Chaining II

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: 2026-03-22.

1 Motivation

In the previous lecture, we introduced chaining as a multiscale refinement of the ε-net method. The main result was Dudley's entropy integral bound, which controls the expected supremum of a Gaussian process in terms of covering numbers of its index set.

There are two important lessons to carry forward.

First, the proof of Dudley's inequality did not really use any deep structural property of Gaussian random variables beyond one basic input: the increments are sub-Gaussian. So the method extends immediately to a wider class of processes.

Second, Dudley's inequality is not just a statement about Gaussian width. It is a general tool for proving *uniform* stochastic bounds over large classes of functions. One of the most important places where this appears is in empirical process theory, where one studies quantities of the form

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

This is exactly the kind of supremum that appears in uniform laws of large numbers, learning theory, and nonparametric statistics.

The goal of this lecture is twofold. First, we record several useful variants and extensions of Dudley's inequality. Second, we apply these ideas to a first empirical-process result: a uniform law of large numbers for Lipschitz functions on the unit interval.

2 Recap: Dudley's inequality

Let $(X_t)_{t \in T}$ be a random process on a metric space (T, d). Recall that we say the process has *sub-Gaussian increments* with constant K if

$$\|X_t - X_s\|_{\psi_2} \leqslant K\, d(t, s) \qquad \text{for all } s, t \in T.$$

The form of Dudley's theorem that is most useful in applications is the following increment version.

Theorem 2.1 (Dudley inequality)**.** *Let* $(X_t)_{t\in T}$ *be a random process on a metric space* (T,d) *with sub-Gaussian increments:*

$$\|X_t - X_s\|_{\psi_2} \leqslant K\, d(t,s) \qquad \text{for all } s,t \in T.$$

Then for every fixed $t_0 \in T$*,*

$$\mathbb{E}\sup_{t\in T}(X_t - X_{t_0}) \;\leqslant\; CK \int_0^{\operatorname{diam}(T,d)} \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\, d\varepsilon,$$

where $C > 0$ *is an absolute constant.*

This is slightly stronger than the centered formulation from the previous lecture. Indeed, the chaining proof directly bounds $\sup_t (X_t - X_{t_0})$, so no centering assumption is really needed.

3 Useful variants of Dudley's inequality

3.1 Suprema of increments

A very common quantity is not $\sup_t X_t$, but rather the oscillation of the process:

$$\sup_{s,t\in T} |X_t - X_s|.$$

Dudley's inequality immediately yields a bound for this as well.

Corollary 3.1 (Supremum of increments)**.** *Under the assumptions of Theorem 2.1,*

$$\mathbb{E}\sup_{s,t\in T} |X_t - X_s| \;\leqslant\; 2CK \int_0^{\operatorname{diam}(T,d)} \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\, d\varepsilon.$$

Proof. Fix $t_0 \in T$. For any $s,t \in T$,

$$X_t - X_s = (X_t - X_{t_0}) - (X_s - X_{t_0}),$$

so

$$|X_t - X_s| \leqslant |X_t - X_{t_0}| + |X_s - X_{t_0}|.$$

Taking suprema and expectations gives

$$\mathbb{E}\sup_{s,t\in T} |X_t - X_s| \leqslant 2\,\mathbb{E}\sup_{t\in T} |X_t - X_{t_0}|.$$

Now apply Theorem 2.1 to the processes $(X_t)_{t\in T}$ and $(-X_t)_{t\in T}$, or equivalently to the positive and negative parts of the supremum. □

This form is often the most natural one for empirical processes, since the absolute value is built in from the start.

3.2 Diameter truncation

A small but useful observation is that the entropy integral automatically truncates at the diameter.

Remark 3.2 (Upper limit of the entropy integral)**.** If $\varepsilon > \mathrm{diam}(T, d)$, then a single ε-ball covers all of T, so

$$\mathcal{N}(T, d, \varepsilon) = 1 \qquad \text{and hence} \qquad \log \mathcal{N}(T, d, \varepsilon) = 0.$$

Therefore the integral in Dudley's inequality may always be written as

$$\int_0^{\mathrm{diam}(T,d)} \sqrt{\log \mathcal{N}(T, d, \varepsilon)}\, d\varepsilon.$$

3.3 A high-probability version

Dudley's theorem controls only the expectation of the supremum. A chaining argument also gives a tail bound. We record the result without proof.

Proposition 3.3 (High-probability Dudley bound)**.** *Let $(X_t)_{t\in T}$ be a random process on (T, d) with sub-Gaussian increments:*

$$\|X_t - X_s\|_{\psi_2} \leqslant K\, d(t, s) \qquad \textit{for all } s, t \in T.$$

Then there exist absolute constants $c, C > 0$ such that for every $u \geqslant 0$,

$$\mathbb{P}\left\{ \sup_{s,t\in T} |X_t - X_s| \geqslant CK \left[\int_0^{\mathrm{diam}(T,d)} \sqrt{\log \mathcal{N}(T, d, \varepsilon)}\, d\varepsilon + u\ \mathrm{diam}(T, d) \right] \right\} \leqslant 2e^{-cu^2}.$$

So Dudley's entropy integral describes the *typical size* of the oscillation, while the diameter controls the sub-Gaussian tail around that size.

3.4 Beyond sub-Gaussian increments

The same chaining proof adapts to weaker increment assumptions. Only the maximal inequality at each scale changes.

Proposition 3.4 (Sub-exponential increments)**.** *Suppose $(X_t)_{t\in T}$ satisfies*

$$\|X_t - X_s\|_{\psi_1} \leqslant K\, d(t, s) \qquad \textit{for all } s, t \in T.$$

Then

$$\mathbb{E} \sup_{t\in T} (X_t - X_{t_0}) \ \leqslant\ CK \int_0^{\mathrm{diam}(T,d)} \log \mathcal{N}(T, d, \varepsilon)\, d\varepsilon \qquad \textit{for every } t_0 \in T.$$

The only change is that the expected maximum of M sub-exponential variables scales like $\log M$, not $\sqrt{\log M}$.

There are also mixed-tail versions, where the increments have a Bernstein-type structure with one metric controlling the Gaussian regime and another controlling the exponential regime. These become useful later for empirical processes built from bounded random variables.

4 Examples and limitations

4.1 Gaussian width as an entropy integral

Let $g \sim \mathcal{N}(0, I_n)$, and let $T \subset \mathbb{R}^n$ be bounded. Recall the Gaussian width

$$w(T) := \mathbb{E} \sup_{t \in T} \langle g, t \rangle.$$

Since $t \mapsto \langle g, t \rangle$ is a Gaussian process with canonical metric

$$d(s, t) = \|s - t\|_2,$$

Dudley's inequality yields:

Corollary 4.1 (Entropy integral bound for Gaussian width)**.** *For every bounded $T \subset \mathbb{R}^n$,*

$$w(T) \;\leqslant\; C \int_0^{\mathrm{diam}(T, \|\cdot\|_2)} \sqrt{\log \mathcal{N}(T, \|\cdot\|_2, \varepsilon)}\, d\varepsilon.$$

Since Gaussian width is translation-invariant in expectation, one may always shift T so that $0 \in T$ when convenient.

4.2 Example: the Euclidean ball

Take $T = B_2^n$. Then

$$w(B_2^n) = \mathbb{E}\|g\|_2.$$

Using the volumetric covering estimate

$$\mathcal{N}(B_2^n, \varepsilon) \leqslant \left(\frac{3}{\varepsilon}\right)^n \qquad (0 < \varepsilon \leqslant 1),$$

we obtain

$$w(B_2^n) \lesssim \int_0^1 \sqrt{n \log(3/\varepsilon)}\, d\varepsilon \lesssim \sqrt{n}.$$

This is the correct order, since $\mathbb{E}\|g\|_2 \asymp \sqrt{n}$.

So for round sets like the Euclidean ball, Dudley's bound is sharp up to constants.

4.3 Dudley can be loose

Dudley's inequality is not always optimal. There are sets T for which the entropy integral overestimates the Gaussian width.

A standard example is the weighted orthonormal set

$$T_n := \left\{ \frac{e_k}{\sqrt{1 + \log k}} \;:\; 1 \leqslant k \leqslant n \right\} \subset \mathbb{R}^n.$$

Then

$$w(T_n) = \mathbb{E} \max_{1 \leqslant k \leqslant n} \frac{g_k}{\sqrt{1 + \log k}} \asymp 1,$$

but the entropy integral grows like $\log \log n$.

So Dudley can miss the truth by more than a constant factor. This is one reason generic chaining is needed later.

5 From random processes to empirical processes

We now turn to a first application of Dudley's inequality outside the Gaussian setting.

5.1 Monte Carlo integration

Let μ be a probability measure on a domain Ω, and let $X \sim \mu$. For an integrable function $f : \Omega \to \mathbb{R}$,

$$\int_\Omega f \, d\mu = \mathbb{E} f(X).$$

If $X_1, \ldots, X_n$ are i.i.d. copies of X, then the Monte Carlo estimator

$$\frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

approximates $\mathbb{E} f(X)$.

For a fixed f, this is just the law of large numbers. A more ambitious question is whether one sample $X_1, \ldots, X_n$ can approximate *simultaneously* the integrals of all functions in a class $\mathcal{F}$.

This leads naturally to empirical processes.

5.2 Empirical measure and empirical process

Define the empirical measure

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

where δ_x is the Dirac mass at x.

Then for any measurable f,

$$\mu_n(f) = \int f \, d\mu_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

So the approximation error is

$$\mu_n(f) - \mu(f).$$

Definition 5.1 (Empirical process)**.** Given a class $\mathcal{F}$ of real-valued 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 $\mathcal{F}$.

The quantity of interest is then

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

A bound that tends to zero with n is called a *uniform law of large numbers* over $\mathcal{F}$.

6 A uniform law of large numbers for Lipschitz functions

We now prove a simple but instructive example.

Theorem 6.1 (Uniform LLN for Lipschitz functions)**.** *Let μ be any probability measure on $[0,1]$, and let $X_1, \dots, X_n$ be i.i.d. with law μ. For $L > 0$, define the anchored Lipschitz class*

$$\mathcal{F}_L := \{ f : [0,1] \to \mathbb{R} : \ f(0) = 0, \ \|f\|_{\mathrm{Lip}} \leqslant L \}.$$

Then

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

where $C > 0$ is an absolute constant.

In other words, the theorem says that a single random sample $X_1, \dots, X_n$ can simultaneously approximate

$$\mathbb{E} f(X) = \int f \, d\mu$$

for *all* Lipschitz functions in the class $\mathcal{F}_L$, with the same $n^{-1/2}$ rate that one expects for a single fixed function.

So the law of large numbers becomes *uniform* over a whole function class without losing the root-n rate. That is the key phenomenon.

Remark 6.2. The anchoring condition $f(0) = 0$ is only for convenience. Indeed, for any constant c,

$$\mu_n(f + c) - \mu(f + c) = \mu_n(f) - \mu(f),$$

so adding constants does not affect the empirical process. Thus the same theorem can be read as a uniform law of large numbers over all Lipschitz functions modulo constants.

The proof has two ingredients:

- a covering number bound for the Lipschitz class in the sup norm;
- an application of Dudley's inequality to the empirical process.

6.1 Covering Lipschitz functions

We first prove the entropy estimate.

Proposition 6.3 (Covering Lipschitz functions)**.** *Let*

$$\mathcal{F}_1 := \{ f : [0,1] \to \mathbb{R} : \ f(0) = 0, \ \|f\|_{\mathrm{Lip}} \leqslant 1 \}.$$

Then for every $\varepsilon \in (0,1]$,

$$\log \mathcal{N}(\mathcal{F}_1, \|\cdot\|_\infty, \varepsilon) \ \leqslant \ \frac{C}{\varepsilon},$$

where $C > 0$ is an absolute constant.

Proof. Fix $\varepsilon \in (0,1]$, and let

$$m := \lceil 1/\varepsilon \rceil, \qquad x_j := \frac{j}{m}, \qquad j = 0, 1, \dots, m.$$

Then the mesh size is

$$\delta := x_{j+1} - x_j = \frac{1}{m} \leqslant \varepsilon.$$

Take any $f \in \mathcal{F}_1$. Since $f(0) = 0$ and $\|f\|_{\mathrm{Lip}} \leqslant 1$, we have

$$|f(x)| \leqslant x \leqslant 1 \qquad \text{for all } x \in [0,1].$$

For each grid point x_j, let q_j be the nearest multiple of ε to $f(x_j)$. Then

$$q_j \in \varepsilon\mathbb{Z} \cap [-1,1], \qquad |f(x_j) - q_j| \leqslant \frac{\varepsilon}{2}.$$

Since $q_0 = 0$, the sequence (q_j) starts at a fixed value.

Now define a piecewise constant approximation g by

$$g(x) := q_j \qquad \text{for } x \in [x_j, x_{j+1}), \qquad g(1) := q_m.$$

If $x \in [x_j, x_{j+1})$, then

$$|f(x) - g(x)| \leqslant |f(x) - f(x_j)| + |f(x_j) - q_j| \leqslant \delta + \frac{\varepsilon}{2} \leqslant \frac{3\varepsilon}{2}.$$

Thus

$$\|f - g\|_\infty \leqslant \frac{3\varepsilon}{2}.$$

It remains to count how many such approximants g can occur. Because f is 1-Lipschitz and $\delta \leqslant \varepsilon$,

$$|f(x_{j+1}) - f(x_j)| \leqslant \delta \leqslant \varepsilon.$$

Hence

$$|q_{j+1} - q_j| \leqslant |q_{j+1} - f(x_{j+1})| + |f(x_{j+1}) - f(x_j)| + |f(x_j) - q_j| \leqslant \frac{\varepsilon}{2} + \varepsilon + \frac{\varepsilon}{2} = 2\varepsilon.$$

Since the q_j's lie on the ε-grid, once q_j is fixed, there are at most five possibilities for q_{j+1}:

$$q_j - 2\varepsilon,\ q_j - \varepsilon,\ q_j,\ q_j + \varepsilon,\ q_j + 2\varepsilon.$$

Therefore the number of possible sequences $(q_0, \ldots, q_m)$ is at most 5^m. So

$$\mathcal{N}(\mathcal{F}_1, \|\cdot\|_\infty, 2\varepsilon) \leqslant 5^m,$$

and hence

$$\log \mathcal{N}(\mathcal{F}_1, \|\cdot\|_\infty, 2\varepsilon) \leqslant m \log 5 \leqslant \frac{C}{\varepsilon}.$$

After adjusting constants and replacing 2ε by ε, the result follows. □

6.2 Proof of the uniform LLN

We now prove Theorem 6.1.

Proof of Theorem 6.1. By scaling, it suffices to consider the case $L = 1$. So let $\mathcal{F}_1$ be as in Proposition 6.3, and define the empirical process

$$Z_f := \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X), \qquad f \in \mathcal{F}_1.$$

We first check sub-Gaussian increments with respect to the metric

$$d(f,g) := \frac{1}{\sqrt{n}}\|f-g\|_\infty.$$

Indeed,

$$Z_f - Z_g = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad Y_i := (f-g)(X_i) - \mathbb{E}[(f-g)(X)].$$

The random variables $Y_1, \dots, Y_n$ are independent, mean zero, and satisfy

$$|Y_i| \leqslant |(f-g)(X_i)| + \mathbb{E}|(f-g)(X)| \leqslant 2\|f-g\|_\infty.$$

Therefore, by Hoeffding's inequality (or Hoeffding's lemma),

$$\|Z_f - Z_g\|_{\psi_2} \leqslant C\,\frac{\|f-g\|_\infty}{\sqrt{n}} = C\,d(f,g).$$

So $(Z_f)_{f\in\mathcal{F}_1}$ is a process with sub-Gaussian increments.

Since the zero function belongs to $\mathcal{F}_1$, we have

$$\sup_{f\in\mathcal{F}_1}|Z_f| \leqslant \sup_{f,g\in\mathcal{F}_1}(Z_f - Z_g).$$

Applying Corollary 3.1,

$$\mathbb{E}\sup_{f\in\mathcal{F}_1}|Z_f| \leqslant \frac{C}{\sqrt{n}}\int_0^{\operatorname{diam}(\mathcal{F}_1,\|\cdot\|_\infty)} \sqrt{\log\mathcal{N}(\mathcal{F}_1,\|\cdot\|_\infty,\varepsilon)}\,d\varepsilon.$$

Because $f(0) = 0$ and $\|f\|_{\mathrm{Lip}} \leqslant 1$, every $f \in \mathcal{F}_1$ satisfies $|f(x)| \leqslant 1$, so

$$\operatorname{diam}(\mathcal{F}_1,\|\cdot\|_\infty) \leqslant 2.$$

Using Proposition 6.3,

$$\mathbb{E}\sup_{f\in\mathcal{F}_1}|Z_f| \leqslant \frac{C}{\sqrt{n}}\int_0^2 \sqrt{\frac{C'}{\varepsilon}}\,d\varepsilon \leqslant \frac{C''}{\sqrt{n}}.$$

This proves the theorem for $L = 1$. Multiplying by L yields the general case. □

6.3 A warning from higher dimensions

The one-dimensional Lipschitz class is still relatively tame. In higher dimensions, the same strategy becomes much harder.

If one considers Lipschitz functions on $[0,1]^d$ with the sup norm, then the metric entropy grows much faster—roughly like

$$\log \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \asymp \varepsilon^{-d}.$$

If one inserts this into Dudley's entropy integral, one encounters

$$\int_0^1 \varepsilon^{-d/2}\, d\varepsilon,$$

which diverges as soon as $d \geqslant 2$.

So the naive Dudley argument breaks down in higher dimensions. This is a first glimpse of the curse of dimensionality. To get meaningful results there, one needs either finer chaining arguments, better metrics, or additional structural assumptions on the function class.

7 Look ahead

This lecture extended Dudley's entropy integral bound from Gaussian processes to the broader setting of processes with sub-Gaussian increments, and it recorded several useful variants: bounds for process oscillations, high-probability versions, and analogues for weaker increment assumptions such as sub-exponential tails.

The main application was a first uniform law of large numbers. By viewing the Monte Carlo approximation error as an empirical process indexed by a Lipschitz class, and then bounding the covering numbers of that class, we obtained a uniform $n^{-1/2}$ bound on the integration error over all Lipschitz functions on $[0,1]$.

In the next lecture, we will continue with empirical processes from a different angle. We will introduce symmetrization and related tools, which often give tighter and more flexible ways to control empirical processes than direct Dudley bounds.

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.