Empirical Process Theory

SDS 391P.6, Spring 2026

Pratik Patil

These notes are a work in progress and are provided as-is for instructional purposes only. They are not (yet) at the level of a scholarly document. In particular, the notes draw from various sources and do not (yet) have sufficient references to the original sources. Additionally, almost surely the notes have errors and they are only probably approximately correct. The notes will be updated regularly as the course progresses. Last updated: 2026-03-30.

1 Motivation

In the previous lecture, we introduced empirical processes through the example of uniform Monte Carlo integration. Given i.i.d. samples $X_1, \dots, X_n \sim \mu$, we studied quantities of the form

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right|,$$

and we proved a uniform law of large numbers for Lipschitz functions on $[0, 1]$ by combining two ingredients:

- Dudley's chaining inequality for processes with sub-Gaussian increments;
- a covering number bound for the class of Lipschitz functions in the uniform norm.

That argument worked well in a one-dimensional geometric setting, but it also had a limitation: it controlled the empirical process using the *ambient metric* on the whole function class, namely the sup norm. For more complicated classes, especially classes of indicator functions, it is often much better to use the randomness of the sample itself.

The first main tool in this direction is *symmetrization*. It allows us to replace the empirical process by a Rademacher process built on the observed sample. Once we condition on the sample, we can bring back the machinery developed earlier for Rademacher processes and sub-Gaussian increments.

The second main tool is *VC dimension*, which measures the combinatorial richness of a Boolean class. VC dimension allows us to bound the number of distinct labelings that the class can produce on a finite sample. This turns metric entropy into a combinatorial quantity.

The goal of this lecture is to connect these ideas:

$$\text{empirical process} \longrightarrow \text{symmetrization} \longrightarrow \text{Rademacher complexity} \longrightarrow \text{VC dimension}$$

and then use them to obtain a first VC-based bound on the uniform empirical error.

2 Empirical measures and empirical processes

Let μ be a probability measure on a measurable space Ω, and let

$$X_1, \dots, X_n \sim \mu \qquad \text{independently.}$$

The associated empirical measure is

$$\mu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i},$$

where δ_x denotes the Dirac mass at x.

For any measurable function $f : \Omega \to \mathbb{R}$,

$$\mu_n(f) := \int f \, d\mu_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i), \qquad \mu(f) := \int f \, d\mu = \mathbb{E}f(X).$$

Thus $\mu_n(f)$ is the empirical average and $\mu(f)$ is the population average.

Given a class $\mathcal{F}$ of real-valued measurable functions on Ω, the basic quantity of interest is the uniform empirical error

$$\sup_{f\in\mathcal{F}} |\mu_n(f) - \mu(f)|.$$

This is the random error of approximating the true measure μ by the empirical measure μ_n, uniformly over the function class.

Definition 2.1 (Empirical process)**.** Given a class $\mathcal{F}$ of measurable functions on Ω, define

$$Z_f := \mu_n(f) - \mu(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X), \qquad f \in \mathcal{F}.$$

The random family $(Z_f)_{f\in\mathcal{F}}$ is called the *empirical process* indexed by $\mathcal{F}$.

So the expected uniform empirical error is simply

$$\mathbb{E}\sup_{f\in\mathcal{F}} |Z_f|.$$

3 Symmetrization

A direct analysis of the empirical process is often awkward because the population mean $\mu(f) = \mathbb{E}f(X)$ is deterministic while the empirical term depends on the sample. The standard way to decouple them is to introduce a *ghost sample*.

Let $X_1', \dots, X_n'$ be an independent copy of $X_1, \dots, X_n$, independent of everything else.

Theorem 3.1 (Giné–Zinn symmetrization)**.** *Let $\mathcal{F}$ be a class of measurable real-valued functions on Ω, and let $\varepsilon_1, \dots, \varepsilon_n$ be i.i.d. Rademacher random variables, independent of everything else. Then*

$$\mathbb{E}\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \big(f(X_i) - \mathbb{E}f(X)\big) \right| \leqslant 2\,\mathbb{E}\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(X_i) \right|.$$

Proof. Since X_i' has the same law as X_i,

$$\mathbb{E}f(X_i) = \mathbb{E}f(X_i').$$

Therefore

$$\mathbb{E}\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \big(f(X_i) - \mathbb{E}f(X_i')\big) \right| = \mathbb{E}\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \big(f(X_i) - \mathbb{E}f(X_i')\big) \right|.$$

Apply Jensen's inequality conditionally on $X_1, \dots, X_n$: since the map

$$(y_1, \dots, y_n) \mapsto \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - y_i) \right|$$

is convex, we obtain

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \big(f(X_i) - \mathbb{E} f(X_i')\big) \right| \leqslant \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \big(f(X_i) - f(X_i')\big) \right|.$$

Now the joint law of (X_i, X_i') is invariant under swapping the two coordinates. Hence

$$\left(\frac{1}{n} \sum_{i=1}^{n} \big(f(X_i) - f(X_i')\big) \right)_{f \in \mathcal{F}} \overset{d}{=} \left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \big(f(X_i) - f(X_i')\big) \right)_{f \in \mathcal{F}}.$$

Therefore

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \big(f(X_i) - f(X_i')\big) \right| = \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \big(f(X_i) - f(X_i')\big) \right|.$$

By the triangle inequality, we have

$$\left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \big(f(X_i) - f(X_i')\big) \right| \leqslant \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i') \right|.$$

Taking suprema over f, then expectations, and using that the two terms have the same distribution, yields the result. □

Symmetrization says that the empirical process is controlled by a Rademacher process built on the sample. This is the first place where *Rademacher complexity* naturally appears.

4 Rademacher complexity

Definition 4.1 (Empirical and expected Rademacher complexity)**.** Let $\mathcal{F}$ be a class of measurable real-valued functions on Ω. Given a sample $X_1, \dots, X_n$, define the empirical Rademacher complexity

$$\widehat{\mathfrak{R}}_n(\mathcal{F}) := \mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right|,$$

where the expectation is taken only over the Rademacher variables $\varepsilon_1, \dots, \varepsilon_n$.

The expected Rademacher complexity is

$$\mathfrak{R}_n(\mathcal{F}) := \mathbb{E}_{X_1, \dots, X_n} \widehat{\mathfrak{R}}_n(\mathcal{F}).$$

Theorem 3.1 can now be rewritten as

$$\mathbb{E} \sup_{f \in \mathcal{F}} |\mu_n(f) - \mu(f)| \leqslant 2\, \mathfrak{R}_n(\mathcal{F}).$$

So Rademacher complexity measures how large the empirical process can be.

4.1 Conditional sub-Gaussian increments

Condition on the observed sample $X_1, \dots, X_n$. Then the process

$$R_f := \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i), \qquad f \in \mathcal{F},$$

is a Rademacher process. Its increments are sub-Gaussian in a data-dependent metric.

Define the empirical L_2 pseudometric

$$\|f - g\|_{L_2(\mu_n)} := \left(\frac{1}{n} \sum_{i=1}^{n} (f(X_i) - g(X_i))^2 \right)^{1/2}.$$

Proposition 4.2 (Conditional chaining bound for Rademacher averages)**.** *Assume $0 \in \mathcal{F}$. Then*

$$\widehat{\mathfrak{R}}_n(\mathcal{F}) \leqslant \frac{C}{\sqrt{n}} \int_0^{\operatorname{diam}(\mathcal{F}, L_2(\mu_n))} \sqrt{\log \mathcal{N}(\mathcal{F}, L_2(\mu_n), \varepsilon)}\, d\varepsilon,$$

where $C > 0$ is an absolute constant.

Proof. Condition on the sample. For $f, g \in \mathcal{F}$,

$$R_f - R_g = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \big(f(X_i) - g(X_i) \big).$$

This is a centered Rademacher series, so by the fixed-series bound from earlier in the course,

$$\|R_f - R_g\|_{\psi_2} \leqslant \frac{C}{n} \left(\sum_{i=1}^{n} (f(X_i) - g(X_i))^2 \right)^{1/2} = \frac{C}{\sqrt{n}} \|f - g\|_{L_2(\mu_n)}.$$

Thus $(R_f)_{f \in \mathcal{F}}$ has sub-Gaussian increments with respect to the metric

$$d_n(f, g) := \frac{1}{\sqrt{n}} \|f - g\|_{L_2(\mu_n)}.$$

Since $0 \in \mathcal{F}$, Dudley's inequality gives

$$\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} |R_f - R_0| \leqslant C \int_0^{\operatorname{diam}(\mathcal{F}, d_n)} \sqrt{\log \mathcal{N}(\mathcal{F}, d_n, \delta)}\, d\delta.$$

Rescaling $\delta = \varepsilon / \sqrt{n}$ yields the result. □

This is a very general bound. The challenge is now to control the empirical covering numbers $\mathcal{N}(\mathcal{F}, L_2(\mu_n), \varepsilon)$.

It is worth pausing to reinterpret what we have done. The quantity

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right|$$

measures how much the class $\mathcal{F}$ can correlate with random signs. Large Rademacher complexity means the class is flexible enough to fit noise well. Small Rademacher complexity means the class is effectively simple from the sample's point of view.

So symmetrization turns the uniform empirical error into a geometric-combinatorial complexity problem. We will next control this complexity for Boolean classes using VC dimension.

5 Boolean classes

In this lecture we focus on Boolean classes, i.e. classes of functions $h : \Omega \to \{0,1\}$. Equivalently, we may identify such a class with a collection of subsets of Ω.

Let $\mathcal{H}$ be a class of Boolean functions on Ω. Given sample points $x_1, \dots, x_n \in \Omega$, define the set of *traces* of $\mathcal{H}$ on the sample by

$$\mathcal{H}|_{x_{1:n}} := \big\{\big(h(x_1), \dots, h(x_n)\big) : \ h \in \mathcal{H}\big\} \subseteq \{0,1\}^n.$$

So $\mathcal{H}|_{x_{1:n}}$ records all labelings that the class can realize on the sample.

For Boolean classes, the empirical L_∞ metric has a particularly simple structure. Indeed, for $h, g \in \mathcal{H}$,

$$\|h - g\|_{L_\infty(\mu_n)} = \max_{1 \leqslant i \leqslant n} |h(X_i) - g(X_i)|.$$

Since $h(X_i), g(X_i) \in \{0,1\}$, this distance is either 0 or 1. It equals 0 exactly when h and g agree on all sample points.

Therefore, for every $\varepsilon \in (0,1)$, we have

$$\mathcal{N}(\mathcal{H}, L_\infty(\mu_n), \varepsilon) = |\mathcal{H}|_{X_{1:n}}|.$$

So for Boolean classes, uniform covering numbers reduce to counting traces.

This motivates the following quantity.

Definition 5.1 (Growth function)**.** For a class $\mathcal{H}$ of Boolean functions, define its growth function by

$$\Pi_{\mathcal{H}}(n) := \sup_{x_1,\dots,x_n \in \Omega} |\mathcal{H}|_{x_{1:n}}|.$$

The growth function measures how many different binary labelings the class can realize on n points.

6 VC dimension

Definition 6.1 (Shattering)**.** Let $\mathcal{H}$ be a class of Boolean functions on Ω, and let $\Lambda = \{x_1, \dots, x_m\} \subseteq \Omega$ be finite. We say that $\mathcal{H}$ *shatters* Λ if every labeling of Λ is realized by some function in $\mathcal{H}$. Equivalently, we have

$$|\mathcal{H}|_\Lambda| = 2^m.$$

Definition 6.2 (VC dimension)**.** The *VC dimension* of $\mathcal{H}$, denoted $\mathrm{VC}(\mathcal{H})$, is the largest integer d such that some subset of Ω of size d is shattered by $\mathcal{H}$. If arbitrarily large finite sets are shattered, we write $\mathrm{VC}(\mathcal{H}) = \infty$.

Examples.

- Half-lines on $\mathbb{R}$:
$$\mathcal{H} = \{\mathbf{1}_{(-\infty,a]} : \ a \in \mathbb{R}\} \qquad \Rightarrow \qquad \mathrm{VC}(\mathcal{H}) = 1.$$
- Intervals on $\mathbb{R}$:
$$\mathcal{H} = \{\mathbf{1}_{[a,b]} : \ a \leqslant b\} \qquad \Rightarrow \qquad \mathrm{VC}(\mathcal{H}) = 2.$$

- Half-spaces in $\mathbb{R}^d$:

$$\mathcal{H} = \left\{ \mathbf{1}_{\{\langle w,x\rangle + b \geqslant 0\}} : \ w \in \mathbb{R}^d, \ b \in \mathbb{R} \right\} \qquad \Rightarrow \qquad \mathrm{VC}(\mathcal{H}) = d + 1.$$

The VC dimension is a combinatorial measure of complexity. The key fact is that finite VC dimension forces the growth function $\Pi_{\mathcal{H}}(n)$ to be polynomial rather than exponential in n.

7 Pajor lemma and Sauer–Shelah theorem

We now state the main combinatorial result.

Lemma 7.1 (Pajor)**.** *Let $\mathcal{H}$ be a finite class of Boolean functions on a finite set Ω. Then*

$$|\mathcal{H}| \leqslant \#\{\Lambda \subseteq \Omega : \ \Lambda \text{ is shattered by } \mathcal{H}\}.$$

Proof. We argue by induction on $|\Omega|$. The case $|\Omega| = 0$ is trivial.

Assume the result holds for sets of size $n-1$, and let $|\Omega| = n$. Pick $x_0 \in \Omega$, and split the class into

$$\mathcal{H}_0 := \{h \in \mathcal{H} : h(x_0) = 0\}, \qquad \mathcal{H}_1 := \{h \in \mathcal{H} : h(x_0) = 1\}.$$

By induction,

$$|\mathcal{H}_0| \leqslant |\operatorname{Sh}(\mathcal{H}_0)|, \qquad |\mathcal{H}_1| \leqslant |\operatorname{Sh}(\mathcal{H}_1)|,$$

where $\operatorname{Sh}(\cdot)$ denotes the family of shattered subsets of $\Omega \backslash \{x_0\}$.

Now any set shattered by $\mathcal{H}_0$ or by $\mathcal{H}_1$ is certainly shattered by $\mathcal{H}$. Moreover, if a set $\Lambda \subseteq \Omega \backslash \{x_0\}$ is shattered by both $\mathcal{H}_0$ and $\mathcal{H}_1$, then $\Lambda \cup \{x_0\}$ is shattered by $\mathcal{H}$: to realize any labeling on $\Lambda \cup \{x_0\}$, use $\mathcal{H}_0$ if the label at x_0 is 0, and $\mathcal{H}_1$ if it is 1.

Therefore,

$$\#\operatorname{Sh}(\mathcal{H}) \geqslant \#\operatorname{Sh}(\mathcal{H}_0) + \#\operatorname{Sh}(\mathcal{H}_1),$$

and hence

$$|\mathcal{H}| = |\mathcal{H}_0| + |\mathcal{H}_1| \leqslant \#\operatorname{Sh}(\mathcal{H}).$$

□

Pajor's lemma immediately yields Sauer–Shelah.

Theorem 7.2 (Sauer–Shelah)**.** *Let $\mathcal{H}$ be a class of Boolean functions with VC dimension $d < \infty$. Then for every $n \geqslant d$,*

$$\Pi_{\mathcal{H}}(n) \leqslant \sum_{k=0}^{d} \binom{n}{k} \leqslant \left(\frac{en}{d}\right)^d.$$

Proof. Fix $x_1, \ldots, x_n \in \Omega$. Apply Pajor's lemma to the restricted class $\mathcal{H}|_{x_{1:n}}$, which is a finite class on an n-point set. Any shattered subset has cardinality at most $d = \mathrm{VC}(\mathcal{H})$, so the number of shattered subsets is at most

$$\sum_{k=0}^{d} \binom{n}{k}.$$

Therefore

$$|\mathcal{H}|_{x_{1:n}}| \leqslant \sum_{k=0}^{d} \binom{n}{k}.$$

Taking the supremum over all $x_1, \dots, x_n$ gives the first inequality.

The second inequality is the standard combinatorial estimate

$$\sum_{k=0}^{d} \binom{n}{k} \leqslant \Big(\frac{en}{d}\Big)^d \qquad (n \geqslant d).$$

□

So VC dimension controls the growth function, and therefore the trace complexity of the class on any sample.

8 A VC-based bound for empirical processes

We now combine everything.

Theorem 8.1 (Preliminary VC bound for the uniform empirical error)**.** *Let $\mathcal{H}$ be a class of Boolean functions on Ω, and assume*

$$d := \mathrm{VC}(\mathcal{H}) < \infty.$$

Let $X_1, \dots, X_n$ be i.i.d. with law μ, and assume $n \geqslant d$. Then

$$\mathbb{E} \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} h(X_i) - \mathbb{E}h(X) \right| \leqslant C \sqrt{\frac{d \log(en/d)}{n}},$$

and in particular

$$\mathbb{E} \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} h(X_i) - \mathbb{E}h(X) \right| \lesssim \sqrt{\frac{d \log n}{n}}.$$

Proof. By symmetrization,

$$\mathbb{E} \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} h(X_i) - \mathbb{E}h(X) \right| \leqslant 2\, \mathfrak{R}_n(\mathcal{H}).$$

Condition on the sample $X_1, \dots, X_n$. The process

$$R_h := \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(X_i), \qquad h \in \mathcal{H},$$

has sub-Gaussian increments with respect to

$$\frac{1}{\sqrt{n}} \|h - g\|_{L_\infty(\mu_n)}.$$

Indeed, since $|h(X_i) - g(X_i)| \leqslant \|h - g\|_{L_\infty(\mu_n)}$, the fixed-series sub-Gaussian bound gives

$$\|R_h - R_g\|_{\psi_2} \leqslant \frac{C}{\sqrt{n}} \|h - g\|_{L_\infty(\mu_n)}.$$

Therefore, by Dudley's inequality,

$$\widehat{\mathfrak{R}}_n(\mathcal{H}) \leqslant \frac{C}{\sqrt{n}} \int_0^{\operatorname{diam}(\mathcal{H}, L_\infty(\mu_n))} \sqrt{\log \mathcal{N}(\mathcal{H}, L_\infty(\mu_n), \varepsilon)}\, d\varepsilon.$$

Since $\mathcal{H} \subset \{0,1\}^\Omega$, the diameter in $L_\infty(\mu_n)$ is at most 1. Moreover, for every $0 < \varepsilon < 1$,

$$\mathcal{N}(\mathcal{H}, L_\infty(\mu_n), \varepsilon) = |\mathcal{H}|_{X_{1:n}}| \leqslant \Pi_{\mathcal{H}}(n).$$

So

$$\widehat{\mathfrak{R}}_n(\mathcal{H}) \leqslant \frac{C}{\sqrt{n}} \int_0^1 \sqrt{\log \Pi_{\mathcal{H}}(n)}\, d\varepsilon = \frac{C}{\sqrt{n}} \sqrt{\log \Pi_{\mathcal{H}}(n)}.$$

Taking expectation over the sample does not change the right-hand side, so

$$\mathfrak{R}_n(\mathcal{H}) \leqslant \frac{C}{\sqrt{n}} \sqrt{\log \Pi_{\mathcal{H}}(n)}.$$

Now use Sauer–Shelah:

$$\Pi_{\mathcal{H}}(n) \leqslant \left(\frac{en}{d}\right)^d.$$

Hence

$$\log \Pi_{\mathcal{H}}(n) \leqslant d \log(en/d),$$

and therefore

$$\mathfrak{R}_n(\mathcal{H}) \leqslant C \sqrt{\frac{d \log(en/d)}{n}}.$$

Finally multiply by 2 using symmetrization. □

This is already a strong result: a purely combinatorial quantity, VC dimension, controls the size of the empirical process.

Remark 8.2 (Sharper VC bounds)**.** The bound in Theorem 8.1 is not optimal. The logarithmic factor appears because we passed to the stronger metric $L_\infty(\mu_n)$. A finer argument, using $L_2(\mu_n)$ covering numbers together with a deeper entropy estimate for VC classes, yields the sharp order

$$\mathbb{E} \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n h(X_i) - \mathbb{E}h(X) \right| \lesssim \sqrt{\frac{d}{n}}.$$

We will not prove that sharper estimate today.

9 Application: Glivenko–Cantelli theorem

Let X be a real random variable with cumulative distribution function

$$F(a) := \mathbb{P}\{X \leqslant a\}, \qquad a \in \mathbb{R}.$$

Given a sample $X_1, \dots, X_n$, define the empirical CDF

$$F_n(a) := \frac{1}{n} \#\{i : \ X_i \leqslant a\}.$$

This corresponds to the class of half-lines

$$\mathcal{H} = \{\mathbf{1}_{(-\infty, a]} : \ a \in \mathbb{R}\},$$

which has VC dimension 1.

Applying Theorem 8.1 gives

$$\mathbb{E} \sup_{a \in \mathbb{R}} |F_n(a) - F(a)| \lesssim \sqrt{\frac{\log n}{n}}.$$

This is already a uniform law of large numbers for distribution functions. The sharp rate is $n^{-1/2}$, and indeed a sharper argument later removes the extra $\sqrt{\log n}$.

10 Look ahead

This lecture introduced the first core tools of empirical process theory. We rewrote the uniform empirical error as the supremum of an empirical process, used symmetrization to compare it with a Rademacher process, and then interpreted the resulting quantity through empirical Rademacher complexity. For Boolean classes, the complexity of the class on a sample reduced to counting traces, and VC dimension entered naturally through shattering and the Sauer–Shelah theorem. Combining these ingredients produced a first VC-based uniform law of large numbers.

The next lecture will turn these ideas toward statistical learning theory. There, the same symmetrization and VC arguments will control uniform deviations of empirical risk from population risk, leading to generalization bounds and sample-complexity guarantees.

Source material

Parts of this lecture are based on references: Vershynin (2018); Tropp (2023), in addition to the author's accumulated experience working on related topics.

References

Tropp, J. A. (2023). Probability in high dimensions. Caltech CMS Lecture Notes 2021-01.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.