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

Lecture 8: November 6, 2026

Lecturer: Wenlong Mou

1 Fat-shattering dimension

How about a function class with infinite VC dimension? In the worst case, an infinite VC dimension function class can shatter any finite set of points, and hence the covering number is unbounded. However, some classes still have structures, e.g.,

$$\mathcal{F} = \Big\{ f : [0,1] \to [0,1] \big| |f(x) - f(y)| \le |x - y|, \forall x, y \Big\}.$$

This is a classical example of nonparametric estimation, and we know that we can estimate functions in this class at a rate of $n^{-1/3}$. So probably there is some structure in this function class that allows us to bound its covering number. For the subgraph class, consider the grid $x_i = i/n$ with common threshold $t_i = 1/2$.

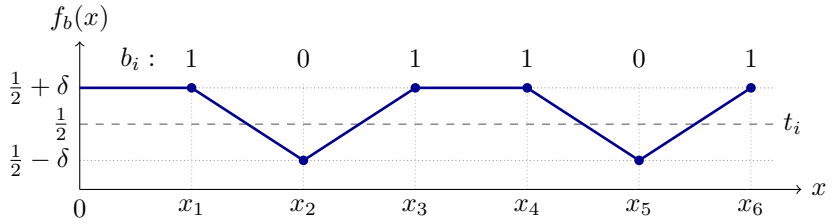


Figure 1. One labeling of the grid ($n = 6$, $x_6 = 1$). For any labeling b, set $f_b(x_i) = \frac{1}{2} + \delta(2b_i - 1)$ with $\delta = 1/(4n)$, interpolate linearly between grid points, and extend constantly on $[0, x_1]$. Each segment has absolute slope at most $2n\delta = 1/2$. The vertical scale is magnified.

Figure 1 shows how a 1-Lipschitz function realizes a labeling by placing its values above or below the common threshold $t_i = 1/2$. The same construction works for every labeling and every n, so the subgraph class has infinite VC dimension. The margin δ shrinks as the grid becomes finer. This motivates the following notion:

Definition 8.1 (ε-fat shattering)**.** *Consider a sequence $(x_i, t_i)_{i=1}^n$, it is ε-fat shattered by $\mathcal{F}$ if for every labeling $(b_1, \ldots, b_n) \in \{0,1\}^n$, there exists a function $f \in \mathcal{F}$ such that*

$$f(x_i) \begin{cases} \ge t_i + \varepsilon, & b_i = 1, \\ \le t_i, & b_i = 0. \end{cases}$$

Definition 8.2 (ε-fat shattering dimension)**.** *The ε-fat shattering dimension of a function class $\mathcal{F}$ is defined as the largest integer d such that there exists a set of d points in $\mathcal{X}$ and d values $\{t_1, \ldots, t_d\}$ that can be ε-fat shattered by $\mathcal{F}$. If no such largest integer exists, we say that the ε-fat shattering dimension is infinite. This is denoted as $\mathrm{fat}_\varepsilon(\mathcal{F})$.*

This can be understood as a scale-sensitive version of VC dimension.

1.1 Using fat-shattering dimension to bound covering/packing numbers

Theorem 8.1 (Mendelson–Vershynin). *If $\mathcal{F}$ is bounded by 1, then there exists universal constants $c_1, c_2 > 0$ such that for any $\varepsilon \in (0, 1]$, we have*

$$\sup_Q N(\mathcal{F}; \mathbb{L}^2(Q), \varepsilon) \le \left(\frac{c_1}{\varepsilon}\right)^{c_2 fat_{c_2\varepsilon}(\mathcal{F})}.$$

There is also an improved version of this result by Rudelson–Vershynin, which gives a better dependence on ε:

Theorem 8.2 (Rudelson–Vershynin). *Suppose that $\mathcal{F}$ is bounded by 1, and suppose that we have $fat_\varepsilon(\mathcal{F}) \le g(\varepsilon)$ for some function g with at most polynomial growth as $\varepsilon \to 0$. Then we have*

$$\sup_Q N(\mathcal{F}; \mathbb{L}^2(Q), \varepsilon) \le \exp\Big(c_1 g(c_2\varepsilon)\Big).$$

In such a case, we can apply the Dudley integral bound to get

$$\mathbb{E} \sup_{f\in\mathcal{F}} \big|(P_n - P)f\big| \le c_1 \int_0^1 \sqrt{\frac{g(c_2\varepsilon)}{n}} d\varepsilon.$$

- If VC dimension is finite, then $\mathrm{fat}_\varepsilon(\mathcal{F})$ is bounded by a constant, and we have the same bound as before.
- If $\mathrm{fat}_\varepsilon(\mathcal{F}) \le c\varepsilon^{-\alpha}$ for some $\alpha \in (0, 2)$, then we have

$$\mathbb{E} \sup_{f\in\mathcal{F}} \big|(P_n - P)f\big| \le c_1 \int_0^1 \sqrt{\frac{c(c_2\varepsilon)^{-\alpha}}{n}} d\varepsilon = c_1\sqrt{\frac{c}{n}} \int_0^1 \varepsilon^{-\alpha/2} d\varepsilon = c_1\sqrt{\frac{c}{n}} \cdot \frac{2}{2-\alpha}.$$

- If we can only show that $\mathrm{fat}_\varepsilon(\mathcal{F}) \le c\varepsilon^{-\alpha}$ for some $\alpha \ge 2$, then the integral diverges, and we can use a truncation argument to bound the empirical process supremum.

1.2 Examples

Consider the one-dimensional Lipschitz function class. Now suppose that $(x_i, t_i)_{i=1}^n$ is ε-fat shattered by this function class. Assume without loss of generality that $x_1 < x_2 < \cdots < x_n$. For any binary sequence $(b_1, \ldots, b_n) \in \{0, 1\}^n$, we can find a function f_b in the class such that

$$f_b(x_i) \begin{cases} \ge t_i + \varepsilon, & b_i = 1, \\ \le t_i, & b_i = 0. \end{cases}$$

Intuitively, two adjacent points x_i and x_{i+1} with different labels must have a gap of at least ε in their function values. If they get too close, then the Lipschitz condition will be violated.

Formally, we fix i, and let f_+ be the function that realizes the labeling $(b_1, \ldots, b_i = 0, b_{i+1} = 1, \ldots, b_n)$, and let f_- be the function that realizes the labeling $(b_1, \ldots, b_i = 1, b_{i+1} = 0, \ldots, b_n)$. Then we have

$$f_+(x_i) \le t_i, \quad f_+(x_{i+1}) \ge t_{i+1} + \varepsilon, \quad f_-(x_i) \ge t_i + \varepsilon, \quad f_-(x_{i+1}) \le t_{i+1}.$$

On the other hand, by the Lipschitz condition, we have

$$|f_+(x_i) - f_+(x_{i+1})| \le |x_i - x_{i+1}|, \quad |f_-(x_i) - f_-(x_{i+1})| \le |x_i - x_{i+1}|.$$

Combining these inequalities, we have

$$|x_i - x_{i+1}| \geq \max\left\{\varepsilon + t_{i+1} - t_i, \varepsilon + t_i - t_{i+1}\right\} = \varepsilon + |t_{i+1} - t_i| \geq \varepsilon.$$

So we have $n \leq 1/\varepsilon$, and hence $\mathrm{fat}_\varepsilon(\mathcal{F}) \leq 1/\varepsilon$. This gives a bound on the covering number of the Lipschitz function class, and hence a bound on the empirical process supremum.

2 Localization and convergence rates

Back to our original problem: convergence rate of M-estimators. We analyze

$$\widehat{\theta}_n := \arg\min_{\theta\in\Theta} F_n(\theta), \quad F_n(\theta) = \frac{1}{n}\sum_{i=1}^n f_\theta(X_i),$$

by comparing with the population version

$$\theta^* := \arg\min_{\theta\in\Theta} F(\theta), \quad F(\theta) = \mathbb{E} f_\theta(X).$$

We used the bound

$$F(\widehat{\theta}_n) - F(\theta^*) \leq 2\sup_{\theta\in\Theta}|F_n(\theta) - F(\theta)|. \tag{8.1}$$

But this bound may be loose in practice: consider a finite-dimensional parametric model Θ, leading to an empirical process bound

$$\mathbb{E}\sup_{\theta\in\Theta}|F_n(\theta) - F(\theta)| \leq c\sqrt{\frac{d}{n}}.$$

In many problems, assuming smoothness assumptions and non-degeneracy at θ^*, locally, we have

$$\mu\left\|\theta - \theta^*\right\|_2^2 \leq F(\theta) - F(\theta^*) \leq \beta\left\|\theta - \theta^*\right\|_2^2.$$

(e.g. you can consider logistic regression as a canonical example)

From this, the best bound we can derive is $\left\|\widehat{\theta}_n - \theta^*\right\|_2 \leq c(d/n)^{1/4}$. This is even worse than the classical asymptotic $1/\sqrt{n}$ convergence.

The key problem here is that the bound (8.1) takes a worst-case supremum over the entire parameter space Θ, which is too conservative. For our estimation problem, we know that $\widehat{\theta}_n$ is close to θ^*, and in the end, we are interested in the gap $F_n - F$ evaluated at $\widehat{\theta}_n$ and θ^*. So probably we can just take a supremum over $\Theta \cap \mathbb{B}(\theta^*, r)$ for some radius r that is shrinking with n.

At a superficial glance, this seems to be a chicken-egg problem: we want to bound $\left\|\widehat{\theta}_n - \theta^*\right\|_2$, but we need to know the radius r to take a supremum over. But this can be resolved by the following theorem.

Theorem 8.3 (Localization)**.** *Assume that F satisfies $F(\theta) - F(\theta^*) \geq \|\theta - \theta^*\|_2^2$ for all $\theta \in \Theta$. Suppose that we have shown an empirical process bound for every $u > 0$:*

$$\mathbb{E}\Big[\sup_{\theta\in\Theta\cap\mathbb{B}(\theta^*,u)}|(P_n - P)(f_\theta - f_{\theta^*})|\Big] \leq \phi_n(u), \tag{8.2}$$

with a function $\phi_n(u)$ satisfying that $\phi_n(cu) \le c^\alpha \phi_n(u)$ for every $u > 0$ and $c > 1$, with some $\alpha < 2$.

Then for r_n being the smallest positive number that satisfies the fixed-point relation

$$\phi_n(r_n) \le r_n^2, \tag{8.3}$$

for any $\delta > 0$, there exists a constant $c_\delta > 0$ (depending only on δ and α) such that

$$\mathbb{P}\Big(\Big\| \widehat{\theta}_n - \theta^* \Big\|_2 \ge c_\delta r_n \Big) \le \delta. \tag{8.4}$$

Parsing this theorem.

- We use the localized and shifted class $\{f_\theta - f_{\theta^*} : \theta \in \Theta \cap \mathbb{B}(\theta^*, u)\}$ instead of the original class $\{f_\theta : \theta \in \Theta\}$. Intuitively, if f is somewhat smooth with respect to θ, then the localized class should have smaller complexity than the original class.
- The convergence rate r_n is determined implicitly by the fixed-point relation $\phi_n(r_n) \le r_n^2$. This solution is known as the critical radius in the literature. The condition $\phi_n(cu) \le c^\alpha \phi_n(u)$ is a technical condition that ensures that the solution to the fixed-point relation is well-defined.
- The result is a finite-sample high-probability bound. There is no limiting arguments, and c_δ does not depend on n, r_n or any other problem parameters.

Proof: Starting with the error decomposition

$$\begin{aligned}
\Big\| \widehat{\theta}_n - \theta^* \Big\|_2^2 \le F(\widehat{\theta}_n) - F(\theta^*) &= F(\widehat{\theta}_n) - F_n(\widehat{\theta}_n) + \underbrace{F_n(\widehat{\theta}_n) - F_n(\theta^*)}_{\le 0} + F_n(\theta^*) - F(\theta^*) \\
&\le P f_{\widehat{\theta}_n} - P_n f_{\widehat{\theta}_n} + P_n f_{\theta^*} - P f_{\theta^*} \\
&= (P - P_n)(f_{\theta^*} - f_{\widehat{\theta}_n}).
\end{aligned}$$

We are interested in the tail probability

$$\mathbb{P}\Big(\Big\| \widehat{\theta}_n - \theta^* \Big\|_2 \ge 2^M r_n \Big) = \sum_{j=M}^{\infty} \mathbb{P}\Big(2^j r_n \le \Big\| \widehat{\theta}_n - \theta^* \Big\|_2 < 2^{j+1} r_n \Big).$$

Because of the upper limit in each term, we can control the empirical process supremum only by looking at a local neighborhood of θ^*.

For each term in the summation, we have

$$\begin{aligned}
&\mathbb{P}\Big(2^j r_n \le \Big\| \widehat{\theta}_n - \theta^* \Big\|_2 < 2^{j+1} r_n \Big) \\
&\le \mathbb{P}\Big(\Big\| \widehat{\theta}_n - \theta^* \Big\|_2 < 2^{j+1} r_n,\ |(P_n - P)(f_{\theta^*} - f_{\widehat{\theta}_n})| \ge 2^{2j} r_n^2 \Big) \\
&\le \mathbb{P}\Big(\sup_{\theta \in \Theta \cap \mathbb{B}(\theta^*, 2^{j+1} r_n)} |(P_n - P)(f_\theta - f_{\theta^*})| \ge 2^{2j} r_n^2 \Big) \\
&\le \frac{1}{2^{2j} r_n^2} \mathbb{E}\Big[\sup_{\theta \in \Theta \cap \mathbb{B}(\theta^*, 2^{j+1} r_n)} |(P_n - P)(f_\theta - f_{\theta^*})| \Big] \\
&\le \frac{1}{2^{2j} r_n^2} \phi(2^{j+1} r_n) \\
&\le \frac{\phi(r_n)}{r_n^2} \cdot 2^{(j+1)\alpha - 2j} \le 4 \cdot 2^{-(2-\alpha)j}.
\end{aligned}$$

Since $\alpha < 2$, we can sum over $j \geq M$ to obtain

$$\mathbb{P}\Big(\Big\| \widehat{\theta}_n - \theta^* \Big\|_2 \geq 2^M r_n \Big) \leq 4 \cdot \frac{2^{-(2-\alpha)M}}{1 - 2^{-(2-\alpha)}}.$$

We can choose M sufficiently large so that the right-hand-side is less than δ, and we have the desired result.

Additional remark: here the dependence on δ comes from Markov inequality, and hence is not optimal. It does not even guarantee a finite mean-squared error. It can only give high-probability bounds. To get better dependence on δ, we can use Talagrand's concentration inequality to get a sharper bound. On the other hand, these concentration inequalities may require additional assumptions on the function class, e.g., boundedness or sub-Gaussian tails.

2.1 Examples

Regular M-estimators: $\Theta = \mathbb{R}^d$, and the loss function satisfies a stochastic Lipschitz assumption:

$$f_{\theta_1}(x) - f_{\theta_2}(x) \leq M(x) \left\| \theta_1 - \theta_2 \right\|_2, \quad \mathbb{E}[M^2(X)] < \infty.$$

Assume furthermore that $F(\theta) - F(\theta^*) \geq \|\theta - \theta^*\|_2^2$. In order to apply the theorem, the key step is to compute the expected supremum of the localized empirical process:

$$\mathbb{E}\Big[\sup_{\|\theta - \theta^*\|_2 \leq u} |(P_n - P)(f_\theta - f_{\theta^*})| \Big].$$

For this quantity, we can use the Dudley integral method. First, we note that for any $\theta \in \mathbb{B}(\theta^*, u)$, we have

$$|f_\theta(X) - f_{\theta^*}(X)| \leq M(X) \left\| \theta - \theta^* \right\|_2 \leq M(X) u.$$

So we can take $uM(X)$ as an envelope function. Defining the localized function class $\mathcal{F}_u = \{f_\theta - f_{\theta^*} : \|\theta - \theta^*\|_2 \leq u\}$, we can apply Dudley's integral bound to get

$$\begin{aligned} &\mathbb{E}\Big[\sup_{\|\theta - \theta^*\|_2 \leq u} |(P_n - P)(f_\theta - f_{\theta^*})| \Big] \\ &\leq \frac{cu}{\sqrt{n}} \left\| M \right\|_{\mathbb{L}^2(P)} \cdot \int_0^1 \sqrt{\log \sup_Q N\big(\mathcal{F}_u; \mathbb{L}^2(Q), \varepsilon u \left\| M \right\|_{\mathbb{L}^2(Q)}\big)} d\varepsilon. \end{aligned}$$

Now for the covering number, we have done this before. Indeed, for $\theta_1, \theta_2 \in \mathbb{B}(\theta^*, u)$, we have

$$\left\| f_{\theta_1} - f_{\theta_2} \right\|_{\mathbb{L}^2(Q)} \leq \left\| M \right\|_{\mathbb{L}^2(Q)} \left\| \theta_1 - \theta_2 \right\|_2,$$

so that we get

$$N\big(\mathcal{F}_u; \mathbb{L}^2(Q), \varepsilon u \left\| M \right\|_{\mathbb{L}^2(Q)}\big) \leq N\big(\mathbb{B}(\theta^*, u); \left\| \cdot \right\|_2, \varepsilon u\big) \leq (c/\varepsilon)^d,$$

for some universal constant $c > 0$. Plugging this into the Dudley integral, we have

$$\mathbb{E}\Big[\sup_{\|\theta - \theta^*\|_2 \leq u} |(P_n - P)(f_\theta - f_{\theta^*})| \Big] \leq cu \cdot \sqrt{d/n} \cdot \left\| M \right\|_{\mathbb{L}^2(P)} =: \phi_n(u).$$

This is linearly growing with u, and hence $\alpha = 1 < 2$. The fixed-point relation becomes

$$cr_n \cdot \sqrt{d/n} \cdot \|M\|_{\mathbb{L}^2(P)} \le r_n^2,$$

where the smallest positive solution is $r_n = c\sqrt{d/n} \cdot \|M\|_{\mathbb{L}^2(P)}$. So we can conclude that

$$\mathbb{P}\Big(\left\|\widehat{\theta}_n - \theta^*\right\|_2 \ge c_\delta \sqrt{d/n} \cdot \|M\|_{\mathbb{L}^2(P)} \Big) \le \delta,$$

where the constant c_δ depends only on δ.

This recovers the classical $1/\sqrt{n}$ convergence rate of regular M-estimators. Also note that compared to classical results derived from Taylor expansion and asymptotic normality, this result has two advantages:

- It is purely non-asymptotic, holding true for any finite n.
- It requires less regularity conditions. We only need a stochastic Lipschitz condition, and no second-order smoothness is required.

e.g., we can consider $f_\theta(x) = \|\theta - x\|_2$, which is not differentiable at $\theta = x$. The classical Taylor expansion argument does not work here, but the localization argument still works.

A "non-regular" M-estimator and its convergence rate: Let us consider $\Theta \in \mathbb{R}$ and

$$f_\theta(x) = -\mathbf{1}_{x \in [\theta-1, \theta+1]}.$$

This can be seen as a soft version of the "mode" of the distribution: we try to find the interval of length 2 that contains the most probability mass.

Population version:

$$\theta^* = \arg\max_{\theta \in \mathbb{R}} \mathbb{P}(X \in [\theta - 1, \theta + 1]).$$

At the empirical level, the estimator becomes

$$\widehat{\theta}_n = \arg\max_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{X_i \in [\theta-1, \theta+1]}.$$

For this problem, the loss function is non-differentiable and discontinuous in θ, but we may still be able to analyze its convergence rate using the localization argument.

To make the argument work, we still need the population-level quantities to be well-behaved. In particular, we need $F(\theta) := -\mathbb{P}(X \in [\theta - 1, \theta + 1])$ to be locally strongly convex and smooth around θ^*. Assuming that X has a density $p(x)$ which is differentiable, straightforward calculation shows that

$$\frac{d^2}{d\theta^2} F(\theta^*) = p'(\theta^* - 1) - p'(\theta^* + 1).$$

We assume that this is positive. By further assuming continuity of p', we can find a neighborhood around θ^* such that $F(\theta)$ is strongly convex and smooth in that neighborhood. (by consistency of $\widehat{\theta}_n$, we can restrict our attention to this neighborhood with high probability).

It remains to study

$$\mathbb{E}\Big[\sup_{|\theta-\theta^*|\le u}|(P_n-P)(f_\theta-f_{\theta^*})|\Big].$$

We need to construct an envelope function for the localized class $\{f_\theta - f_{\theta^*} : |\theta-\theta^*|\le u\}$. $G(x)\equiv 1$ is a trivial envelope function, but it is too loose. We can do better by noting that for any $\theta\in[\theta^*-u,\theta^*+u]$, we have

$$|f_\theta(x)-f_{\theta^*}(x)| = \mathbf{1}_{x\in[\theta-1,\theta+1]\Delta[\theta^*-1,\theta^*+1]} \le \mathbf{1}_{x\in[\theta^*-1-u,\theta^*-1+u]}\vee \mathbf{1}_{x\in[\theta^*+1-u,\theta^*+1+u]} =: G_u(x).$$

By Dudley's integral bound, we have

$$\mathbb{E}\Big[\sup_{|\theta-\theta^*|\le u}|(P_n-P)(f_\theta-f_{\theta^*})|\Big] \le c\sqrt{\frac{\mathbb{E}[G_u(X)^2]}{n}}\cdot\int_0^1\sqrt{\log\sup_Q N\big(\mathcal{F}_u;\mathbb{L}^2(Q),\varepsilon\,\|G_u\|_{\mathbb{L}^2(Q)}\big)}d\varepsilon.$$

We note that

- For the envelope function, we have

  $$\mathbb{E}[G_u(X)^2] = \mathbb{P}(X\in[\theta^*-1-u,\theta^*-1+u]\cup[\theta^*+1-u,\theta^*+1+u]) \le 4p_{\max}u,$$

  by assuming that the density $p(x)$ is bounded by $p_{\max}$.

- For the integral, we note that the VC dimension of intervals is 2, and therefore the VC dimension of the localized class $\mathcal{F}_u$ is at most 8. So we have:

  $$\log\sup_Q N\big(\mathcal{F}_u;\mathbb{L}^2(Q),\varepsilon\,\|G_u\|_{\mathbb{L}^2(Q)}\big)\le c'\log(1/\varepsilon).$$

Putting them together, we have

$$\mathbb{E}\Big[\sup_{|\theta-\theta^*|\le u}|(P_n-P)(f_\theta-f_{\theta^*})|\Big]\le c\sqrt{\frac{p_{\max}u}{n}} =: \phi_n(u),$$

which is growing at a rate $\alpha=1/2<2$. The fixed-point relation becomes

$$c\sqrt{\frac{p_{\max}r_n}{n}} \le (p'(\theta^*-1)-p'(\theta^*+1))r_n^2.$$

Solving this gives $r_n = c'n^{-1/3}$, which implies that

$$|\widehat{\theta}_n-\theta^*|\le c_\delta n^{-1/3},$$

with probability at least $1-\delta$.

Indeed, we know that this rate of convergence is tight for this estimator. It is known that we have

$$n^{1/3}(\widehat{\theta}_n-\theta^*)\xrightarrow{d}\text{ a functional of Brownian motion.}$$