Notes on Generalization Theory

Daniel López Montero

January 4, 2026

Abstract

These notes includes a self-contained collection of the most relevant results used in generalization bounds and empirical process theory.

1 Preliminaries and Notation

We work on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. A *random* (or *stochastic*) process is a collection of random variables $\{X_t : t \in T\}$ indexed by a set T. In many applications T is a subset of the real line (time), but later we also consider general index sets equipped with a metric.

When T is ordered, a filtration is a family of sub-σ-algebras $\{\mathcal{F}_t\}_{t\in T}$ with $\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{A}$ for $s \leq t$. A process $\{X_t\}_{t\in T}$ is *adapted* to $\{\mathcal{F}_t\}_{t\in T}$ if X_t is $\mathcal{F}_t$-measurable for each t. The *natural filtration* of $\{X_t\}$ is $\mathcal{F}_t := \sigma(\{X_s\}_{s\leq t})$.

In discrete time, i.e., $T = \{0, 1, 2, \dots\}$, a sequence $\{M_k\}_{k\geq 0}$ is a (discrete) *martingale* with respect to $\{\mathcal{F}_k\}$ if M_k is $\mathcal{F}_k$-measurable and $\mathbb{E}[M_k \mid \mathcal{F}_{k-1}] = M_{k-1}$ for all $k \geq 1$. The increments $X_k := M_k - M_{k-1}$ then satisfy $\mathbb{E}[X_k \mid \mathcal{F}_{k-1}] = 0$ and are called a *martingale difference sequence.*

2 Subgaussian Random Variables

Subgaussian random variables are those whose centered moment generating function is controlled by that of a Gaussian. For a centered normal random variable $G \sim \mathcal{N}(0, \sigma^2)$, we have

$$\mathbb{E}\left[e^{\lambda G}\right] = e^{\sigma^2\lambda^2/2}.$$

The definition below abstracts this inequality as a convenient tail/concentration condition.

Definition 1 (Subgaussian random variable)**.** Let X be a random variable, we define the *log-moment generating function* ψ of X as

$$\psi(\lambda) := \log \mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}]$$

and we say that X random variable is σ^2-subgaussian if its log-moment generating function satisfies $\psi(\lambda) \leq \lambda^2\sigma^2/2$ for all $\lambda \in \mathbb{R}$, and the smallest σ^2 for which that holds is called the variance proxy of X.

The following result is a variant of the Markov's inequality written in terms of the log-moment generating function.

Lemma 2 (Chernoff bound)**.** Let X be a σ^2-subgaussian random variable. Then, the following inequality holds

$$\mathbb{P}[X - \mathbb{E}X \geq a] \leq \exp\left\{\frac{-a^2}{2\sigma^2}\right\}.$$

In particular, $\mathbb{P}[|X - \mathbb{E}X| \geq a] \leq 2\exp\{-a^2/2\sigma^2\}$.

Proof. We exponentiate inside the probability before applying Markov's inequality. For any $\lambda \geq 0$, we have

$$\mathbb{P}[X - \mathbb{E}X \geq a] = \mathbb{P}[e^{\lambda(X-\mathbb{E}X)} \geq e^{\lambda a}] \leq e^{-\lambda a}\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] = e^{\psi(\lambda)-\lambda a}.$$

Since this holds for every $\lambda \geq 0$, we can choose λ to obtain the best bound. The optimal choice is $\lambda = \frac{a}{\sigma^2}$. Substituting this into the inequality and using the bound $\psi(\lambda) \leq \lambda^2\sigma^2/2$, we obtain

$$\mathbb{P}[X - \mathbb{E}X \geq a] \leq \exp\left\{\left(\frac{a}{\sigma^2}\right)^2\sigma^2/2 - \left(\frac{a}{\sigma^2}\right)a\right\} = \exp\left\{\frac{-a^2}{2\sigma^2}\right\}.$$

□

Lemma 3 (Maximal inequality)**.** Let $\{X_k\}_{1\leq k\leq n}$ be a collection of σ^2-subgaussian random variables with the same variance, satisfying $\mathbb{E}[X_k] = 0$ for all $k = 1, \ldots, n$. Then

$$\mathbb{E}\left[\sup_{1\leq k\leq n} X_k\right] \leq \sqrt{2\sigma^2\log n}.$$

Proof. Since $-\log x$ is convex, by Jensen's inequality, we have for any $\lambda > 0$

$$\begin{aligned}\mathbb{E}\left[\sup_k X_k\right] &= \mathbb{E}\left[\frac{1}{\lambda}\log(e^{\lambda \sup_k X_k})\right] \leq \frac{1}{\lambda}\log\mathbb{E}[e^{\lambda\sup_k X_k}] \\ &\leq \frac{1}{\lambda}\log\sum_{1\leq k\leq n}\mathbb{E}[e^{\lambda X_k}] \leq \frac{1}{\lambda}\log\left(ne^{\lambda^2\sigma^2/2}\right) = \frac{\log n}{\lambda} + \frac{\lambda\sigma^2}{2}.\end{aligned}$$

Since this holds for every $\lambda > 0$, we can now optimize over λ on the right hand side. Differentiating and setting it equal to zero, we obtain the minimum at $\lambda = \frac{\sqrt{2\log n}}{\sigma}$. Substituting this value into the equation, we obtain the desired result

$$\mathbb{E}\left[\sup_{1\leq k\leq n} X_k\right] \leq \sqrt{2\sigma^2\log n}.$$

□

Lemma 4 (Hoeffding lemma)**.** Let X be a random variable such that $a \le X \le b$ a.s. for some $a, b \in \mathbb{R}$. Then,

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \le e^{\lambda^2(b-a)^2/8}.$$

In other words, X is a $(b-a)^2/4$-subgaussian random variable.

Proof. Let us define the mean-centered random variable $Y := X - \mathbb{E}[X]$, then $\tilde{a} \le Y \le \tilde{b}$ where $\tilde{a} := a - \mathbb{E}[X]$ and $\tilde{b} := b - \mathbb{E}[X]$. By definition, log-moment generating function of X is given by $\psi(\lambda) = \log \mathbb{E}[e^{\lambda Y}]$, along with its derivatives

$$\psi'(\lambda) = \frac{\mathbb{E}[Ye^{\lambda Y}]}{\mathbb{E}[e^{\lambda Y}]}, \qquad \psi''(\lambda) = \frac{\mathbb{E}[Y^2 e^{\lambda Y}]}{\mathbb{E}[e^{\lambda Y}]} - \left(\frac{\mathbb{E}[Ye^{\lambda Y}]}{\mathbb{E}[e^{\lambda Y}]}\right)^2.$$

We can interpret $\psi''(\lambda)$ as the variance of the random variable Y under another probability measure. We define a new probability measure $\mathbb{Q}$ by setting $d\mathbb{Q} := \frac{e^{\lambda Y}}{\mathbb{E}[e^{\lambda Y}]} d\mathbb{P}$. The definition is well-posed because the Radon-Nikodym derivative $\frac{e^{\lambda Y}}{\mathbb{E}[e^{\lambda Y}]}$ is positive and integrates to 1. From this, we deduce that

$$\mathbb{E}_{\mathbb{Q}}[g(Y)] = \frac{\mathbb{E}[g(Y)e^{\lambda Y}]}{\mathbb{E}[e^{\lambda Y}]}.$$

Therefore,

$$\mathrm{Var}_{\mathbb{Q}}(Y) = \mathbb{E}_{\mathbb{Q}}[Y^2] - (\mathbb{E}_{\mathbb{Q}}[Y])^2 = \frac{\mathbb{E}[Y^2 e^{\lambda Y}]}{\mathbb{E}[e^{\lambda Y}]} - \left(\frac{\mathbb{E}[Ye^{\lambda Y}]}{\mathbb{E}[e^{\lambda Y}]}\right)^2 = \psi''(\lambda).$$

We can bound the variance of Y as follows

$$\mathrm{Var}_{\mathbb{Q}}(Y) = \mathrm{Var}_{\mathbb{Q}}(Y - z) \le \mathbb{E}_{\mathbb{Q}}[(Y-z)^2].$$

Since z can be any real number, we let $z = (\tilde{a} + \tilde{b})/2$. Given that $\tilde{a} \le Y \le \tilde{b}$ a.s., we obtain

$$\mathbb{E}_{\mathbb{Q}}[(Y-z)^2] = \frac{1}{4}\mathbb{E}_{\mathbb{Q}}[(2Y - \tilde{a} - \tilde{b})^2] \le \frac{(\tilde{b} - \tilde{a})^2}{4} = \frac{(b-a)^2}{4}.$$

Since Y is centered, $\psi(0) = 0$ and $\psi'(0) = \mathbb{E}[Y] = 0$. Using the bound $\psi''(\lambda) \le \frac{(b-a)^2}{4}$ and the Fundamental Theorem of Calculus, we obtain

$$\psi(\lambda) = \int_0^\lambda \int_0^\mu \psi''(\rho) d\rho d\mu \le \frac{\lambda^2(b-a)^2}{8}.$$

Therefore, X is a σ^2-subgaussian random variable with $\sigma^2 = (b-a)^2/4$. □

3 Martingale Concentration and Bounded Differences

Hoeffding's lemma gives subgaussian bounds for bounded random variables. Combined with a filtration and a martingale difference sequence, it yields concentration for dependent sums (Azuma–Hoeffding). A standard application is McDiarmid's bounded differences inequality for functions of independent variables.

Lemma 5 (Azuma)**.** Let $\{X_k\}_{1\le k\le n}$ be a stochastic process adapted to the natural filtration $\{\mathcal{F}_k\}_{1\le k\le n}$, i.e., $\mathcal{F}_k = \sigma(X_1,\ldots,X_k)$. Assume that the random variables satisfy:

$$\mathbb{E}[e^{\lambda X_k} \mid \mathcal{F}_{k-1}] \le e^{\lambda^2\sigma_k^2/2} \quad a.s. \qquad\qquad \text{for all } \lambda\in\mathbb{R},\ k=1,\ldots,n. \tag{1}$$

where $\sigma_k^2 \ge 0$ is a (deterministic) variance proxy for X_k. Then, $\mathbb{E}[X_k \mid \mathcal{F}_{k-1}] = 0$ (and hence $\mathbb{E}[X_k] = 0$) and the sum $\sum_{k=1}^n X_k$ is σ^2-subgaussian with variance proxy $\sigma^2 := \sum_{k=1}^n \sigma_k^2$.

Proof. First, we prove that $\mathbb{E}[X_k \mid \mathcal{F}_{k-1}] = 0$. Since equality holds in (1) at $\lambda = 0$, differentiating both sides at $\lambda = 0$ yields

$$\mathbb{E}[X_k \mid \mathcal{F}_{k-1}] = \frac{d}{d\lambda}\Big|_{\lambda=0} \mathbb{E}\left[e^{\lambda X_k} \mid \mathcal{F}_{k-1}\right] \le \frac{d}{d\lambda}\Big|_{\lambda=0} e^{\lambda^2\sigma_k^2/2} = 0.$$

Applying the same argument to $-X_k$ gives $\mathbb{E}[-X_k \mid \mathcal{F}_{k-1}] \le 0$, hence $\mathbb{E}[X_k \mid \mathcal{F}_{k-1}] = 0$.

Next, we bound the moment generating function of $S_k := \sum_{i=1}^k X_i$. Since S_{k-1} is $\mathcal{F}_{k-1}$-measurable, the tower property gives

$$\begin{aligned}\mathbb{E}[e^{\lambda S_k}] &= \mathbb{E}\left[\mathbb{E}\left[e^{\lambda S_{k-1}} e^{\lambda X_k} \mid \mathcal{F}_{k-1}\right]\right] = \mathbb{E}\left[e^{\lambda S_{k-1}}\mathbb{E}\left[e^{\lambda X_k} \mid \mathcal{F}_{k-1}\right]\right] \\ &\le e^{\lambda^2\sigma_k^2/2}\, \mathbb{E}[e^{\lambda S_{k-1}}].\end{aligned}$$

Iterating this inequality yields $\mathbb{E}[e^{\lambda S_n}] \le e^{\lambda^2 \sum_{i=1}^n \sigma_i^2/2} = e^{\lambda^2\sigma^2/2}$. Since $\mathbb{E}[S_n] = 0$, this is exactly the subgaussian bound for S_n. □

Corollary 6 (Azuma-Hoeffding inequality)**.** Let $\{\mathcal{F}_k\}_{1\le k\le n}$ be a filtration and let $\{X_k\}_{1\le k\le n}$ be an adapted process such that $\mathbb{E}[X_k \mid \mathcal{F}_{k-1}] = 0$ for $k = 1,\ldots,n$. Assume that there exist $\mathcal{F}_{k-1}$-measurable random variables A_k, B_k such that $A_k \le X_k \le B_k$ a.s.

Then, $\sum_{k=1}^n X_k$ is σ^2-subgaussian with $\sigma^2 = \frac{1}{4}\sum_{k=1}^n \|B_k - A_k\|_\infty^2$. In particular, for every $t \ge 0$,

$$\mathbb{P}\left[\sum_{k=1}^n X_k \ge t\right] \le \exp\left(-\frac{2t^2}{\sum_{k=1}^n \|B_k - A_k\|_\infty^2}\right).$$

The same bound holds for $\mathbb{P}\left[\left|\sum_{k=1}^n X_k\right| \ge t\right]$ up to an extra factor 2.

Proof. Conditionally on $\mathcal{F}_{k-1}$, the bounds $A_k \le X_k \le B_k$ are deterministic and $\mathbb{E}[X_k \mid \mathcal{F}_{k-1}] = 0$. Applying Hoeffding's Lemma 4 to this conditional distribution yields, for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda X_k} \mid \mathcal{F}_{k-1}\right] \le \exp\left(\frac{\lambda^2(B_k - A_k)^2}{8}\right) \le \exp\left(\frac{\lambda^2\|B_k - A_k\|_\infty^2}{8}\right).$$

Therefore (1) holds with $\sigma_k^2 = \|B_k - A_k\|_\infty^2/4$, and Lemma 5 implies that $\sum_{k=1}^n X_k$ is σ^2-subgaussian with $\sigma^2 = \frac{1}{4}\sum_{k=1}^n \|B_k - A_k\|_\infty^2$.

The one-sided tail bound follows from Lemma 2 (applied to $\sum_{k=1}^n X_k$), and the two-sided bound follows by also applying it to $-\sum_{k=1}^n X_k$. □

Definition 7 (Discrete derivative)**.** Let $f \in C(\mathbb{R}^n, \mathbb{R})$. We define the *discrete derivative* of f with respect to variable x_k at the point $x \in \mathbb{R}^n$ as follows:

$$\mathfrak{D}_k f(x) := \sup_z f(x_1,\ldots,x_{k-1},z,x_{k+1},\ldots,x_n) - \inf_z f(x_1,\ldots,x_{k-1},z,x_{k+1},\ldots,x_n).$$

Theorem 8 (McDiarmid)**.** Let $X_1, \dots, X_n$ be independent random variables and let $f \in C(\mathbb{R}^n, \mathbb{R})$. Then, $f(X_1, \dots, X_n)$ is σ^2-subgaussian with $\sigma^2 = \frac{1}{4}\sum_{k=1}^n \|\mathfrak{D}_k f\|_\infty^2$ where $\mathfrak{D}_k f$ is the discrete derivative (Definition 7).

Proof. Let $\mathcal{F}_k := \sigma(X_1, \dots, X_k)$ be the natural filtration and define the Doob martingale

$$M_k := \mathbb{E}[f(X_1, \dots, X_n) \mid \mathcal{F}_k], \qquad k = 0, 1, \dots, n.$$

Set $Y_k := M_k - M_{k-1}$ for $k = 1, \dots, n$. Then $\mathbb{E}[Y_k \mid \mathcal{F}_{k-1}] = 0$ and the sum telescopes to

$$\sum_{k=1}^n Y_k = f(X_1, \dots, X_n) - \mathbb{E} f(X_1, \dots, X_n). \tag{2}$$

Fix $k \in \{1, \dots, n\}$. Define the $\mathcal{F}_{k-1}$-measurable function

$$h_k(z) := \mathbb{E}[f(X_1, \dots, X_{k-1}, z, X_{k+1}, \dots, X_n) \mid \mathcal{F}_{k-1}] \,.$$

By independence, $M_k = h_k(X_k)$ and $M_{k-1} = \mathbb{E}[h_k(X_k) \mid \mathcal{F}_{k-1}]$. Let

$$\ell_k := \inf_z h_k(z), \qquad u_k := \sup_z h_k(z).$$

Then $\ell_k \le M_k \le u_k$ and $\ell_k \le M_{k-1} \le u_k$, hence

$$\ell_k - M_{k-1} \le Y_k \le u_k - M_{k-1}.$$

Therefore $A_k := \ell_k - M_{k-1}$ and $B_k := u_k - M_{k-1}$ are $\mathcal{F}_{k-1}$-measurable and satisfy $A_k \le Y_k \le B_k$ almost surely, with

$$B_k - A_k = u_k - \ell_k \le \mathbb{E}[\mathfrak{D}_k f(X_1, \dots, X_n) \mid \mathcal{F}_{k-1}] \le \|\mathfrak{D}_k f\|_\infty.$$

Applying the Azuma–Hoeffding inequality (Corollary 6) to $\sum_{k=1}^n Y_k$ and using (2) concludes the proof. □

4 Metric Entropy: Covering and Packing Numbers

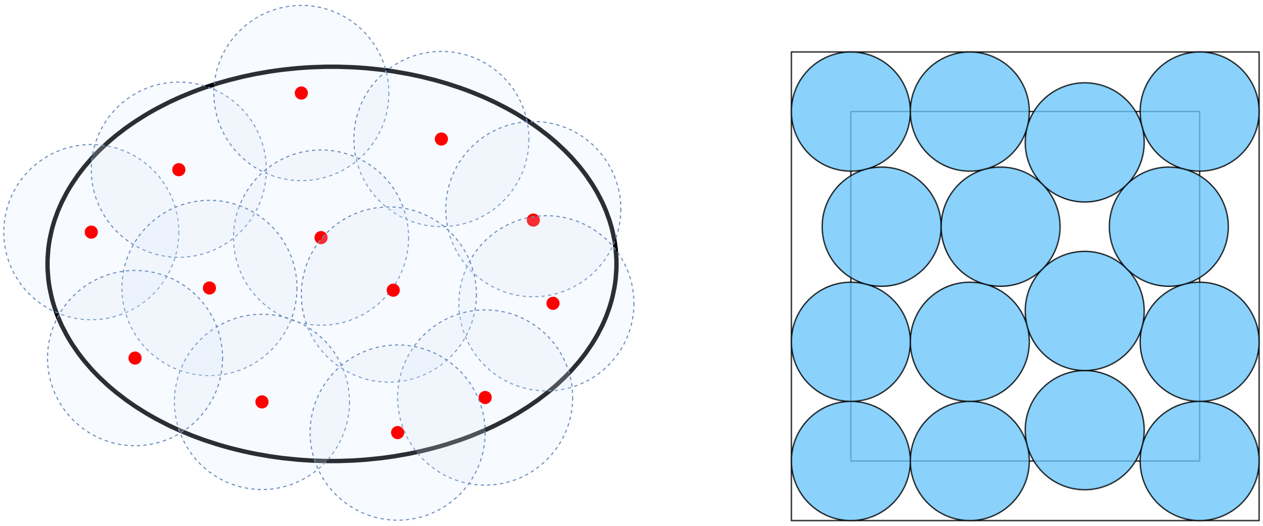
To control suprema of random processes indexed by a metric space, we need a notion of the "size" or "complexity" of the index set (E, d). Covering and packing numbers capture this idea by counting how many metric balls are needed to cover a set, or how many well-separated points it contains (Figure 1).

Definition 9 ($\epsilon-$net and covering number)**.** A set $N \subseteq E$ is called a $\epsilon-$*net* for (E, d) if for every $x \in E$, there exists $\pi(x) \in N$ such that $d(x, \pi(x)) \le \epsilon$. The smallest cardinality of an ϵ-net for (E, d) is called the *covering number*

$$N(E, d, \epsilon) := \inf\{|N| : N \text{ is an } \epsilon-\text{net for } (E, d)\}.$$

Definition 10 ($\epsilon-$packing and packing number)**.** A set $N \subseteq E$ is called an $\epsilon-$*packing* of (E, d) if $d(x, x') > \epsilon$ for every $x, x' \in N$, $x \neq x'$. The largest cardinality of an $\epsilon-$packing of (E, d) is called the *packing number*

$$D(E, d, \epsilon) := \sup\{|N| : N \text{ is an } \epsilon-\text{packing of } (E, d).\}.$$



(a) The ellipse represents E and the red dots are the elements of an ϵ-net. The circles are balls of radius ϵ covering the set.

(b) The optimal packing problem in a square using the blue balls.

Figure 1: Packing vs Covering number

In the literature, the binary logarithm of the covering number of a set E is commonly referred to as the $\boldsymbol{\epsilon}$**−entropy** of E, denoted by

$$H_\epsilon(E) := \log_2 N(E, d, \epsilon).$$

Similarly, the binary logarithm of the packing number is typically called the $\boldsymbol{\epsilon}$**−capacity** of E, defined as

$$C_\epsilon(E) := \log_2 D(E, d, \epsilon).$$

The base-2 logarithm is common in information-theoretic contexts, where the natural unit is the bit.

The following lemma illustrates the relationship between the covering number and the packing number.

Lemma (Duality between covering and packing number)**.** For every $\epsilon > 0$

$$D(E, d, 2\epsilon) \leq N(E, d, \epsilon) \leq D(E, d, \epsilon).$$

Proof. (1) Let D be a $2\epsilon-$packing and let N be an $\epsilon-$net. For every $x \in D$, choose $\pi(x) \in N$ such that $d(x, \pi(x)) \leq \epsilon$. Then, for every $x' \in D$ such that $x \neq x'$, we have

$$2\epsilon < d(x, x') \leq d(x, \pi(x)) + d(\pi(x), \pi(x')) + d(\pi(x'), x') \leq 2\epsilon + d(\pi(x), \pi(x')),$$

which implies that $\pi(x) \neq \pi(x')$ (see Figure 2). Thus, the function $\pi : D \to N$ is injective, and thus, $|D| \leq |N|$. In other words, $D(E, d, 2\epsilon) \leq N(E, d, \epsilon)$.

(2) Let D be a *maximal* $\epsilon-$packing with $|D| = D(E, d, \epsilon)$. We claim that D is necessarily an $\epsilon-$net. Indeed, suppose for contradiction that there exists a point $x \in E$ such that $d(x, x') > \epsilon$ for every $x' \in D$. This would imply that $D \cup \{x\}$ is a *larger* $\epsilon-$packing, contradicting the maximality of D. Therefore, every point in E must be within a distance at most of ϵ from some point in D, confirming that D is an $\epsilon-$net. □

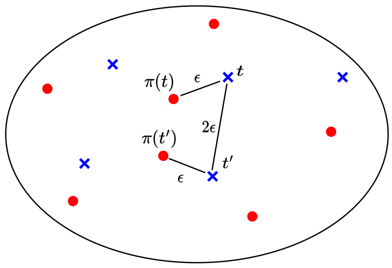


Figure 2: distance between x and x'.

We are now ready to establish an upper bound on the covering number of the Euclidean ball B_2^n with respect to the Euclidean distance. The proof of this fundamental result employs a clever technique known as a *volume argument*.

Lemma. Let B_2^n be the n-dimensional Euclidean ball centered at zero with radius 1, i.e., $B_2^n := \{x \in \mathbb{R}^n : \|x\|_2 < 1\}$. Then, $N(B_2^n, \|\cdot\|_2, \epsilon) = 1$ for $\epsilon \geq 1$ and

$$\left(\frac{1}{\epsilon}\right)^n \leq N(B_2^n, \|\cdot\|_2, \epsilon) \leq \left(\frac{3}{\epsilon}\right)^n \qquad \text{for } 0 < \epsilon < 1.$$

Proof. Case $\epsilon \geq 1$: For $\epsilon \geq 1$, we have $N(B_2^n, \|\cdot\|_2, \epsilon) = 1$ since $\{0\}$ is an ϵ-net: $\|x\|_2 < 1 \leq \epsilon$ for every $x \in B_2^n$.

Case $0 < \epsilon < 1$: We begin with the upper bound. Let D be a 2ϵ-packing of B_2^n. Since $\|x - x'\|_2 > 2\epsilon$ for all $x \neq x'$ in D, the balls $\{B(x, \epsilon) : x \in D\}$ are disjoint. Moreover, for any $x \in B_2^n$ we have $B(x, \epsilon) \subseteq B(0, 1 + \epsilon)$. Therefore,

$$\sum_{x \in D} \lambda(B(x, \epsilon)) = \lambda\left(\bigcup_{x \in D} B(x, \epsilon)\right) \leq \lambda(B(0, 1 + \epsilon))$$

where λ denotes Lebesgue measure on $\mathbb{R}^n$. Using homogeneity $\lambda(B(0, \alpha)) = \alpha^n \lambda(B(0, 1))$, we obtain

$$|D|\, \lambda(B(0, \epsilon)) \leq \lambda(B(0, 1 + \epsilon)) \quad \Rightarrow \quad |D|\, \epsilon^n \lambda(B(0, 1)) \leq (1 + \epsilon)^n \lambda(B(0, 1)).$$

Therefore,

$$|D| \leq \left(\frac{1 + \epsilon}{\epsilon}\right)^n.$$

We have established that for every 2ϵ-packing D of B_2^n, we obtain an upper bound for $D(B_2^n, \|\cdot\|_2, 2\epsilon)$. This leads to the following chain of inequalities

$$N(B_2^n, \|\cdot\|_2, 2\epsilon) \overset{\text{Lemma 4}}{\leq} D(B_2^n, \|\cdot\|_2, 2\epsilon) \leq \left(1 + \frac{1}{\epsilon}\right)^n \leq \left(\frac{3}{2\epsilon}\right)^n \qquad \text{for } 2\epsilon < 1.$$

Relabeling 2ϵ as ϵ completes the proof.

We proceed similarly to obtain the lower bound. Let N be an $\epsilon-$net for B_2^n. Then,

$$\lambda(B_2^n) \leq \lambda\left(\bigcup_{x \in N} B(x, \epsilon)\right) \leq \sum_{x \in N} \lambda(B(x, \epsilon)) = |N|\, \lambda(B(0, \epsilon)).$$

Hence,

$$|N| \geq \frac{\lambda(B(0, 1))}{\lambda(B(0, \epsilon))} = \left(\frac{1}{\epsilon}\right)^n.$$

This inequality holds for every ϵ-net N, so we conclude that $N(B_2^n, \|\cdot\|_2, \epsilon) \geq (1/\epsilon)^n$. □

The following Lemma finds a bound for the growing rate of the $\epsilon-$entropy of the Sobolev space H^{m+1}. The proof can be found in [Nickl and Pötscher, 2007, Corollary 4].

Lemma (ϵ-Entropy of $H^{m+1}(\Omega,\mathbb{R}^{d_2})$)**.** Let $\Omega \subseteq \mathbb{R}^{d_2}$ be a Lipschitz domain. For $m \geq 1$, one has

$$\log N(B_{H^{m+1}(\Omega)}(1), \|\cdot\|_{H^{m+1}(\Omega)}, \epsilon) = \mathcal{O}_{\epsilon\to 0}(\epsilon^{-d_2/(m+1)}).$$

5 Subgaussian Processes and Chaining

We now return to random processes indexed by a metric space (T, d). Dudley's inequality bounds $\mathbb{E}[\sup_{t\in T} X_t]$ for a subgaussian process in terms of the metric entropy of (T, d), as developed in the previous section.

Definition 11 (Subgaussian process)**.** A random process $\{X_t\}_{t\in T}$ on a metric space (T, d) is called *subgaussian* if $\mathbb{E}[X_t] = 0$ for all $t \in T$ and

$$\mathbb{E}[e^{\lambda(X_t - X_s)}] \leq e^{\lambda^2 d(t,s)^2/2} \qquad \text{for all } t, s \in T,\ \lambda \in \mathbb{R}. \tag{3}$$

Notice that for every $s, t \in T$, the random variable $X_t - X_s$ is a $d(t, s)^2$-subgaussian random variable.

Definition (Separable process)**.** A random process $\{X_t\}_{t\in T}$ is called separable (with respect to d) if there exists a countable set $T_0 \subseteq T$ such that for every $t \in T$ there exists a sequence $\{t_n\}_{n\in\mathbb{N}} \subseteq T_0$ with $d(t_n, t) \to 0$ and $X_{t_n} \to X_t$ almost surely.

Remark. The assumption of separability is almost always satisfied. For example, assuming that $t \to X_t$ is continuous and T is a separable metric space (as is the case in this manuscript), then we can take T_0 to be any countable dense subset of T, allowing us to verify the separability assumption.

For the next theorem, we need $\sup_{t\in T} X_t$ to be measurable. If T is uncountable, the pointwise supremum need not be measurable in general. Under separability, however, we have $\sup_{t\in T} X_t = \sup_{t\in T_0} X_t$ almost surely, and the right-hand side is a supremum of countably many measurable random variables, hence measurable.

Theorem 12 (Dudley)**.** Let $\{X_t\}_{t\in T}$ be a separable subgaussian process in the metric space (T, d). Then,

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq 6 \sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log N(T, d, 2^{-k})}.$$

Proof. We begin by proving the result for the case where $|T| < \infty$. Let $k_0 \in \mathbb{Z}$ be the largest integer such that $2^{-k_0} \geq \operatorname{diam}(T)$. It is clear that for every $t_0 \in T$, the set $N_0 := \{t_0\}$ forms a $2^{-k_0}-$net and $\pi_0(t) \equiv t_0$.

For $k > k_0$, let N_k be a $2^{-k}-$net such that $|N_k| = N(T, d, 2^{-k})$. We denote $\pi_k(t)$ as the element in N_k that satisfies $d(t, \pi_k(t)) \leq 2^{-k}$. Using a chaining argument up to the scale 2^{-n}, we proceed as follows

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in T} X_t\right] &= \mathbb{E}\left[\sup_{t\in T}\left\{X_{\pi_0(t)} + \left(\sum_{k=k_0+1}^{n} X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right) + X_t - X_{\pi_n}(t)\right\}\right] \\ &\leq \mathbb{E}[X_{t_0}] + \sum_{k=k_0+1}^{n} \mathbb{E}\left[\sup_{t\in T}\{X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\}\right] + \mathbb{E}\left[\sup_{t\in T}\{X_t - X_{\pi_n(t)}\}\right].\end{aligned}$$

By definition of subgaussian process, $\mathbb{E}[X_{t_0}] = 0$. Since $|T| < \infty$, we can choose n sufficiently large so that $N_n = T$, and hence $\pi_n(t) = t$, meaning that the third term vanishes. Next, we bound the second term. By definition, $X_{\pi_k(t)} - X_{\pi_{k-1}}(t)$ is a $d(\pi_k(t), \pi_{k-1}(t))$-subgaussian random variable. We can readily estimate the variance,

$$d(\pi_k(t), \pi_{k-1}(t)) \leq d(\pi_k(t), t) + d(t, \pi_{k-1}(t)) \leq 2^{-k} + 2^{-(k-1)} = 3 \times 2^{-k}.$$

Moreover, we can control the number of terms in the sum, note that $\{X_{\pi_k(t)} - X_{\pi_{k-1}(t)} : t \in T\}$ contains at most $|N_k||N_{k-1}|$ which is bounded by $|N_k|^2$ terms. Applying Maximal Inequality Lemma 3 to these terms, we obtain

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in T} X_t\right] &\leq \sum_{k=k_0+1}^{n} \sqrt{2d(\pi_k(t), \pi_{k-1}(t))^2 \log |N_k|^2} \leq 6 \sum_{k=k_0+1}^{n} 2^{-k}\sqrt{\log |N_k|} \\ &\leq 6 \sum_{k=k_0+1}^{n} 2^{-k}\sqrt{\log N(T, d, 2^{-k})}.\end{aligned}$$

To prove the result when T is infinite, let $T_0 = \{t_1, t_2, \dots\} \subseteq T$ be a countable set witnessing separability, so that $\sup_{t\in T} X_t = \sup_{t\in T_0} X_t$ a.s. For $m \geq 1$ let $T_m := \{t_1, \dots, t_m\}$. Then $\sup_{t\in T_m} X_t \uparrow \sup_{t\in T_0} X_t$ almost surely, and by monotone convergence,

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] = \mathbb{E}\left[\sup_{t\in T_0} X_t\right] = \sup_{m\geq 1} \mathbb{E}\left[\sup_{t\in T_m} X_t\right].$$

Applying the finite case to each T_m and using $N(T_m, d, \epsilon) \leq N(T, d, \epsilon)$ yields the same bound. □

Corollary (Entropy integral)**.** Let $\{X_t\}_{t\in T}$ be a separable subgaussian process on the metric space (T, d). Then,

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq 12 \int_0^\infty \sqrt{\log N(T, d, \epsilon)} d\epsilon.$$

Proof. Since $N(T, d, \cdot)$ is decreasing, we obtain the following chains of inequalities

$$\begin{aligned}\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log N(T, d, 2^{-k})} &= 2\sum_{k\in\mathbb{Z}} \int_{2^{-k-1}}^{2^{-k}} \sqrt{\log N(T, d, 2^{-k})} d\epsilon \\ &\leq 2\sum_{k\in\mathbb{Z}} \int_{2^{-k-1}}^{2^{-k}} \sqrt{\log N(T, d, \epsilon)} d\epsilon \\ &= 2\int_0^\infty \sqrt{\log N(T, d, \epsilon)} d\epsilon.\end{aligned}$$

□

References

Richard Nickl and Benedikt M. Pötscher. Bracketing Metric Entropy Rates and Empirical Central Limit Theorems for Function Classes of Besov- and Sobolev-Type. *Journal of Theoretical Probability*, 20(2):177–199, June 2007. ISSN 1572-9230. doi: 10.1007/s10959-007-0058-1. URL `https://doi.org/10.1007/s10959-007-0058-1`.

Ramon Van Handel. Probability in High Dimension:. Technical report, Fort Belvoir, VA, June 2014. URL `http://www.dtic.mil/docs/citations/ADA623999`.