n - P)f|\big]$. To start with, we need to construct an envelope function for the function class $\mathcal{F}$. We can take

$$G(x) = |f_{\theta_0}(x)| + 2RM(x),$$

for some fixed $\theta_0 \in \Theta$. Then we have

$$|f_\theta(x)| \le G(x), \quad \forall \theta \in \Theta.$$

Now we need to bound the covering number of $\mathcal{F}$ with respect to the $\mathbb{L}^2(Q)$ metric, for any distribution Q. First, we note that

$$N(\mathcal{F}, \mathbb{L}^2(Q), \varepsilon\|G\|_{\mathbb{L}^2(Q)}) \le N(\mathcal{F}, \mathbb{L}^2(Q), \varepsilon\|M\|_{\mathbb{L}^2(Q)}R).$$

Now we need to bound the right-hand-side by direct construction of a covering set. Let $(\theta_i)_{i=1}^N$ be an δ-covering of Θ with respect to the ℓ_2 norm. For any $\theta \in \Theta$, there exists some θ_i such that $\|\theta - \theta_i\|_2 \le \delta$. Then we have

$$\|f_\theta - f_{\theta_i}\|^2_{\mathbb{L}^2(Q)} = \int |f_\theta(x) - f_{\theta_i}(x)|^2 dQ(x) \le \int M^2(x)\|\theta - \theta_i\|_2^2 dQ(x) \le \delta^2 \|M\|^2_{\mathbb{L}^2(Q)}.$$

That means $(f_{\theta_i})_{i=1}^N$ is a $\delta\|M\|_{\mathbb{L}^2(Q)}$-covering of $\mathcal{F}$ with respect to the $\mathbb{L}^2(Q)$ metric. Therefore, by taking $\delta = \varepsilon R$, we have

$$N(\mathcal{F}, \|\cdot\|_{\mathbb{L}^2(Q)}, \varepsilon R\|M\|_{\mathbb{L}^2(Q)}) \le N(\Theta, \|\cdot\|_2, \varepsilon R) \le N(\mathbb{B}_2^d(R), \|\cdot\|_2, \varepsilon R) \le \Big(1 + \frac{2}{\varepsilon}\Big)^d.$$

Substituting this bound into the main theorem, we have

$$\begin{aligned}
\mathbb{E}\Big[\sup_{f\in\mathcal{F}} |(P_n - P)f|\Big] &\le c\sqrt{\frac{\mathbb{E}[G^2(X)]}{n}} \int_0^1 \sqrt{\log\big(2\sup_Q N(\mathcal{F}; \mathbb{L}^2(Q), \|G\|_{\mathbb{L}^2(Q)}\,\epsilon)\big)}d\epsilon \\
&\le c'\sqrt{\frac{\mathbb{E}[G^2(X)]}{n}} \int_0^1 \sqrt{d\log(1/\epsilon)}d\epsilon \\
&\le c'' \cdot \big(\|f_{\theta_0}\|_{\mathbb{L}^2(\mathbb{P})} + R\|M\|_{\mathbb{L}^2(\mathbb{P})}\big)\sqrt{\frac{d}{n}}.
\end{aligned}$$

This applies to a broad class of Lipschitz models. On the other hand, some function classes may not be continuously parametrized by a finite-dimensional parameter θ.

e.g. consider the function class

$$\mathcal{F} = \left\{ \mathbf{1}_{x \in [\theta_1, \theta_2]} : \theta_1, \theta_2 \in \mathbb{R} \right\},$$

is a very simple function class parametrized by two parameters θ_1, θ_2, but not continuously. The above argument does not apply. On the other hand, covering/packing number is a robust notion — as long as the function class is "simple", we can still get a good bound.

2 Combinatorial notion of complexity: VC dimension

First, let us consider binary functions mapping from $\mathcal{X}$ to $\{0, 1\}$ (we will extend back to real-valued functions later). Consider a class of functions $\mathcal{F}$.

Definition 7.1 (shattering)**.** *A set of points $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ is said to be shattered by $\mathcal{F}$ if for every subset $S \subset \{1, \ldots, n\}$, there exists a function $f \in \mathcal{F}$ such that*

$$f(x_i) = \begin{cases} 1, & i \in S \\ 0, & i \notin S \end{cases}, \quad i = 1, \ldots, n.$$

Intuitively, a function class $\mathcal{F}$ needs to be "rich enough" to realize all possible labelings of the points in the set $\{x_1, \ldots, x_n\}$, in order to shatter it. From machine learning perspective, you can also think of shattering as a classifier being able to fit any possible labeling of the points in the set $\{x_1, \ldots, x_n\}$.

Based on the notion of shattering, we can define the Vapnik-Chervonenkis (VC) dimension of a function class $\mathcal{F}$.

Definition 7.2 (VC dimension)**.** *The VC dimension of a function class $\mathcal{F}$ is defined as the largest integer d such that there exists a set of d points in $\mathcal{X}$ that can be shattered by $\mathcal{F}$. If no such largest integer exists, we say that the VC dimension is infinite.*

Intuitively, this is the largest number of data points that the binary classifier class can fit perfectly for any possible labeling.

Example 7.1. Indicator of line segments in $\mathbb{R}$: $\mathcal{F} = \{\mathbf{1}_{x \in [\theta_1, \theta_2]} : \theta_1, \theta_2 \in \mathbb{R}\}$. Consider any 3 points $x_1 < x_2 < x_3$. For the labeling $(1, 0, 1)$, there is no line segment that can realize this labeling. Therefore, the VC dimension of this function class is at most 2. On the other hand, for any 2 points $x_1 < x_2$, we can realize all possible labelings $(0, 0), (0, 1), (1, 0), (1, 1)$ by choosing appropriate line segments. Therefore, the VC dimension of this function class is exactly 2.

More generally, you can consider union of K disjoint intervals in $\mathbb{R}$:

$$\mathcal{F} = \left\{ \mathbf{1}_{x \in \bigcup_{k=1}^{K} [\theta_{2k-1}, \theta_{2k}]} : \theta_1 < \theta_2 < \cdots < \theta_{2K} \in \mathbb{R} \right\}.$$

By putting $2K + 1$ points on real line with alternating labeling $(1, 0, 1, 0, \ldots, 0, 1)$, we can see that the VC dimension of this function class is at most $2K$.

2.1 Structural results on VC dimension

Idea: If a function class has finite VC dimension, then it is "simple" in some sense, and we can bound its covering number.

Key result:

Theorem 7.1 (Sauer's lemma)**.** *Let $\mathcal{F}$ be a class of binary functions with VC dimension d. Then for any $n \geq d$, we have*

$$\left|\left\{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\right\}\right| \leq \underbrace{\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}}_{\Phi_d(n)} \leq \left(\frac{en}{d}\right)^d.$$

Remark: consider the cardinality of all possible labelings of n points by a function class $\mathcal{F}$.

- When $n \leq d$, all the 2^n possible labelings can be realized by $\mathcal{F}$.
- The definition of VC dimension only tells us that for $n = d + 1$, there exists at least one labeling that cannot be realized by $\mathcal{F}$.
- Sauer's lemma tells us that once we pass the threshold $n > d$, the number of realizable labelings is bounded by a polynomial in n (with degree d), which is much smaller than the total number of possible labelings 2^n.

Proof idea: induction on $(n + d)$. The result is trivial when $n = d$ or when $d = 0$.

By induction hypothesis, let us assume that the conclusion holds for $(n-1, d-1)$ and $(n-1, d)$. Consider the set of n points $\{x_1, \ldots, x_n\}$. We separate the labeling of the last point x_n from the first $n - 1$ points.

Consider any label sequence $b_1, \ldots, b_{n-1} \in \{0, 1\}^{n-1}$, there are 3 possibilities:

1. There is no function $f \in \mathcal{F}$ that realizes the labeling $(b_1, \ldots, b_{n-1})$ on the first $n - 1$ points.
2. Both labelings $(b_1, \ldots, b_{n-1}, 0)$ and $(b_1, \ldots, b_{n-1}, 1)$ can be realized by some functions in $\mathcal{F}$.
3. Only one of the labelings $(b_1, \ldots, b_{n-1}, 0)$ and $(b_1, \ldots, b_{n-1}, 1)$ can be realized by some function in $\mathcal{F}$.

The idea is to form a decomposition of the set $\{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}$ based on the labeling of the first $n - 1$ points. The first case does not contribute to the cardinality. For the second and third cases, we can decompose the class into $\mathcal{F}_1$ and $\mathcal{F}_2$ according to the following rule (here we abuse the notation a bit, and think of $\mathcal{F}_1$ and $\mathcal{F}_2$ are defined on the domain $\{x_1, \ldots, x_n\}$, and equivalently, they are just labeling sequences):

- In case 2, suppose $f_1, f_2 \in \mathcal{F}$ realize the labelings $(b_1, \ldots, b_{n-1}, 0)$ and $(b_1, \ldots, b_{n-1}, 1)$ respectively. We add f_1 to $\mathcal{F}_1$ and f_2 to $\mathcal{F}_2$.
- In case 3, there is only one function $f \in \mathcal{F}$ that realizes the labeling $(b_1, \ldots, b_{n-1}, b_n)$ for some $b_n \in \{0, 1\}$. We add f to $\mathcal{F}_1$.

By this construction, we have

$$|\mathcal{F}(x_1, \ldots, x_n)| = |\mathcal{F}_1| + |\mathcal{F}_2|.$$

For the set $\mathcal{F}_1$, we note that this set is formed by mapping all the achievable labelings of the first $n-1$ points to an element in $\mathcal{F}_1$. So we have

$$|\mathcal{F}_1| = |\mathcal{F}(x_1, \ldots, x_{n-1})| \leq \Phi_d(n-1),$$

by induction hypothesis.

For the set $\mathcal{F}_2$, we claim that

$$\text{VCdim}(\mathcal{F}_2) \leq d-1. \tag{7.1}$$

To prove this claim, we argue by contradiction. Suppose that there exists $(x_{k_1}, x_{k_2}, \cdots, x_{k_d})$ that can be shattered by $\mathcal{F}_2$. Because $\mathcal{F}_2$ is formed by sequences that allow free labeling of the last point x_n, we can add this free labeling to the set of d points. This means that $(x_{k_1}, x_{k_2}, \cdots, x_{k_d}, x_n)$ can be shattered by $\mathcal{F}$. This contradicts the assumption that $\mathcal{F}$ has VC dimension d. Therefore, we have proved the claim (7.1). By induction hypothesis, we have

$$|\mathcal{F}_2| \leq \Phi_{d-1}(n-1).$$

So we conclude that

$$|\mathcal{F}(x_1, \ldots, x_n)| = |\mathcal{F}_1| + |\mathcal{F}_2| \leq \Phi_d(n-1) + \Phi_{d-1}(n-1) = \Phi_d(n),$$

which completes the induction step. The last inequality in the theorem follows from the standard bound on binomial coefficients.

Application of Sauer's lemma: VC dimension of composite classes Consider two classes of binary functions $\mathcal{F}_1$ and $\mathcal{F}_2$, with VC dimensions d_1 and d_2 respectively. For any binary operation $\clubsuit$ (e.g. AND, OR, XOR), we can define a composite class of functions

$$\mathcal{F} = \{f_1 \clubsuit f_2 : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}.$$

Can we bound the VC dimension of $\mathcal{F}$ in terms of d_1 and d_2?

Given $(x_1, \ldots, x_n)$, we can consider the set of achievable labelings by $\mathcal{F}$:

$$\mathcal{F}(x_1, \ldots, x_n) = \{(f_1(x_1) \clubsuit f_2(x_1), \ldots, f_1(x_n) \clubsuit f_2(x_n)) : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}.$$

The cardinality of this set is bounded by

$$|\mathcal{F}(x_1, \ldots, x_n)| \leq |\mathcal{F}_1(x_1, \ldots, x_n)| \cdot |\mathcal{F}_2(x_1, \ldots, x_n)| \leq \left(\frac{en}{d_1}\right)^{d_1} \cdot \left(\frac{en}{d_2}\right)^{d_2}.$$

On the other hand, let d be the VC dimension of $\mathcal{F}$. By definition, we have

$$2^d = \max_{x_1, \ldots, x_d} |\mathcal{F}(x_1, \ldots, x_d)| \leq \left(\frac{ed}{d_1}\right)^{d_1} \cdot \left(\frac{ed}{d_2}\right)^{d_2}.$$

Solving for d through this inequality, there exists a universal constant $c > 0$ such that

$$d \leq c(d_1 + d_2).$$

Additional examples of VC dimension: Consider the class of linear classifiers in $\mathbb{R}^d$:

$$\mathcal{F} = \Big\{ f_\theta(x) = \mathbf{1}_{\theta^\top x \geq 0} : \theta \in \mathbb{R}^d \Big\}.$$

We claim that $\mathrm{VCdim}(\mathcal{F}) = d$. For the upper bound, consider any $(d+1)$ points in $\mathbb{R}^d$: $x_1, \ldots, x_{d+1}$. For any possible labeling $(b_1, b_2, \cdots, b_{d+1}) \in \{0,1\}^{d+1}$, suppose that they can be realized by some linear classifier f_θ.

On the other hand, we note that these $(d+1)$ points are linearly dependent in $\mathbb{R}^d$. That means there exists some not-all-zero coefficients $(\alpha_1, \ldots, \alpha_{d+1})$ such that

$$\sum_{i=1}^{d+1} \alpha_i x_i = 0.$$

Multiplying all coefficients by -1 if necessary, we may assume that $\alpha_j < 0$ for some j. Let us consider the labeling $b_i = \mathbf{1}_{\alpha_i \geq 0}$, and let θ be the linear classifier that realizes this labeling. Then we have

$$\sum_{i=1}^{d+1} \alpha_i \theta^\top x_i = \theta^\top \sum_{i=1}^{d+1} \alpha_i x_i = 0, \quad \text{and} \quad \alpha_i \theta^\top x_i \geq 0, \quad \forall i = 1, \ldots, d+1.$$

The only possibility is that $\alpha_i \theta^\top x_i = 0$ for all $i = 1, \ldots, d+1$. For each i, either $\alpha_i = 0$ or $\theta^\top x_i = 0$.

However, $\alpha_j < 0$ gives $b_j = 0$, so $\theta^\top x_j < 0$ and hence $\alpha_j \theta^\top x_j > 0$, a contradiction. Thus no set of $d+1$ points can be shattered, and $\mathrm{VCdim}(\mathcal{F}) \leq d$.

For the lower bound, take the standard basis vectors $e_1, \ldots, e_d$. For any labeling $(b_1, \ldots, b_d) \in \{0,1\}^d$, choose $\theta = (2b_1 - 1, \ldots, 2b_d - 1)^\top$. Then $f_\theta(e_i) = b_i$ for every i. Hence these d points are shattered, proving $\mathrm{VCdim}(\mathcal{F}) = d$.

2.2 VC dimension and covering number

Recall that the goal is to bound the $\mathbb{L}^2(Q)$ covering number of a function class $\mathcal{F}$ for any distribution Q (for binary classes, $G \equiv 1$ is an envelope function). We have the following result:

Theorem 7.2. *For any Q, we have*

$$\log N(\mathcal{F}; \mathbb{L}^2(Q), \varepsilon) \leq c \cdot \mathit{VCdim}(\mathcal{F}) \cdot \log(1/\varepsilon),$$

for some universal constant $c > 0$.

Proof: Let $f_1, f_2, \cdots, f_N$ be a maximal ε-packing of $\mathcal{F}$ with respect to the $\mathbb{L}^2(Q)$ metric. For any $i, j \in [N]$, we have

$$\varepsilon^2 \leq \|f_i - f_j\|^2_{\mathbb{L}^2(Q)} = \int |f_i(x) - f_j(x)|^2 dQ(x) = Q(\{x : f_i(x) \neq f_j(x)\}).$$

The idea is to use probabilistic method: we draw n i.i.d. samples $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} Q$. For any $i, j \in [N]$, we have

$$\mathbb{P}\Big(\big(f_i(X_1), f_i(X_2), \ldots, f_i(X_n)\big) = \big(f_j(X_1), f_j(X_2), \ldots, f_j(X_n)\big) \Big) \leq (1 - \varepsilon^2)^n \leq \exp(-n\varepsilon^2).$$

Then we can take union bound over all pairs (i, j), and we have

$$\mathbb{P}\Big(\exists i, j \in [N], \text{ not distinguishable under} \big\{X_1, \ldots, X_n\big\}\Big) \le \binom{N}{2} \exp(-n\varepsilon^2) \le N^2 \exp(-n\varepsilon^2).$$

We choose $n = \frac{4 \log N}{\varepsilon^2}$, then the right-hand-side is less than $1/2$. That means there exists a set of n points $\{x_1, \ldots, x_n\}$ such that all the functions in the packing set are distinguishable on this set of points. That means

$$N = \Big|\Big\{(f_i(x_1), \ldots, f_i(x_n)) : i \in [N]\Big\}\Big| \le \Big|\Big\{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\Big\}\Big| \le \Phi_d(n) \le \Big(\frac{en}{d}\Big)^d = \Big(\frac{4e \log N}{d\varepsilon^2}\Big)^d.$$

Solving for N, we have

$$\log N \le cd \log(1/\varepsilon),$$

for some universal constant $c > 0$.

e.g. for linear classifiers in $\mathbb{R}^d$, $\mathcal{F} = \{f(x) = \mathbf{1}_{\theta^\top x > 0} : \theta \in \mathbb{R}^d\}$, we have

$$\mathbb{E}\big[\sup_{f \in \mathcal{F}} |(P_n - P)f|\big] \le \frac{c}{\sqrt{n}} \int_0^1 \sqrt{d \log(1/\varepsilon)} d\varepsilon \le c' \sqrt{\frac{d}{n}}.$$

This can also be used to bound the generalization error (gap between training and test error) of linear classifiers in $\mathbb{R}^d$.

2.3 VC subgraph dimension for real-valued function classes

Idea: to extend the notion of VC dimension to real-valued function classes.

Definition 7.3 (VC subgraph dimension)**.** *For a real-valued function class $\mathcal{F}$, we define*

$$VCdim(\mathcal{F}) := VCdim(\{(t, x) \mapsto \mathbf{1}_{f(x) \ge t} : f \in \mathcal{F}\}).$$

Intuitively, we augment one dimension to the input space to transform the real-valued function class into a binary function class, while keeping its complexity.

e.g. consider the class of linear functions in $\mathbb{R}^d$:

$$\mathcal{F} = \{f_\theta(x) = \theta^\top x : \theta \in \mathbb{R}^d\}.$$

We note that $t_i \ge f_\theta(x_i)$ is equivalent to the fact that

$$\begin{bmatrix} -\theta^\top & 1 \end{bmatrix} \begin{bmatrix} x_i \\ t_i \end{bmatrix} \ge 0,$$

which becomes a subclass of the $(d+1)$-dimensional linear classifiers. So we have $\text{VCdim}(\mathcal{F}) \le d+1$.

Furthermore, suppose that ϕ is a strictly increasing function, we can consider a generalized linear function class

$$\mathcal{F} = \{f_\theta(x) = \phi(\theta^\top x) : \theta \in \mathbb{R}^d\}.$$

We note that $t_i \ge f_\theta(x_i)$ is equivalent to the fact that

$$\begin{bmatrix} -\theta^\top & 1 \end{bmatrix} \begin{bmatrix} x_i \\ \phi^{-1}(t_i) \end{bmatrix} \ge 0,$$

which is also a subclass of the $(d+1)$-dimensional linear classifiers.

The nice thing about VC subgraph dimension is that it bounds the $\mathbb{L}^2(Q)$ covering number with respect to an envelope function, which is exactly what we need for the general empirical process bound. We have the following result:

Theorem 7.3. *Let $\mathcal{F}$ be a real-valued function class with envelope function G. Then we have*

$$\sup_Q \log N(\mathcal{F}; \mathbb{L}^2(Q), \varepsilon \|G\|_{\mathbb{L}^2(Q)}) \leq \left(c_1/\varepsilon\right)^{c_2\, VCdim(\mathcal{F})},$$

for some universal constants $c_1, c_2 > 0$.

Proof: for a pair of functions $f, g \in \mathcal{F}$, we have

$$\begin{aligned}
\|f-g\|^2_{\mathbb{L}^2(Q)} &= \int |f(x)-g(x)|^2 dQ(x) \leq 2\int G(x)|f(x)-g(x)|dQ(x) \\
&= 2\int G(x) \int |\mathbf{1}_{t\leq f(x)} - \mathbf{1}_{t\leq g(x)}| dt\; dQ(x) \\
&= 2 \iint_{(t,x):|t|\leq G(x)} G(x)\cdot |\mathbf{1}_{t\leq f(x)} - \mathbf{1}_{t\leq g(x)}| dt\; dQ(x).
\end{aligned}$$

(The idea is to relate the $\mathbb{L}^2(Q)$ distance between f and g to the the indicator functions)

By Cauchy–Schwarz inequality, we have

$$\begin{aligned}
&2 \iint_{(t,x):|t|\leq G(x)} G(x)\cdot |\mathbf{1}_{t\leq f(x)} - \mathbf{1}_{t\leq g(x)}| dt\; dQ(x) \\
&\leq 2\sqrt{\iint_{(t,x):|t|\leq G(x)} \left(\sqrt{G(x)}\right)^2 dt\; dQ(x)} \cdot \sqrt{\iint_{(t,x):|t|\leq G(x)} \left(\sqrt{G(x)}|\mathbf{1}_{t\leq f(x)} - \mathbf{1}_{t\leq g(x)}|\right)^2 dt\; dQ(x)}.
\end{aligned}$$

For the first integral, we note that

$$\iint_{(t,x):|t|\leq G(x)} \left(\sqrt{G(x)}\right)^2 dt\; dQ(x) = 2\int G^2(x) dQ(x) = 2\|G\|^2_{\mathbb{L}^2(Q)}.$$

For the second integral, it is

$$\iint_{(t,x):|t|\leq G(x)} |\mathbf{1}_{t\leq f(x)} - \mathbf{1}_{t\leq g(x)}|^2 G(x) dt\; dQ(x),$$

which is the $\mathbb{L}^2(G(x)dtdQ(x))$-distance of two indicator functions. Indeed, we can construct a probability measure

$$d\widetilde{Q}(t,x) \propto G(x) dt\; dQ(x).$$

We have

$$\iint_{(t,x):|t|\leq G(x)} |\mathbf{1}_{t\leq f(x)} - \mathbf{1}_{t\leq g(x)}|^2 G(x) dt\; dQ(x) = \left\|\mathbf{1}_{t\leq f(x)} - \mathbf{1}_{t\leq g(x)}\right\|^2_{\mathbb{L}^2(\widetilde{Q})} \cdot 2\|G\|^2_{\mathbb{L}^2(Q)}.$$

That means an ε-covering of the class of indicator functions $\{(t,x) \mapsto \mathbf{1}_{t\leq f(x)} : f \in \mathcal{F}\}$ with respect to the $\mathbb{L}^2(\widetilde{Q})$ metric leads to a $2\,\|G\|_{\mathbb{L}^2(Q)}\sqrt{\varepsilon}$-covering of the class $\mathcal{F}$ with respect to the $\mathbb{L}^2(Q)$ metric. By the previous result on VC dimension, we have the final bound.