Chaining I

SDS 391P.6, Spring 2026

Pratik Patil

1 Motivation

In several earlier lectures, we encountered quantities of the form

$$\sup_{t \in T} X_t,$$

where the index set T is large, often infinite, and $(X_t)_{t \in T}$ is a family of random variables. This is the basic object of study in the theory of *random processes*.

A familiar example already appeared in our covariance estimation lecture. There, we wrote an operator norm as a supremum over the sphere:

$$\|\widehat{\Sigma} - \Sigma\| = \sup_{x \in S^{d-1}} \left| x^\top (\widehat{\Sigma} - \Sigma) x \right|.$$

To handle this quantity, we discretized the sphere by an ε-net and then used a union bound. That was effective because we had strong concentration for each *fixed* direction x, and the metric entropy of the sphere was manageable.

But there is a serious limitation to the one-net method. A single ε-net only approximates T at one scale. If the process has structure at many different scales (and this is often what happens for Gaussian and sub-Gaussian processes that we will study), then a single discretization is too crude. Coarse nets have small cardinality but poor approximation error; fine nets have good approximation error but enormous cardinality. The right method must use *all scales at once*.

That is the idea of *chaining*. Instead of approximating each point $t \in T$ once, we approximate it successively: first coarsely, then more finely, then more finely still. This turns a one-scale ε-net argument into a multiscale argument. The result is Dudley's inequality, which bounds the expected supremum of a Gaussian (or more generally sub-Gaussian) process by an *entropy integral*:

$$\mathbb{E} \sup_{t \in T} X_t \lesssim \int_0^{\operatorname{diam}(T)} \sqrt{\log \mathcal{N}(T, d, \varepsilon)}\, d\varepsilon.$$

This lecture introduces that method. We begin with random processes and their induced metric structure, then discuss Gaussian processes as the main motivating example, explain why expected suprema are important, and finally prove Dudley's inequality through the chaining method.

2 Random processes and their geometry

2.1 Random processes

A random process is simply a family of random variables indexed by a set.

Definition 2.1 (Random process)**.** Let T be a nonempty set. A *random process* indexed by T is a family of real-valued random variables

$$(X_t)_{t\in T}.$$

Historically, the index t often represented time, but in our setting T can be much more general. For example:

- if $T = \{1, \dots, n\}$, then a random process is just a random vector;
- if $T \subset \mathbb{R}^d$, one may think of a random field;
- if T is a class of functions, then $(X_t)_{t\in T}$ may encode empirical errors, stochastic integrals, or Gaussian widths.

The central quantity we will study is the expected supremum

$$\mathbb{E}\sup_{t\in T} X_t.$$

This quantity measures the typical size of the largest fluctuation of the process over its index set.

2.2 Increments and the induced metric

The behavior of a random process is controlled not so much by the individual variables X_t, but by the *increments*

$$X_t - X_s, \qquad s, t \in T.$$

These increments induce a natural pseudo-metric on the index set.

Definition 2.2 (Canonical pseudo-metric)**.** Let $(X_t)_{t\in T}$ be a square-integrable random process. Define

$$d_X(s,t) \;:=\; \|X_t - X_s\|_{L_2} = \big(\mathbb{E}|X_t - X_s|^2\big)^{1/2}, \qquad s, t \in T.$$

This is always a pseudo-metric: symmetry and the triangle inequality follow from the corresponding properties of the L_2 norm. It may happen that $d_X(s,t) = 0$ for distinct $s \neq t$, namely when $X_s = X_t$ almost surely. In that case, one can quotient out such redundant points, so this causes no real difficulty.

The key message is that the expected supremum of a process is governed by the geometry of the index set T under the metric d_X. This is the starting point of the metric approach to random processes.

3 Gaussian processes

3.1 Definition

Among all random processes, Gaussian processes are the cleanest and most important class.

Definition 3.1 (Gaussian process)**.** A random process $(X_t)_{t\in T}$ is a *Gaussian process* if every finite subcollection

$$(X_{t_1}, \dots, X_{t_m})$$

has a multivariate Gaussian distribution.

As usual, the most convenient setting is the centered one.

Definition 3.2 (Centered Gaussian process)**.** A Gaussian process $(X_t)_{t\in T}$ is *centered* if

$$\mathbb{E}X_t = 0 \qquad \text{for all } t \in T.$$

A centered Gaussian process is determined by its covariance function

$$\Sigma(s,t) := \mathbb{E}[X_s X_t].$$

Equivalently, it is determined by its canonical metric

$$d_X(s,t)^2 = \mathbb{E}(X_t - X_s)^2 = \Sigma(t,t) - 2\Sigma(s,t) + \Sigma(s,s).$$

3.2 Examples

Brownian motion. If $(B_t)_{t\in[0,1]}$ is standard Brownian motion, then

$$B_t - B_s \sim \mathcal{N}(0, |t-s|),$$

so the canonical metric is

$$d_B(s,t) = \sqrt{|t-s|}.$$

Thus Brownian motion lives on the interval $[0,1]$, but the geometry relevant to its supremum is *not* the Euclidean metric $|t-s|$, but rather its square root.

Canonical Gaussian process on $\mathbb{R}^n$. Let $g \sim \mathcal{N}(0, I_n)$, and let $T \subset \mathbb{R}^n$. Define

$$X_t := \langle g, t\rangle, \qquad t \in T.$$

Then $(X_t)_{t\in T}$ is a centered Gaussian process, and

$$X_t - X_s = \langle g, t-s\rangle \sim \mathcal{N}(0, \|t-s\|_2^2).$$

Therefore,

$$d_X(s,t) = \|t-s\|_2.$$

So the canonical metric is just the Euclidean distance on T.

This example is especially important because the expected supremum is exactly the *Gaussian width*:

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

Gaussian width is one of the central geometric complexity parameters in high-dimensional probability.

Operator norm of a Gaussian matrix. Let $G \in \mathbb{R}^{m\times n}$ have i.i.d. standard normal entries. Then

$$\|G\| = \sup_{u\in S^{m-1},\, v\in S^{n-1}} \langle Gu, v\rangle.$$

For fixed (u, v), the random variable $\langle Gu, v\rangle$ is Gaussian, so $\|G\|$ is the supremum of a Gaussian process indexed by $S^{m-1} \times S^{n-1}$. This is one of the main reasons Gaussian processes show up in random matrix theory.

4 A maximal inequality for finite sub-Gaussian families

Before turning to chaining, we recall a basic estimate for finite families of sub-Gaussian random variables. It will serve as the one-scale building block inside the chaining proof.

Lemma 4.1 (Finite maximal inequality)**.** *Let $Z_1, \dots, Z_M$ be centered random variables such that*

$$\|Z_j\|_{\psi_2} \leqslant L \qquad \text{for all } j = 1, \dots, M.$$

Then

$$\mathbb{E} \max_{1\leqslant j\leqslant M} Z_j \;\leqslant\; C\, L \sqrt{\log M},$$

where $C > 0$ is an absolute constant.

Proof. Fix $\theta > 0$. By Jensen and the elementary inequality $\max_j a_j \leqslant \theta^{-1} \log \sum_j e^{\theta a_j}$, we have

$$\mathbb{E} \max_{1\leqslant j\leqslant M} Z_j \leqslant \frac{1}{\theta} \log \mathbb{E} \sum_{j=1}^{M} e^{\theta Z_j} = \frac{1}{\theta} \log \sum_{j=1}^{M} \mathbb{E} e^{\theta Z_j}.$$

Since each Z_j is sub-Gaussian with ψ_2-norm at most L, we get

$$\mathbb{E} e^{\theta Z_j} \leqslant e^{C_1 \theta^2 L^2}$$

for an absolute constant C_1. Therefore

$$\mathbb{E} \max_{1\leqslant j\leqslant M} Z_j \leqslant \frac{\log M}{\theta} + C_1 \theta L^2.$$

Optimize in θ by choosing $\theta \asymp \sqrt{\log M}/L$, which gives

$$\mathbb{E} \max_{1\leqslant j\leqslant M} Z_j \leqslant C\, L \sqrt{\log M}.$$

□

This inequality is exactly what a one-net argument uses: the price for replacing a supremum by a maximum over M points is $\sqrt{\log M}$.

5 Why a single ε-net is not enough

Suppose $(X_t)_{t\in T}$ is a centered process on a metric space (T, d), and suppose it has sub-Gaussian increments:

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

A first attempt to bound $\mathbb{E}\sup_{t\in T} X_t$ is to choose an ε-net $T_\varepsilon \subset T$ and approximate each t by a nearby point $\pi_\varepsilon(t) \in T_\varepsilon$.

Then

$$X_t = X_{\pi_\varepsilon(t)} + \big(X_t - X_{\pi_\varepsilon(t)}\big),$$

so

$$\sup_{t\in T} X_t \leqslant \max_{s\in T_\varepsilon} X_s + \sup_{t\in T}\big(X_t - X_{\pi_\varepsilon(t)}\big). \tag{1}$$

The first term is manageable:

$$\mathbb{E}\max_{s\in T_\varepsilon} X_s \lesssim K\ \mathrm{diam}(T)\,\sqrt{\log|T_\varepsilon|},$$

at least after subtracting a reference point. The second term is the problem. Although each increment $X_t - X_{\pi_\varepsilon(t)}$ has ψ_2-norm at most $K\varepsilon$, the supremum still ranges over all $t \in T$, so a single-scale argument does not resolve it.

This is exactly the obstruction that a single net reduces complexity, but does not eliminate the residual supremum. The only natural solution is to repeat the approximation again on the residual term, and then again, and then again. That iteration is chaining.

6 Chaining: the multiscale idea

The chaining idea is very simple to state. Choose a sequence of finer and finer nets

$$T_0,\ T_1,\ T_2,\ \ldots$$

for the metric space T, where T_k is an ε_k-net with $\varepsilon_k \downarrow 0$. For each $t \in T$, let $\pi_k(t) \in T_k$ be a nearest net point at scale ε_k.

Instead of approximating t once, we approximate it progressively:

$$\pi_0(t),\ \pi_1(t),\ \pi_2(t),\ \ldots,\ t.$$

Then

$$X_t - X_{\pi_0(t)} = \sum_{k\geqslant 1}\big(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\big).$$

This is the chain.

The point is that each link in the chain is small:

$$\|\ X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\ \|_{\psi_2} \lesssim K\,\varepsilon_{k-1}.$$

At the same time, the number of possible links at level k is controlled by the size of the nets: roughly

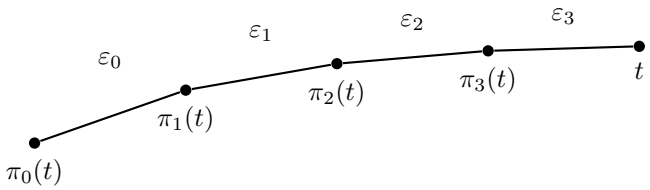
$$|T_k|\cdot|T_{k-1}|.$$

So each scale contributes something like

$$K\,\varepsilon_k\,\sqrt{\log \mathcal{N}(T,d,\varepsilon_k)}.$$

Summing over scales leads to Dudley's inequality.

This multiscale decomposition is illustrated schematically in Figure 1.



A chain approximates t at progressively finer scales.
The links get shorter, but the nets get larger.

Figure 1: The chaining idea: approximate each point t by a sequence of progressively finer net points $\pi_k(t)$.

7 Dudley's inequality: discrete form

We now state and prove the discrete version first, since it makes the chaining structure most transparent.

Theorem 7.1 (Dudley's discrete inequality)**.** *Let $(X_t)_{t\in T}$ be a centered random process on a metric space (T,d) such that*

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

Assume for simplicity that T is finite. Then

$$\mathbb{E}\sup_{t\in T} X_t \;\leqslant\; CK\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log \mathcal{N}(T,d,2^{-k})},$$

where $C>0$ is an absolute constant.

Proof. Let

$$\varepsilon_k := 2^{-k}, \qquad k\in\mathbb{Z}.$$

For each k, choose an ε_k-net $T_k \subset T$ with

$$|T_k| = \mathcal{N}(T,d,\varepsilon_k).$$

Since T is finite, there exist integers $k_0 < k_1$ such that:

- $T_{k_0} = \{t_0\}$ is a singleton;
- $T_{k_1} = T$.

For each $t\in T$, choose $\pi_k(t)\in T_k$ such that

$$d(t,\pi_k(t)) \leqslant \varepsilon_k.$$

Also note that $\pi_{k_1}(t) = t$.

Because the process is centered, we have

$$\mathbb{E}X_{t_0} = 0,$$

so

$$\mathbb{E}\sup_{t\in T} X_t = \mathbb{E}\sup_{t\in T}(X_t - X_{t_0}).$$

Now telescope:

$$X_t - X_{t_0} = X_{\pi_{k_1}(t)} - X_{\pi_{k_0}(t)} = \sum_{k=k_0+1}^{k_1} \big(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\big).$$

Taking supremum in t and using subadditivity of the supremum, we get

$$\mathbb{E}\sup_{t\in T}(X_t - X_{t_0}) \leqslant \sum_{k=k_0+1}^{k_1} \mathbb{E}\sup_{t\in T}\big(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\big).$$

We bound each scale separately. For fixed k, the number of distinct pairs

$$(\pi_k(t), \pi_{k-1}(t))$$

is at most

$$|T_k|\,|T_{k-1}| \leqslant |T_k|^2.$$

Also, by the triangle inequality, we have

$$d(\pi_k(t), \pi_{k-1}(t)) \leqslant d(\pi_k(t), t) + d(t, \pi_{k-1}(t)) \leqslant \varepsilon_k + \varepsilon_{k-1} \leqslant 3\varepsilon_{k-1}.$$

Therefore each increment

$$X_{\pi_k(t)} - X_{\pi_{k-1}(t)}$$

is centered sub-Gaussian with ψ_2-norm at most $3K\varepsilon_{k-1}$. Applying Lemma 4.1 to the family of all such increments gives

$$\mathbb{E}\sup_{t\in T}\big(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\big) \leqslant CK\varepsilon_{k-1}\sqrt{\log|T_k|^2}.$$

Since $\sqrt{\log|T_k|^2} = \sqrt{2\log|T_k|} \lesssim \sqrt{\log|T_k|}$, this becomes

$$\mathbb{E}\sup_{t\in T}\big(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\big) \leqslant CK\varepsilon_{k-1}\sqrt{\log\mathcal{N}(T, d, \varepsilon_k)}.$$

Summing over k and recalling $\varepsilon_k = 2^{-k}$, we obtain

$$\mathbb{E}\sup_{t\in T} X_t \leqslant CK \sum_{k=k_0+1}^{k_1} 2^{-k}\sqrt{\log\mathcal{N}(T, d, 2^{-k})}.$$

Extending the sum to all $k \in \mathbb{Z}$ only increases the right-hand side, and this proves the theorem. □

8 Dudley's inequality: entropy integral form

The discrete theorem is already enough for most purposes. The integral form is often cleaner to state and easier to manipulate.

Theorem 8.1 (Dudley's entropy integral inequality)**.** *Let $(X_t)_{t\in T}$ be a centered random process on a metric space (T,d) such that*

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

Assume T is finite (or more generally, that the process is separable). Then

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

where $C>0$ is an absolute constant.

Proof. Starting from Theorem 7.1, we have

$$\mathbb{E}\sup_{t\in T} X_t \leqslant CK \sum_{k\in\mathbb{Z}} 2^{-k} \sqrt{\log \mathcal{N}(T,d,2^{-k})}.$$

Now write

$$2^{-k} = 2\int_{2^{-k-1}}^{2^{-k}} d\varepsilon.$$

Since $\varepsilon \mapsto \mathcal{N}(T,d,\varepsilon)$ is decreasing in ε, we get

$$\sqrt{\log \mathcal{N}(T,d,2^{-k})} \leqslant \sqrt{\log \mathcal{N}(T,d,\varepsilon)} \qquad \text{for } \varepsilon \in [2^{-k-1}, 2^{-k}].$$

Therefore

$$2^{-k}\sqrt{\log \mathcal{N}(T,d,2^{-k})} \leqslant 2\int_{2^{-k-1}}^{2^{-k}} \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\, d\varepsilon.$$

Summing over k gives

$$\mathbb{E}\sup_{t\in T} X_t \leqslant CK \int_0^{\infty} \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\, d\varepsilon.$$

Finally, if $\varepsilon > \operatorname{diam}(T,d)$, then one ball covers all of T, so

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

Thus the integral truncates at $\operatorname{diam}(T,d)$. □

Remark 8.2. For Gaussian processes, one usually applies Dudley with the canonical metric

$$d(s,t) = \|X_t - X_s\|_{L_2}.$$

Because Gaussian increments are sub-Gaussian with ψ_2-norm comparable to their L_2-norm, Theorem 8.1 immediately gives an upper bound on centered Gaussian processes.

9 Example: canonical Gaussian process on the Euclidean ball

Let $g \sim \mathcal{N}(0, I_n)$, and take $T = B_2^n \subset \mathbb{R}^n$. Then the canonical Gaussian process is

$$X_t = \langle g, t \rangle, \qquad t \in B_2^n.$$

Its expected supremum is

$$\mathbb{E} \sup_{t \in B_2^n} \langle g, t \rangle = \mathbb{E}\|g\|_2,$$

which is of order $\sqrt{n}$.

Let us recover this from Dudley's inequality.

The metric is Euclidean:

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

From the volumetric bound for covering numbers of the Euclidean ball,

$$\mathcal{N}(B_2^n, \|\cdot\|_2, \varepsilon) \leqslant \left(\frac{3}{\varepsilon}\right)^n, \qquad 0 < \varepsilon \leqslant 1.$$

Therefore

$$\log \mathcal{N}(B_2^n, \varepsilon) \leqslant n \log(3/\varepsilon).$$

Applying Dudley, we get

$$\mathbb{E}\|g\|_2 = \mathbb{E} \sup_{t \in B_2^n} \langle g, t \rangle \leqslant C \int_0^1 \sqrt{n \log(3/\varepsilon)}\, d\varepsilon.$$

Since the singularity at 0 is integrable, we have

$$\int_0^1 \sqrt{\log(3/\varepsilon)}\, d\varepsilon \leqslant C',$$

and so

$$\mathbb{E}\|g\|_2 \leqslant C'' \sqrt{n}.$$

This matches the correct order.

So in this example, Dudley's bound is sharp up to constants.

10 Strengths and limitations of Dudley's inequality

Dudley's inequality has a very useful interpretation:

$$\mathbb{E} \sup_{t \in T} X_t \lesssim \text{total metric complexity of } T \text{ across all scales.}$$

The quantity

$$\sqrt{\log \mathcal{N}(T, d, \varepsilon)}$$

measures the complexity of T at resolution ε, and Dudley's integral adds those complexities over all ε.

This is the first major place in the course where metric entropy becomes a multiscale object. In the covariance-estimation lecture, covering numbers entered at one chosen scale. Here, the whole curve

$$\varepsilon \longmapsto \log \mathcal{N}(T, d, \varepsilon)$$

matters.

That is why chaining is more powerful than a one-net argument: it does not commit to one scale. Instead, it allows coarse approximations where coarse approximations are enough, and fine approximations only where they are needed.

Dudley's inequality is a beautiful theorem, but it is not the end of the story. It is an upper bound, and in general it need not be sharp. There are processes for which the Dudley integral overestimates the true expected supremum, sometimes by logarithmic factors. This is not a defect of the proof so much as a sign that the simple chaining argument, where we move the supremum through the telescoping sum, still loses information.

11 Look ahead

This lecture introduced random processes through the lens of geometry. For Gaussian processes, and more generally for processes with sub-Gaussian increments, the expected supremum is governed by the metric structure induced by the increments. That geometric point of view naturally leads to covering numbers and metric entropy.

A single ε-net is often too crude, because it only captures one scale of approximation. Chaining fixes this by combining nets at all scales. The result is Dudley's entropy integral inequality, which bounds the expected supremum by a multiscale integral of $\sqrt{\log \mathcal{N}(T, d, \varepsilon)}$.

In the next lecture, we will develop useful variants of Dudley's bound and begin applying chaining ideas to empirical processes and uniform laws of large numbers.

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.