CS 839 Probability and Learning in High Dimension Lecture 14 - 21/03/2022

Lecture 14: Random Processes: Chaining and Additional Tools

Lecturer: Yudong Chen *Scribe: Changho Shin*

In this lecture,[1] we will introduce Dudley's upper bound on the supremum $\mathbb{E}\sup_{\theta\in T} Z_\theta$. The upper bound is proved via a Chaining argument. We then apply Dudley's bound to derive a uniform law of large numbers. Finally, we will discuss additional tools for studying the suprema of random processes.

1 Dudley's Upper Bound

Recall: the process $(Z_\theta)_{\theta\in T}$ is said to have sub-Gaussian increment w.r.t. the metric ρ if for each $\theta, \theta' \in T$, $Z_\theta - Z_{\theta'}$ is sub-Gaussian with parameter $\rho(\theta,\theta')^2$. We have the following upper bound.

Theorem 1 (Dudley's entropy integral bound)**.** *Suppose that $(Z_\theta)_{\theta\in T}$ is zero-mean and has sub-Gaussian increment w.r.t. ρ. Then,*

$$\mathbb{E}\sup_{\theta\in T} Z_\theta \lesssim \int_0^\infty \sqrt{\log N(\varepsilon, T, \rho)}\, \mathrm{d}\varepsilon.$$

Remark We omit a separability assumption (so that we can take $\varepsilon \to 0$); See HW1 for details.

Remark Recall Sudakov's lower bound from last lecture:

$$\mathbb{E}\sup_{\theta\in T} Z_\theta \gtrsim \sup_{\varepsilon>0} \varepsilon\sqrt{\log N(\varepsilon, T, \rho)}.$$

Figure 1 provides a comparison between the upper bound in Theorem 1 and Sudakov's lower bound.

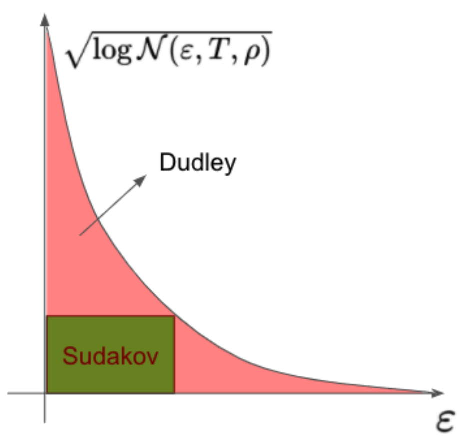

Figure 1: Dudley's inequality bounds $\mathbb{E}\sup_{\theta\in T} Z_\theta$ by the area under the curve. Sudakov's inequality bounds it below by the largest area of a rectangle under the curve, up to constants. Note that they are not necessarily tight — there can be a gap between the upper and lower bounds.

[1]References:

- Section 8.1 in "High -Dimensional Probability: An Introduction with Applications in Data Science", Roman Vershynin, Cambridge University Press, 2018.
- Section 5.3. in "High-Dimensional Statistics: A Non-Asymptotic Viewpoint", Martin J. Wainwright, Cambridge University Press, 2019.

The proof of Theorem 1 uses the "chaining" technique, a multi-scale ε-net argument. To motivate, consider one-step ε-net argument:

$$\sup_{\theta \in T} Z_\theta \leq \max_{\theta \in T_\varepsilon} Z_\theta + \sup_{\substack{\theta, \theta' \in T; \\ \rho(\theta, \theta') < \varepsilon}} |Z_\theta - Z_{\theta'}|.$$

We can bound 1st RHS term by finite Gaussian maxima, and 2nd term by some worst case bound. The chaining idea is to bound 2nd term also by an ε-net argument and repeat.

1.1 Proof of Theorem 1 by Chaining

Proof

First, some notations. Let $D \triangleq \sup_{\theta \in T} \rho(\theta, \theta')$ be diameter of T w.r.t. ρ. Define the dyadic scale

$$\varepsilon_k = D2^{-k}, \quad k = 0, 1, 2, \ldots$$

Let T_k be the smallest ε_k-net of T, so $|T_k| = \mathcal{N}(\varepsilon_k, T, \rho)$. For each $\theta \in T$, let $\pi_k(\theta)$ be the closest point in T_k, so

$$\rho(\theta, \pi_k(\theta)) \leq \varepsilon_k, \quad \forall \theta \in T, \forall k.$$

Note that $T_0 = \{\theta_0\}$ for some $\theta_0 \in T$, and $\pi_0(\theta) = \theta_0$, $\forall \theta \in T$.

Since the process is zero-mean, we have $\mathbb{E} \sup_{\theta \in T} Z_\theta = \mathbb{E} \sup_{\theta \in T} (Z_\theta - Z_{\theta_0})$. We write $Z_\theta - Z_{\theta_0}$ as a telescoping sum:

$$Z_\theta - Z_{\theta_0} = (Z_{\pi_1(\theta)} - Z_{\pi_0(\theta)}) + (Z_{\pi_2(\theta)} - Z_{\pi_1(\theta)}) + \cdots + (Z_\theta - Z_{\pi_M(\theta)}),$$

where $M > 0$ is a large constant. See Figure 2 for an illustration.

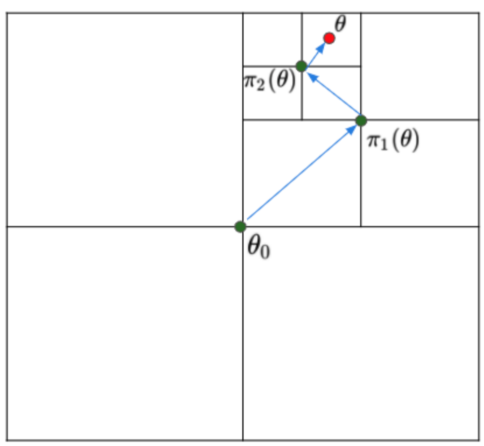

Figure 2: Illustration of chaining. A walk from a fixed point θ_0 to an arbitrary point θ in T along elements $\pi_k(\theta)$ of progressively finer nets of T

More succinctly, we have $Z_\theta - Z_{\theta_0} = \sum_{k=1}^{M} (Z_{\pi_k(\theta)} - Z_{\pi_{k-1}(\theta)}) + (Z_\theta - Z_{\pi_M(\theta)})$, which implies

$$\mathbb{E} \sup_{\theta \in T} (Z_\theta - Z_{\theta_0}) \leq \sum_{k=1}^{M} \mathbb{E} \sup_{\theta \in T} (Z_{\pi_k(\theta)} - Z_{\pi_{k-1}(\theta)}) + \mathbb{E} \sup_{\theta \in T} (Z_\theta - Z_{\pi_M(\theta)}). \tag{1}$$

Consider the k-th term in the summation above:

$$\mathbb{E}\sup_{\theta\in T}(\underbrace{Z_{\pi_k(\theta)}}_{\substack{|T_k|\\ \text{possible values}}} - \underbrace{Z_{\pi_{k-1}(\theta)}}_{\substack{|T_{k-1}|\\ \text{possible values}}}).$$

We see that this is the supremum of $|T_k|\cdot|T_{k-1}|$ random variables. For each fixed θ, the random variable $Z_{\pi_k(\theta)} - Z_{\pi_{k-1}(\theta)}$ is sub-Gaussian with parameter

$$\begin{aligned}\rho(\pi_k(\theta),\pi_{k-1}(\theta)) &\le \rho(\pi_k(\theta),\theta)+\rho(\pi_{k-1}(\theta),\theta)\\ &\le \varepsilon_k+\varepsilon_{k-1}\le 2\varepsilon_{k-1} \qquad \text{by triangle inequality and } \varepsilon_{k-1}>\varepsilon_k\end{aligned}$$

Therefore, we need to bound the maximum of finitely many random variables, each of which is sub-Gaussian with parameter $(2\epsilon_{k-1})^2$. Applying the bound on (sub-)Gaussian maximum from last lecture, we obtain

$$\begin{aligned}\mathbb{E}\sup_{\theta\in T}(Z_{\pi_k(\theta)} - Z_{\pi_{k-1}(\theta)}) &\lesssim \varepsilon_{k-1}\sqrt{\log(|T_k||T_{k-1}|)}\\ &\le \varepsilon_{k-1}\sqrt{\log|T_k|^2}\\ &= \varepsilon_{k-1}\sqrt{2\log\mathcal{N}(\varepsilon_k,T,\rho)}.\end{aligned}$$

Plugging these bounds into equation (1), we get

$$\begin{aligned}\mathbb{E}\sup_{\theta\in T}(Z_\theta - Z_{\theta_0}) &\lesssim \sum_{k=1}^{M}\varepsilon_{k-1}\sqrt{\log\mathcal{N}(\varepsilon_k,T,\rho)}+\mathbb{E}\sup_{\theta\in T}(Z_\theta - Z_{\pi_M(\theta)})\\ &\le \sum_{k=1}^{M} D2^{-(k-1)}\sqrt{\log\mathcal{N}(D2^{-k},T,\rho)}+\mathbb{E}\sup_{\theta\in T}(Z_\theta - Z_{\pi_M(\theta)})\\ &\lesssim \int_{D_{2^{-M-1}}}^{D}\sqrt{\log\mathcal{N}(\varepsilon,T,\rho)}d\varepsilon+\mathbb{E}\sup_{\theta\in T}(Z_\theta - Z_{\pi_M(\theta)}),\end{aligned}$$

where in the last step we bound sum by integral (for aesthetic consideration).

Let $M\to\infty$, then $\mathbb{E}\sup_{\theta\in T}(Z_\theta - Z_{\pi_M(\theta)})\to 0$ (require a separability assumption; See HW1). This completes the proof of Theorem 1. □

Exercise You may compare Theorem 1 with an upper bound obtained via one-step discretization, e.g., from Math 888 Fall 21, Lecture 18, Theorem 5.

> **Definition** The process $(X_t)_{t\in T}$ is L-Lipschitz if there exists a random variable L such that $|X_\theta - X_{\theta'}|\le L\rho(\theta,\theta')$ for all $\theta,\theta'\in T$ almost surely.
>
> **(Math 888 Fall 21, Lecture 18, Theorem 5)**. Suppose that a random process $(X_\theta)_{\theta\in T}$ is L-Lipschitz, mean zero, and that $\|X_\theta\|_{\psi_2}\le\sigma$ for all $\theta\in T$. Then
>
> $$\mathbb{E}\sup_{\theta\in T}X_\theta \lesssim \inf_{\epsilon>0}\left\{\epsilon\,\mathbb{E}[L]+\sigma\sqrt{\log\mathcal{N}(\varepsilon,T,\rho)}\right\}.$$

2 Application: Uniform Law of Large Numbers

Let $X_1, \cdots, X_n$ be i.i.d. unif$[0,1]$ random variables. For a fixed function f, the usual law of large numbers ensures that

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \mathbb{E}\, f(X_1) \text{ as } n \to \infty, \quad \text{almost surely.}$$

Can we prove convergence uniformly over a class of functions $\mathcal{F}$? Below we use Dudley's upper bound to derive one such result,

Theorem 2. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables taking values in $[0,1]$, and $\mathcal{F} := \{f : [0,1] \to \mathbb{R}, f$ is 1-Lipschitz$\}$. Then*

$$\mathbb{E} \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X_1) \right| \lesssim \frac{1}{\sqrt{n}}.$$

Remark (Connection to Wasserstein Distance) Let μ be the distribution of X_i, and let μ_n be the empirical distribution defined as

$$\mu_n := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i}.$$

Note that μ_n is a random quantity. With this notation, the LHS in Theorem 2 can be written as

$$\mathbb{E} \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X_1) \right| = \mathbb{E} \sup_{f\in\mathcal{F}} \left| \int f \mathrm{d}\mu_n - \int f \mathrm{d}\mu \right|,$$

which is the *Wasserstein distance* between μ_n and μ. (The definition is equivalent to the one using transportation cost, by Kantorovich-Rubinstein duality).

2.1 Proof of Theorem 2

Proof Observe that

$$\forall f \in \mathcal{F} : \left| \sup_x f(x) - \inf_x f(x) \right| \leq 1.$$

Therefore, without loss of generality, it suffices to consider 1-Lipschitz functions of the form $f : [0,1] \to [0,1]$; otherwise, just shift the function by letting $f' = f - \inf_x f(x)$.

Consider the empirical process $(Z_f)_{f\in\mathcal{F}}$ where

$$Z_f \triangleq \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X_1).$$

Clearly, the $\mathbb{E}[Z_f] = 0$. Moreover, for each $f, g \in \mathcal{F}$, we have

$$Z_f - Z_g = \frac{1}{n}\sum_{i=1}^{n}(f-g)(X_i) - \mathbb{E}(f-g)(X_1).$$

It follows that

$$
\begin{aligned}
\underbrace{\|Z_f - Z_g\|_{\psi_2}}_{\substack{\text{sub-Gaussian}\\ \text{parameter of } Z_f - Z_g}} &\lesssim \left\| \frac{1}{n}\sum_{i=1}^{n}(f-g)(x_i) \right\|_{\psi_2} && \text{(Centering does not change sub-Gaussian parameter, up to a constant)}\\
&\lesssim \frac{1}{n}\sqrt{\sum_{i=1}^{n} \|(f-g)(x_i)\|_{\psi_2}^2} && \text{(Hoeffding)}\\
&\lesssim \frac{1}{n}\sqrt{\sum_{i=1}^{n} \|f-g\|_{\infty}^2} && \text{(Bounded RVs are sub-Gaussian)}\\
&= \frac{1}{\sqrt{n}}\|f-g\|_{\infty}.
\end{aligned}
$$

We conclude that the process $(Z_f)_{f\in\mathcal{F}}$ has sub-Gaussian increments w.r.t. $\rho(f,g) := \|f-g\|_{\infty}/\sqrt{n}$. Applying Dudley's upper bound (Theorem 1), we obtain

$$
\mathbb{E}\sup_{f\in\mathcal{F}}|Z_f| \lesssim \frac{1}{\sqrt{n}}\int_0^1 \sqrt{\log\mathcal{N}(\varepsilon,\mathcal{F},\|\cdot\|_{\infty})}\,\mathrm{d}\varepsilon, \tag{2}
$$

where we use that fact that diameter$(\mathcal{F}) \le 1$ so the upper limit of the integral can be taken to be 1.

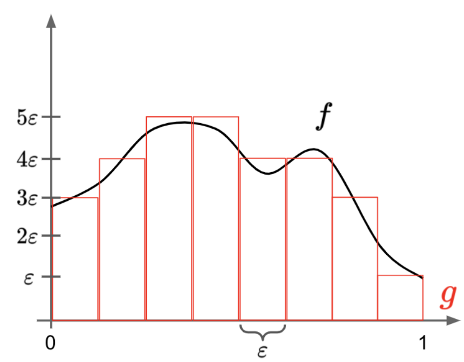

Figure 3: Illustration of covering $\mathcal{F}$ with step functions g's

It remains to bound the covering number $\mathcal{N}(\varepsilon,\mathcal{F},\|\cdot\|_{\infty})$. Here we construct an exterior ε-net $\mathcal{F}_\varepsilon$ of $\mathcal{F}$ (i.e., $\mathcal{F}_\varepsilon$ is not necessarily a subset of $\mathcal{F}$); construction of a usual ε-net is left to HW 1. In particular ,we can cover $\mathcal{F}$ using step functions g's as illustrated in Figure 3. The function g satisfies

$$
\begin{aligned}
\|f-g\|_{\infty} = \sup_{x\in[0,1]}|f(x)-g(x)| &\le 2\max_{k=0,1,\cdots,\frac{1}{\varepsilon}}\ \sup_{x\in[k\varepsilon,(k+1)\varepsilon]}|f(x)-g(x)|\\
&\le \sup_{|x-y|\le\varepsilon}|f(x)-f(y)| \le \varepsilon,
\end{aligned}
$$

so it indeed covers $\mathcal{F}$ in $\|\cdot\|_{\infty}$ norm up to an ϵ error. It is easy to see that $|\mathcal{F}_\varepsilon| \le \left(\frac{1}{\varepsilon}\right)^{1/\varepsilon}$, hence

$$
\log\mathcal{N}(\varepsilon,\mathcal{F},\|\cdot\|_{\infty}) \le \log|\mathcal{F}_\varepsilon| = \frac{1}{\varepsilon}\log\frac{1}{\varepsilon}.
$$

Plugging this bound into equation (2), we obtain

$$\mathbb{E}\sup_{f\in\mathcal{F}} Z_f \lesssim \frac{1}{\sqrt{n}}\int_0^1 \sqrt{\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}}d\varepsilon \lesssim \frac{1}{\sqrt{n}}$$

as desired. □

2.2 Tail Bound Version

Using Theorem 2, we can further obtain a tail bound version of the uniform law of large numbers.

Theorem 3. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables taking values in $[0,1]$, and $\mathcal{F} := \{f : [0,1] \to \mathbb{R}, f$ is 1-Lipschitz$\}$. Then for any $t \geq 0$, we have*

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}\, f(X_1)\right| \lesssim \frac{1}{\sqrt{n}} + t$$

with probability at least $1 - 2\exp(-2nt^2)$. Consequently, we have

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}\, f(X_1)\right| \to 0 \quad \textit{as } n \to \infty, \quad \textit{almost surely.}$$

Proof In order to simplify notation, define the centered functions $\bar{f}(x) \triangleq f(x) - \mathbb{E}[f(X_1)]$. Thinking of the samples $\{X_i\}$ as fixed for the moment, consider the function

$$G(x_1, ...x_n) \triangleq \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \bar{f}(x_i)\right|.$$

We claim that G satisfies the property required to apply the bounded differences inequality. Since the function G is invariant to permutation of its coordinates, it suffices to bound the difference when the first coordinate x_1 is perturbed. Accordingly, we define the vector $y \in \mathbb{R}^n$ with $y_i = x_i$ for all $i \neq 1$, and seek to bound the difference $|G(x) - G(y)|$. For any function $\bar{f} = f - \mathbb{E}[f]$ with $f \in \mathcal{F}$, we have

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n \bar{f}(x_i)\right| - \sup_{g\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \bar{g}(y_i)\right| \\
\leq &\left|\frac{1}{n}\sum_{i=1}^n \bar{f}(x_i)\right| - \left|\frac{1}{n}\sum_{i=1}^n \bar{f}(y_i)\right| \\
\leq &\frac{1}{n}\left|\bar{f}(x_1) - \bar{f}(y_1)\right| && x_i = y_i \text{ except for } i = 1 \\
\leq &\frac{1}{n}. && |\bar{f}(x_1) - \bar{f}(y_1)| = |f(x_1) - f(y_1)| \leq 1 \text{ because } f \text{ is 1-Lipschitz}
\end{aligned}$$

Since the above inequality holds for any function $f \in \mathcal{F}$, we may take the supremum over $f \in \mathcal{F}$ on both sides, which yields $G(x) - G(y) \leq \frac{1}{n}$. Since the same argument may be applied with the n roles of x and y reversed, we conclude that $|G(x) - G(y)| \leq 1$. Then, by the bounded difference inequality (Lecture 12, Theorem 4), we have

$$
\begin{aligned}
&\Pr\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X_1)\right| \gtrsim \frac{1}{\sqrt{n}} + t\right) \\
\le&\Pr\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X_1)\right| \ge \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X_1)\right| + t\right) && \text{By Theorem 2} \\
=&\Pr\Big(G(X_1,\dots,X_n) \ge \mathbb{E}\left[G(X_1,\dots,X_n)\right] + t\Big) \\
\le&2\exp(-2nt^2), && \text{Bounded difference inequality}
\end{aligned}
$$

valid for any $t \ge 0$. This proves the first part of the theorem. Combining with the Borel-Cantelli Lemma, we establish the second part on almost sure convergence. □

Appendices

A Supremum of Random Processes: Additional Tools

In this section, we discuss additional techniques for studying the supremum of random processes.

References:

- Chapter 8.5 in High -Dimensional Probability: An Introduction with Applications in Data Science, Roman Vershynin, Cambridge University Press, 2018.
- Section 4.2, 5.4.3 in High-Dimensional Statistics: A Non-Asymptotic Viewpoint, Martin J. Wainwright, Cambridge University Press, 2019.
- (Additional reading) Probability in High Dimension: APC 550 Lecture Notes, Ramon van Handel, Princeton University, 2016

A.1 Generic Chaining

Sudakov's lower bound and Dudley's upper bound are both loose in the worst case. It is possible to obtain tight bounds using the generic chaining technique.

Consider a metric space (T, ρ). An *admissible sequence* is a sequence of sets $(T_k, k = 0, 1, \dots)$ with $T_k \subset T$ and $|T_k| = 2^{2^k}$ (and as a convention $|T_0| = |\{\theta_0\}| = 1$.) Define the γ_2 functional

$$\gamma_2(T, \rho) := \inf_{(T_k)} \sup_{\theta\in T} \sum_{k=0}^{\infty} 2^{k/2} \cdot \rho(\theta, T_k),$$

where the infimum above is over all admissible sequences and $\rho(\theta, T_k) := \inf_{\theta'\in T_k} \rho(\theta, \theta')$. (Note that the supremum above is *outside* the summation; compare with the proof of Dudley.)

We have the following upper and lower bounds in terms of γ_2. The upper bound applies to any *sub-Gaussian* process.

Theorem 4 (Generic chaining upper bound)**.** *If $(Z_\theta)_{\theta\in T}$ is a zero-mean process with sub-Gaussian increment w.r.t. some ρ, then*

$$\mathbb{E}\sup_{\theta\in T} Z_\theta \lesssim \gamma_2(T, \rho).$$

The lower bound applies to *Gaussian* processes.

Theorem 5 (Talagrand's majorizing measure theorem)**.** *If $(Z_\theta)_{\theta\in T}$ is a zero-mean Gaussian process with metric $\rho(\theta,\theta') := \sqrt{\mathbb{E}(Z_\theta - Z_{\theta'})^2}$, then*

$$\mathbb{E}\sup_{\theta\in T} Z_\theta \gtrsim \gamma_2(T,\rho).$$

For Gaussian processes, we see that the upper and lower bounds match up to a universal constant.

In general, the quantity $\gamma_2(T,\rho)$ is more difficult to compute than metric entropy integral. However, even without knowing how to compute γ_2, we can still deduce from the above theorems the following very useful comparison inequality.

Corollary 1 (Talagrand's sub-Gaussian comparison inequality)**.** *If $(X_\theta)_{\theta\in T}$ is a zero-mean process with sub-Gaussian increment w.r.t. some ρ, $(Y_\theta)_{\theta\in T}$ is a zero-mean Gaussian process, and*

$$\rho(\theta,\theta') \lesssim \sqrt{\mathbb{E}(Y_\theta - Y_{\theta'})^2}$$

then

$$\mathbb{E}\sup_{\theta\in T} X_\theta \lesssim \mathbb{E}\sup_{\theta\in T} Y_\theta.$$

Remark Corollary 1 allows one to reduce a sub-Gaussian problem to a Gaussian one, for which we have many tools.

Remark A special case of Corollary 1 is when $X_\theta = \langle \epsilon, \theta\rangle$ is canonical Rademacher process with $\epsilon \sim \operatorname{unif}\{\pm 1\}^n$, and $Y_\theta = \langle g, \theta\rangle$ is a canonical Gaussian process with $g \sim N(0, I_n)$.

A.2 Contraction

Below, we assume that $\epsilon \sim \operatorname{unif}\{\pm 1\}^n$ and $g \sim N(0, I_n)$ are vectors of iid Rademacher and standard Gaussian variables, respectively.

Theorem 6 (Gaussian Contraction Principle)**.** *Let $T \subset \mathbb{R}^n$ and $\phi_i : \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz for $i = 1, \ldots, n$. Then*

$$\mathbb{E}\sup_{\theta\in T}\sum_{i=1}^n g_i\phi_i(\theta_i) \le \mathbb{E}\sup_{\theta\in T}\sum_{i=1}^n g_i\theta_i.$$

Proof We shall use Gaussian comparison inequality to compare the two Gaussian processes

$$X_\theta = \sum_{i=1}^n g_i\phi_i(\theta_i) \qquad \text{and} \qquad Y_\theta = \sum_{i=1}^n g_i\theta_i.$$

For $\theta, \tilde\theta \in T$, the corresponding increments satisfy

$$\begin{aligned}
\mathbb{E}\left(X_\theta - X_{\tilde\theta}\right)^2 &= \sum_{i=1}^n \left(\phi_i(\theta_i) - \theta_i(\tilde\theta_i)\right)^2 \\
&\le \sum_{i=1}^n \left(\theta_i - \tilde\theta_i\right)^2 && \phi_i \text{ is 1-Lipschitz} \\
&= \mathbb{E}\left(Y_\theta - Y_{\tilde\theta}\right)^2.
\end{aligned}$$

Applying Sudakov-Fernique Gaussian comparison inequality proves the theorem. □

We also have a Rademacher version of the contraction inequality.

Theorem 7 (Ledoux-Talagrand Contraction Principle)**.** *Let $T \subset \mathbb{R}^n$ and $\phi_i : \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz and centered ($\phi_i(0) = 0$) for $i = 1, \ldots, n$. Then*

$$\mathbb{E} \sup_{\theta \in T} \left| \sum_{i=1}^{n} \epsilon_i \phi_i(\theta_i) \right| \le 2 \mathbb{E} \sup_{\theta \in T} \left| \sum_{i=1}^{n} \epsilon_i \theta_i \right|.$$

There is no Rademacher version of the Sudakov-Fernique inequality, so the proof of Theorem 7 is more involved and we will not present it here.

Remark The LHS of the bound in Theorem 7 can be written as a canonical process's supremum, $\mathbb{E} \sup_{\beta \in \phi(T)} |\sum_i \epsilon_i \beta_i|$, where

$$\phi(T) := \Big\{ \big(\phi_1(\theta_1), \ldots, \phi_n(\theta_n) \big) : \theta \in T \Big\}.$$

Therefore, Theorem 7 says that when ϕ is 1-Lipschitz, the composite set $\phi(T)$ is "no larger" than the original (and usually simpler) set T, in the sense of process supremum.

A.3 Symmetrization

We have seen many tools for Gaussian and Rademacher processes, including various concentration, comparison and contraction inequalities. Below we discuss symmetrization, which allows one to *extract Gaussianity* (or Rademacher randomness) from a general process.

Again assume that $\epsilon \sim \text{unif}\{\pm 1\}^n$ and $g \sim N(0, I_n)$ are vectors of iid Rademacher and standard Gaussian variables, respectively, that are independent of everything else.

Theorem 8 (Symmetrization)**.** *Let $X_1, \ldots, X_n$ be i.i.d. RVs taking values in $\mathbb{X}$, and $\mathcal{F}$ be a class of functions on $\mathbb{X}$. Then we have*

$$\mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \{ f(X_i) - \mathbb{E} f(X_i) \} \right] \overset{(a)}{\le} 2 \mathbb{E}_{X,\epsilon} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \epsilon_i f(X_i) \right] \overset{(b)}{\le} \sqrt{2\pi} \mathbb{E}_{X,g} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} g_i f(X_i) \right]$$

and

$$\mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \{ f(X_i) - \mathbb{E} f(X_i) \} \right| \right] \overset{(c)}{\le} 2 \mathbb{E}_{X,\epsilon} \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \right] \overset{(d)}{\le} \sqrt{2\pi} \mathbb{E}_{X,g} \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} g_i f(X_i) \right| \right].$$

Inequalities (a) and (b) can be found as Lemma 7.4 in van Handel's book "Probability in High Dimension". Below we prove (c) and (d).

Proof [Proof of (c) and (d)]

Let $(Y_1, \ldots, Y_n)$ be an independent copy of $(X_1, \ldots, X_n)$. We have the following chain of inequlaities

$$\begin{aligned}
&\mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \{ f(X_i) - \mathbb{E} f(X_i) \} \right| \right] \\
=&\mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \{ f(X_i) - \mathbb{E}_Y f(Y_i) \} \right| \right] && X_i \overset{\mathrm{d}}{=} Y_i \\
\le&\mathbb{E}_X \mathbb{E}_Y \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \{ f(X_i) - f(Y_i) \} \right| \right] && \text{Jensen's} \\
=&\mathbb{E}_X \mathbb{E}_Y \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \epsilon_i \{ f(X_i) - f(Y_i) \} \right| \right] && f(X_i) - f(Y_i) \overset{\mathrm{d}}{=} \epsilon_i \{ f(X_i) - f(Y_i) \} \\
\le&2 \mathbb{E}_X \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \right]. && \text{triangle inequality}
\end{aligned}$$

Above, $\overset{\mathrm{d}}{=}$ means "equal in distribution". We have proved (a).

We recall that $\mathbb{E}\,|g_i| = \sqrt{\frac{2}{\pi}}$ from the property of half Normal distribution, so

$$\begin{aligned} 2\mathbb{E}_{X,\epsilon}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\right] &= 2\sqrt{\frac{\pi}{2}}\mathbb{E}_{X,\epsilon}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i\cdot\mathbb{E}\,|g_i|\cdot f(X_i)\right|\right] \\ &\le \sqrt{2\pi}\mathbb{E}_{X,\epsilon,g}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i\cdot|g_i|\cdot f(X_i)\right|\right] && \text{Jensen's inequality} \\ &= \sqrt{2\pi}\mathbb{E}_{X,g}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}g_i\cdot f(X_i)\right|\right]. && g_i \overset{\mathrm{d}}{=} \epsilon_i\,|g_i| \end{aligned}$$

We have proved (d). □

The symmetrization argument is typically used by *conditioning on* (X_i). For example, we can write

$$\mathbb{E}_{X,g}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}g_i f(X_i)\right] = \mathbb{E}\left[\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}g_i f(X_i)\,\middle|\,X_1,\ldots,X_n\right]\right].$$

Conditioned on (X_i), the quantity $\sum_{i=1}^{n} g_i f(X_i)$ is Gaussian, so one can bound the inner expectation using any results for Gaussian processes.