Metric Entropy and Its Uses

zhenduo li, han zhang

2022.10.5

Table of Contents

Motivation

Recall the Glivenko-Cantelli law via Rademacher complexity:

Theorem 4.10

For any b-uniformly bounded class of functions $\mathcal{F}$, any positive integer $n \geq 1$ and any scalar $\delta \geq 0$, we have

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2\mathcal{R}_n(\mathcal{F}) + \delta$$

with $\mathcal{P}$-probability at least $1 - exp(-\frac{n\delta^2}{2b^2})$. Here $\mathcal{R}_n(\mathcal{F})$ is the Rademacher complexity of $\mathcal{F}$.

Bounding the Rademacher complexity of $\mathcal{F}$:

- Polynomial discrimination $\Rightarrow$ additional control of the l_2-radius;
- VC dimension $\Rightarrow$ exclusive for classes of binary-valued functions.

Motivation

Other random variables indexed by a set, say $\mathbb{T}$:

- Operator norm of a random matrix

$$\|\hat{A} - A\|_{op} := \sup_{\|x\|_2=1} \|(\hat{A} - A)x\|_2, \ \mathbb{T} = \{x \in \mathbb{R}^p, \|x\|_2 = 1\}.$$

- Loss function indexed by parameters

$$(\mathbb{P}_n - \mathbb{P})f_\theta, \ \theta \in \mathbb{T} \subset \mathbb{R}^q.$$

Target: bounds for more general classes

Motivation

Given a metric space $(\mathbb{T}, \rho)$, $\mathbb{T} \neq \emptyset$ and $\rho : \mathbb{T} \times \mathbb{T} \to \mathbb{R}$ is a metric. How to measure the size of $\mathbb{T}$?

$\Rightarrow$ Cover a set with balls that are centered inside and of the same radius, and count the number of these balls.

Covering number and packing number

Definition 5.1 (Covering number)

A δ-cover of a set $\mathbb{T}$ with respect to a metric ρ is a set $\{\theta^1, \ldots, \theta^n\} \subset \mathbb{T}$ such that for each $\theta \in \mathbb{T}$, there exists some $i \in \{1, \ldots, N\}$ such that $\rho(\theta, \theta^i)$. The δ-covering number $N(\delta; \mathbb{T}, \rho)$ is the cardinality of the smallest δ-cover.

Definition 5.4 (Packing number)

A δ-packing of a set $\mathbb{T}$ with respect to a metric ρ is a set $\{\theta^1, \ldots, \theta^M\} \subset \mathbb{T}$ such that $\rho(\theta^i, \theta^j)\rangle\delta$ for all distinct $i, j \in \{1, 2 \ldots, M\}$. The δ-packing number $M(\delta; \mathbb{T}, \rho)$ is the cardinality of the largest δ-packing.

Covering number and packing number

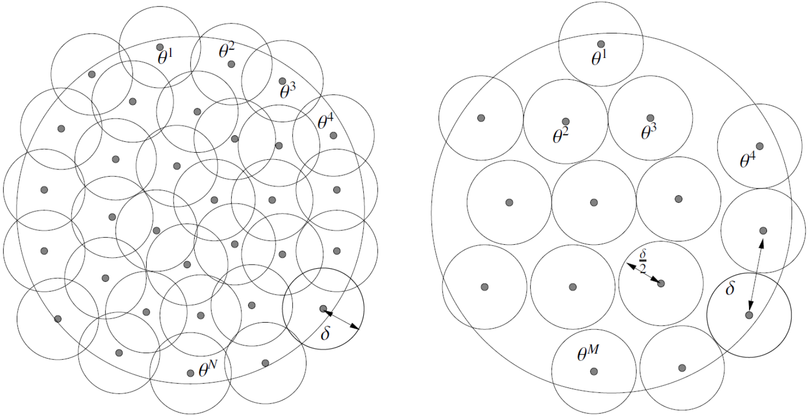

Figure: The δ-covering (left) and δ-packing (right) of a circular

Covering number and packing number

- If $N(\delta; \mathbb{T}, \rho) < \infty$ for all $\delta > 0$, we say $\mathbb{T}$ is totally bounded,
- For a totally bounded set $\mathbb{T}$ with respect to ρ, $\log N(\delta; \mathbb{T}, \rho)$ is called the metric entropy of $\mathbb{T}$ with respect to ρ.
- The covering and packing numbers are essentially equivalent.

Lemma 5.5

For all $\delta > 0$, the packing and covering numbers are related as follows:

$$M(2\delta; \mathbb{T}, \rho) \leq N(\delta; \mathbb{T}, \rho) \leq M(\delta; \mathbb{T}, \rho).$$

Estimation

Example 5.2 & 5.6 (unit cubes)
Set $\mathbb{T} = [-1, 1]$, $\rho(\theta, \theta') = |\theta - \theta'|$. Let $\theta^i = -1 + 2(i-1)\delta$ for $i = 1, \ldots, L$. Here $L := \lfloor \frac{1}{\delta} \rfloor + 1$. Then $\{\theta^1, \ldots, \theta^L\}$ forms a δ-cover of $\mathbb{T}$ with respect to ρ, which implies

$$N(\delta; [-1, 1], |\cdot|) \leq \frac{1}{\delta} + 1.$$

Note that $\{\theta^1, \ldots, \theta^{L-1}\}$ is a 2δ-packing of $\mathbb{T}$, by Lemma 5.5 we have

$$N(\delta; [-1, 1], |\cdot|) \geq M(2\delta; [-1, 1], |\cdot|) \geq \lfloor \frac{1}{\delta} \rfloor.$$

To sum up, we have $N(\delta; [-1, 1], |\cdot|) \sim 1/\delta$.

Estimation

Lemma 5.7

(Volume ratios and metric entropy) Consider a pair of norms $\|\cdot\|$ and $\|\cdot\|'$ on $\mathbb{R}^d$, and let $\mathbb{B}$ and $\mathbb{B}'$ be their corresponding unit balls (i.e., $\mathbb{B} = \{\theta \in \mathbb{R}^d \mid \|\theta\| \leq 1\}$, with $\mathbb{B}'$ similarly defined). Then the δ-covering number of $\mathbb{B}$ in the $\|\cdot\|'$-norm obeys the bounds

$$\left(\frac{1}{\delta}\right)^d \frac{\mathrm{vol}(\mathbb{B})}{\mathrm{vol}(\mathbb{B}')} \leq N(\delta; \mathbb{B}, \|\cdot\|') \leq \frac{\mathrm{vol}(\frac{2}{\delta}\mathbb{B} + \mathbb{B}')}{\mathrm{vol}(\mathbb{B}')}$$

Example 5.8

Taking $\mathbb{B} = \mathbb{B}'$ yields bounds on the metric entropy of a unit ball in terms of its own metrics:

$$d \log(1/\delta) \leq \log N(\delta; \mathbb{B}, \|\cdot\|) \leq d \log(1 + 2/\delta).$$

Estimation

Example 5.9 (A parametric class of functions)
Define $f_\theta(x) = 1 - e^{-\theta x}$, consider the function class

$$\mathscr{F} := \{f_\theta : [0,1] \to \mathbb{R} \mid \theta \in [0,1]\}.$$

A bound for $N(\delta; \mathscr{F}, \|\cdot\|_\infty)$ can be estimated as following.

Upper bound

Given $\delta \in (0,1)$, let $T = \lfloor \frac{1}{2\delta} \rfloor$, $\theta^i = 2\delta i$ for $i = 0, 1, \ldots, T$ and $\theta^{T+1} = 1$.
Then $\{f_{\theta^0}, \ldots, f_{\theta^{T+1}}\}$ form a δ-cover for $\mathscr{F}$. Thus,

$$\|f_{\theta^i} - f_\theta\|_\infty = \max_{x \in [0,1]} |e^{-\theta^i |x|} - e^{-\theta |x|}| \leq |\theta^i - \theta| \leq \delta,$$

which implies that $N(\delta; \mathscr{F}, \|\cdot\|_\infty) \leq T + 2 \leq \frac{1}{2\delta} + 2$.

Estimation

Lower bound

Let $\theta_0 = 0$, $\theta^i = -\log(1-\delta i)$ for all i such that $\theta^i \leq 1$ and $\theta^{T+1} = 1$. We have $\|f_{\theta^i} - f_{\theta^j}\|_\infty \geq |f_{\theta^i}(1) - f_{\theta^j}(1)| \geq \delta$.
Thus, $\{f_{\theta^0}, \ldots, f_{\theta^{T+1}}\}$ is a δ-packing of $(\mathscr{F}, \|\cdot\|_\infty)$ and $M(\delta; \mathscr{F}, \|\cdot\|_\infty) \geq \lfloor \frac{1-1/e}{\delta} \rfloor + 1$.
By Lemma 5.5, we have

$$N(\delta; \mathscr{F}, \|\cdot\|_\infty) \geq M(2\delta; \mathscr{F}, \|\cdot\|_\infty) \geq \lfloor \frac{1-1/e}{2\delta} \rfloor + 1.$$

Application

Discretization (δ-cover argument): Operator norm on a cover

Let A be an $m \times n$ random matrix. Then, for any δ-cover $\mathcal{N}$ of

$$\mathbb{S}^{n-1} := \{\theta \in \mathbb{R}^n : \|\theta\|_2 = 1\},$$

we have

$$\sup_{x \in \mathcal{N}} \|Ax\|_2 \leq \|A\|_{op} \leq \frac{1}{1-\delta} \sup_{x \in \mathcal{N}} \|Ax\|_2.$$

The lower bound is trivial by the definition of the operator norm.
To prove the upper bound, fix a $x \in \mathbb{S}^{n-1}$ for which $\|A\|_{op} = \|Ax\|_2$, we can find a $x_0 \in \mathcal{N}$ such that $\|x - x_0\|_2 \leq \delta$. By the definition of the operator norm, this implies

$$\|Ax - Ax_0\|_2 \leq \|A\|_{op}\|x - x_0\|_2 \leq \delta\|A\|_{op}.$$

Using triangle inequality, we have

$$\|Ax_0\|_2 \geq \|Ax\|_2 - \|Ax - Ax_0\|_2 \geq \|A\|_{op} - \delta\|A\|_{op} = (1-\delta)\|A\|_{op}.$$

Table of Contents

Gaussian and Rademacher complexity

Consider two random process G_θ and R_θ indexed by $\theta \in \mathbb{T} \subset \mathbb{R}^d$:

$$G_\theta := \langle w,\ \theta \rangle = \sum_{i=1}^{d} w_i \theta_i,$$

$$R_\theta := \langle \varepsilon,\ \theta \rangle = \sum_{i=1}^{d} \varepsilon_i \theta_i,$$

where $w_1, \ldots, w_d \overset{i.i.d}{\sim} N(0,1)$ and $\varepsilon_1, \ldots, \varepsilon_d \overset{i.i.d}{\sim} U(\{-1,1\})$. G_θ and R_θ are called, respectively, the canonical Gaussian process and the Rademacher process associated with $\mathbb{T}$.

The Gaussian complexity and the Rademacher complexity of $\mathbb{T}$ are defined, respectively, as

$$\mathcal{G}(\mathbb{T}) := \mathbb{E}(\sup_{\theta \in \mathbb{T}} G_\theta),$$

$$\mathcal{R}(\mathbb{T}) := \mathbb{E}(\sup_{\theta \in \mathbb{T}} R_\theta).$$

Estimation

Example 5.13 (Rademacher/Gaussian complexity of Euclidean ball)
Let $\mathbb{B}_2^d = \{\theta \in \mathbb{R}^d \mid \|\theta\|_2 \leq 1\}$.
By the Cauchy–Schwarz inequality,

$$\mathcal{R}(\mathbb{B}_2^d) = \mathbb{E}(\sup_{\|\theta\|_2 \leq 1} R_\theta) = \mathbb{E}\left[\left(\sum_{i=1}^d \varepsilon_i^2\right)^{1/2}\right] = \sqrt{d}.$$

By Jensen's inequality,

$$\mathcal{G}(\mathbb{B}_2^d) = \mathbb{E}(\sup_{\|\theta\|_2 \leq 1} G_\theta) = \mathbb{E}\left[\left(\sum_{i=1}^d w_i^2\right)^{1/2}\right] \leq \sqrt{\mathbb{E}\|w\|_2^2} = \sqrt{d}.$$

$*$ It can be shown that $\mathbb{E}\|w\|_2 \geq \sqrt{d}(1 - o(1))$, so the Rademacher and Gaussian complexities of $\mathbb{B}_2^d$ are essentially equivalent.

Estimation

Example 5.14 (Rademacher/Gaussian complexity of l_1-ball)
Let $\mathbb{B}_1^d = \{\theta \in \mathbb{R}^d \mid \|\theta\|_1 \leq 1\}$.
By Hölder's inequality,

$$\mathcal{R}(\mathbb{B}_1^d) = \mathbb{E}(\sup_{\|\theta\|_1 \leq 1} R_\theta) = \mathbb{E}\|\varepsilon\|_\infty = 1.$$

By the inequality for Gaussian maxima in Exercise 2.11,

$$\mathcal{G}(\mathbb{B}_1^d) = \mathbb{E}(\sup_{\|\theta\|_1 \leq 1} G_\theta) = \mathbb{E}\|w\|_\infty = \sqrt{2\log d} + o(1).$$

Table of Contents

Sub-Gaussian processes

The canonical Gaussian process and the Rademacher process are particular examples of sub-Gaussian processes defined as following.

Definition 5.16

A collection of zero-mean random variables $\{X_\theta, \theta \in \mathbb{T}\}$ is a sub-Gaussian process with respect to a metric ρ on $\mathbb{T}$ if

$$\mathbb{E}[e^{\lambda(X_\theta - X_{\tilde{\theta}})}] \leq e^{\frac{\lambda^2 \rho^2(\theta,\tilde{\theta})}{2}} \quad \text{for all } \theta, \tilde{\theta} \in \mathbb{T} \text{ and } \lambda \in \mathbb{R}.$$

$*$ The chernoff bound implies an equivalent way to define a sub-Gaussian process by increment:

$$\mathbb{P}(|X_\theta - X_{\tilde{\theta}}| \geq t) \leq 2e^{-\frac{t^2}{2\rho^2(\theta,\tilde{\theta})}}.$$

$*$ Let $\{X_\theta, \theta \in \mathbb{T}\}$ be a Gaussian process, define $\rho(\theta, \tilde{\theta}) = \|X_\theta - X_{\tilde{\theta}}\|_{L^2}$, then $\{X_\theta, \theta \in \mathbb{T}\}$ is a sub-Gaussian process.

Upper bound by one-step discretization

Proposition 5.17 (One-step discretization bound)

Denote by $D := \sup_{\theta,\tilde{\theta}\in\mathbb{T}} \rho(\theta,\tilde{\theta})$ *the diameter of* $\mathbb{T}$. *Let* $\{X_\theta, \theta \in \mathbb{T}\}$ *be a zero-mean sub-Gaussian process with respect to the metric* ρ. *Then for any* $\delta \in [0, D]$ *such that* $N(\delta;\rho,\mathbb{T}) \geq 10$, *we have*

$$\mathbb{E}[\sup_{\theta,\tilde{\theta}\in\mathbb{T}} (X_\theta - X_{\tilde{\theta}})] \leq 2\mathbb{E}[\sup_{\substack{\gamma,\gamma'\in\mathbb{T},\\ \rho(\gamma,\gamma')\leq\delta}} (X_\gamma, X'_\gamma)] + 4\sqrt{D^2 \log N(\delta;\mathbb{T},\rho)}.$$

$*$ The zero-mean condition implies an upper bound on $\mathbb{E}[\sup_{\theta\in\mathbb{T}} X_\theta]$.
$*$ The second term is the error incurred by approximating $\mathbb{T}$ with its δ-cover (discretization).

Upper bound by one-step discretization

Sketched Proof

For a given $\delta \geq 0$, let $\{\theta^1, \ldots, \theta^N\}$ be a δ-cover of $\mathbb{T}$. Then

$$X_\theta - X_{\theta^1} \leq \sup_{\substack{\gamma,\gamma' \in \mathbb{T} \\ \rho(\gamma,\gamma') \leq \delta}} (X_\gamma - X_{\gamma'}) + \max_{i=1,\ldots,N} |X_{\theta^i} - X_{\theta^1}|.$$

The same bound holds for any other $\tilde{\theta} \in \mathbb{T}$, adding the two bounds together yields

$$\sup_{\theta,\tilde{\theta} \in \mathbb{T}} (X_\theta - X_{\tilde{\theta}}) \leq 2 \sup_{\substack{\gamma,\gamma' \in \mathbb{T} \\ \rho(\gamma,\gamma') \leq \delta}} (X_\gamma - X_{\gamma'}) + 2 \max_{i=1,\ldots,N} |X_{\theta^i} - X_{\theta^1}|.$$

By the sub-Gaussian nature of $\{X_\theta, \theta \in \mathbb{T}\}$ and the upper bound for sub-Gaussian maxima (Exercise 2.12(b)), we have

$$\mathbb{E}\left(\max_{i=1,\ldots,N} |X_\theta^i - X_\theta^1| \right) \leq 2\sqrt{D^2 \log N}.$$

Upper bound by one-step discretization

Controlling the first term (localized complexity)
When $\mathbb{T} \subset \mathbb{R}^d$, let $\tilde{\mathbb{T}}(\delta) := \{\gamma - \gamma' \mid \gamma, \gamma' \in \mathbb{T},\ \|\gamma - \gamma'\|_2 \leq \delta\}$, a specific form of the inequality writes

$$\mathcal{G}(\mathbb{T}) \leq \min_{\delta \in [0,D]} \left\{ \mathcal{G}(\tilde{\mathbb{T}}(\delta)) + 2\sqrt{D^2 \log N(\delta; \mathbb{T}, \|\cdot\|_2)} \right\}.$$

Note that

$$\mathcal{G}(\tilde{\mathbb{T}}(\delta)) = \mathbb{E}[\sup_{\theta \in \tilde{\mathbb{T}}(\delta)} \langle \theta,\ w \rangle] \leq \delta \mathbb{E}[\|w\|_2] \leq \delta \sqrt{d},$$

we have

$$\mathcal{G}(\mathbb{T}) \leq \min_{\delta \in [0,D]} \left\{ \delta\sqrt{d} + 2\sqrt{D^2 \log N(\delta; \mathbb{T}, \|\cdot\|_2)} \right\}.$$

Example

Example 5.18 (Gaussian complexity of unit ball)
Note that $N(\delta; \mathbb{B}_2^d, \|\|_2) \leq d \log(1 + 2/\delta)$, setting $\delta = 1/2$ yields

$$\mathcal{G}(\mathbb{B}_2^d) \leq \sqrt{d}(1 + 2\sqrt{2 \log 5}).$$

Chaining and Dudley's entropy integral

The one-step discretization approximate the supremum with a finite maximum.
A more accurate approximation: sum of finite maxima over sequentially refined sets (chaining).
Let $\{X_\theta, \theta \in \mathbb{T}\}$ be a zero-mean sub-Gaussian process with respect to a metric ρ. Define the δ-truncated Dudley's entropy integral

$$\mathcal{J}(\delta; D) := \int_\delta^D \sqrt{\log N(u; \mathbb{T}, \rho)}du.$$

Chaining and Dudley's entropy integral

Theorem 5.22 (Dudley's entropy integral bound)

Let $\{X_\theta, \theta \in \mathbb{T}\}$*be a zero-mean sub-Gaussian process with respect to a metric* ρ*. Then for any* $\delta \in [0, D]$*, we have*

$$\mathbb{E}\Big[\sup_{\theta,\tilde{\theta}}(X_\theta - X_{\tilde{\theta}})\Big] \leq 2\mathbb{E}\Big[\sup_{\substack{\gamma,\gamma'\in\mathbb{T},\\ \rho(\gamma,\gamma')\leq\delta}}(X_\gamma, X'_\gamma)\Big] + 32\mathcal{J}(\delta/4; D).$$

Sketched proof

$$\sup_{\theta,\tilde{\theta}\in\mathbb{T}}(X_\theta - X_{\tilde{\theta}}) \leq 2\sup_{\substack{\gamma,\gamma'\in\mathbb{T}\\ \rho(\gamma,\gamma')\leq\delta}}(X_\gamma - X_{\gamma'}) + 2\max_{i=1,\ldots,N}|X_{\theta^i} - X_{\theta^1}|.$$

Let $\mathbb{U} = \{\theta_1, \ldots, \theta^N\}$, $\mathbb{U}_m$ be any $D2^{-m}$-covering set in the metric ρ for $m = 1, \ldots, L$. Since $\mathbb{U}$ is finite, $\exists L < \infty$ s.t. $\mathbb{U}_L = \mathbb{U}$.
For $m = 1, \ldots, L$, define $\pi_m : \mathbb{U} \to \mathbb{U}_m$ via

$$\pi_m(\theta) = \arg\min_{\beta\in\mathbb{U}_m}\rho(\theta, \beta).$$

Chaining and Dudley's entropy integral

Define $\gamma^L = \theta$ and $\gamma^{m-1} = \pi_{m-1}(\gamma^m)$.
Decompose X_θ by the chaining relation:

$$X_\theta - X_{\gamma^1} = \sum_{m=2}^{L} (X_{\gamma^m} - X_{\gamma^{m-1}}),$$

we have $|X_\theta - X_{\gamma^1}| \le \sum_{m=2}^{L} \max_{\beta \in \mathbb{U}_m} |X_\beta - X_{\pi_{m-1}(\beta)}|$.
For any other $\tilde{\theta} \in \mathbb{U}$, similarly define a chain $\{\tilde{\gamma}^1, \ldots, \tilde{\gamma}^L\}$.
Note that

$$|X_\theta - X_{\tilde{\theta}}| \le |X_{\gamma^1} - X_{\tilde{\gamma}^1}| + |X_\theta - X_{\gamma^1}| + |X_{\tilde{\theta}} - X_{\tilde{\gamma}^1}|,$$

we have

$$\max_{\theta, \tilde{\theta} \in \mathbb{U}} (X_\theta - X_{\tilde{\theta}}) \le \max_{\gamma, \gamma' \in \mathbb{U}_1} |X_\gamma - X_{\tilde{\gamma}}| + 2 \sum_{m=2}^{L} \max_{\beta \in \mathbb{U}_m} |X_\beta - X_{\pi_{m-1}(\beta)}|.$$

Chaining and Dudley's entropy integral

By the sub-Gaussian nature of $\{X_\theta,\ \theta \in \mathbb{T}\}$,

$$\mathbb{E}[\max_{\gamma,\gamma' \in \mathbb{U}_1} |X_\gamma - X_{\tilde{\gamma}}|] \leq 2D\sqrt{\log N(D/2; \mathbb{T}, \rho}).$$

Similrly,

$$\mathbb{E}[\max_{\beta \in \mathbb{U}_m} |X_\beta - X_{\pi_{m-1}(\beta)}|] \leq 2D2^{-(m-1)}\sqrt{\log N(D2^{-m}); \mathbb{T}, \rho}.$$

Thus, we have

$$\begin{aligned}
\mathbb{E}[\max_{\theta,\tilde{\theta} \in \mathbb{U}} |X_\theta - X_{\tilde{\theta}}|] &\leq 4\sum_{m=1}^{L} D2^{-(m-1)}\sqrt{\log N(D2^{-m}; \mathbb{T}, \rho)} \\
&\leq 4\sum_{m=1}^{L}\left(4\int_{D2^{-(m+1)}}^{D} 2^{-m}\sqrt{logN(u; \mathbb{T}, \rho)}du\right) \\
&= 16\int_{\delta/4}^{D}\sqrt{logN(u; \mathbb{T}, \rho)}du.
\end{aligned}$$

Substituting in the last inequality completes the proof.

Application of Dudley's inequality

Example 5.24 (Bounds for Vapnik–Chervonenkis classes)
Recall the upper bound for empirical Rademacher complexity in Lemma 4.14, we have

$$\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right] \leq 4D(x_1^n)\sqrt{\frac{v\log(n+1)}{n}},$$

where $D(x_1^n) := \sup_{f\in\mathscr{F}}\sqrt{\frac{\sum_{i=1}^{n} f(x_i)}{n}}$ is the l_2-radius of $\mathscr{F}(x_1^n)/\sqrt{n}$.
Define $Z_f = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i f(x_i)$, it can be shown that $Z_f - Z_g$ is sub-Gaussian with parameter

$$\|f-g\|_{\mathbb{P}_n}^2 := \frac{1}{n}\sum_{i=1}^{n}(f(x_i)-g(x_i))^2.$$

Application of Dudley's inequality

By Dudley's entropy integral, we have

$$\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)\right|\right] \le \frac{24}{\sqrt{n}}\int_0^{2b}\sqrt{\log N(t;\mathscr{F},\|\cdot\|_{\mathbb{P}_n})}dt.$$

Note that $N(t;\mathscr{F},\|\cdot\|_{\mathbb{P}_n}) \le Cv(16e)^v\left(\frac{b}{t}\right)^{2v}$, we have

$$\begin{aligned}\mathbb{E}\|\mathbb{P}_n-\mathbb{P}\|_{\mathscr{F}} &\le 2\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right]\\ &\le c_0\sqrt{\frac{v}{n}}\left[1+\int_0^{2b}\sqrt{\log(b/t)}dt\right] = c_0'\sqrt{\frac{v}{n}}.\end{aligned}$$

$*$ The bound is sharper than the one in Lemma 4.14.

$*$ For the set of indicator functions ($v=1$), combining this bound with Theorem 4.10 yields

$$\mathbb{P}\left(\|\hat{F}_n-F\|_{\infty} \ge \frac{c}{\sqrt{n}}+\delta\right) \le 2e^{-\frac{n\delta^2}{8}} \quad \text{for all } \delta \ge 0.$$

Table of Contents

A Gaussian comparison principle

Theorem 5.25

Let $(X_1, \ldots, X_N)$ and $(Y_1, \ldots, Y_N)$ be a pair of centered Gaussian random vectors, and suppose that there exist disjoint subsets A and B of [N] × [N] such that

$$\begin{aligned} E[X_iX_j] &\leq E[Y_iY_j] && \text{for all } (i,j) \in A, \\ E[X_iX_j] &\geq E[Y_iY_j] && \text{for all } (i,j) \in B, \\ E[X_iX_j] &= E[Y_iY_j] && \text{for all } (i,j) \notin A \cup B. \end{aligned}$$

Let $F : \mathbb{R}^N \to \mathbb{R}$ be a twice-differentiable function, and suppose that

$$\begin{aligned} \frac{\partial^2 F}{\partial u_i \partial u_j}(u) &\geq 0 && \text{for all } (i,j) \in A, \\ \frac{\partial^2 F}{\partial u_i \partial u_j}(u) &\leq 0 && \text{for all } (i,j) \in B. \end{aligned}$$

Then we are guaranteed that

$$E[F(X)] \leq E[F(Y)].$$

Proof of Theorem 5.25

Proof of the theorem:

- Define $Z(t) = \sqrt{1-t}X + \sqrt{t}Y$, for each $t \in [0, 1]$, and consider the function $\phi : [0, 1] \to \mathbb{R}$ given by $\phi(t) = E[F(Z(t))]$.
- Prove that $\phi'(t) \geq 0$:
    - First, we have $\phi'(t) = \Sigma_{j=1}^{N} E[\frac{\partial F}{\partial z_j}(Z(t))Z_t'(t)]$.
    - We write $Z_i(t) = \alpha_{ij}Z_j'(t) + W_{ij}$, where the random vector $W(j) := (W_{1j}, ..., W_{Nj})$ is independent of $Z_j'(t)$. as

$$E[Z_i(t)Z_j'(t)] = E[(\sqrt{1-t}X + \sqrt{t}Y)(-\frac{1}{2\sqrt{1-t}}X_j + \frac{1}{2\sqrt{t}}Y_j)]$$

$$= \frac{1}{2}(E[Y_iY_j] - E[X_iX_j]),$$

we have $\alpha_{ij} \geq 0$ for $(i, j) \in A$, $\alpha_{ij} \leq 0$ for $(i, j) \in B$, and $\alpha_{ij} = 0$ if $(i, j) \notin A \cup B$.

Proof of Theorem 5.25

- We apply a first-order Taylor series to the function $\partial F/\partial z_j$ between the points W(j) and Z(t):

$$\frac{\partial F}{\partial z_j}(Z(t)) = \frac{\partial F}{\partial z_j}(W(j)) + \Sigma_{i=1}^{N} \frac{\partial^2 F}{\partial z_j \partial z_i}(U)\alpha_{ij} Z_j'(t),$$

  where $U \in \mathbb{R}^N$ is some intermediate point between W(j) and Z(t).
- We have

$$\phi'(t) = E[\frac{\partial F}{\partial z_j}(W(j))Z_j'(t)] + \Sigma_{i=1}^{N} E[\frac{\partial^2 F}{\partial z_j \partial z_i}(U)\alpha_{ij}(Z_j'(t))^2]$$

$$= \Sigma_{i=1}^{N} E[\frac{\partial^2 F}{\partial z_j \partial z_i}(U)\alpha_{ij}(Z_j'(t))^2],$$

  since W(j) and $Z_j'(t)$ are independent, and $Z_j'(t)$ is zero-mean.
- For any $(i,j) \in [N] \times [N]$, we have $\frac{\partial^2 F}{\partial z_j \partial z_i}(U)\alpha_{ij} \geq 0$, then we may conclude that $\phi'(t) \geq 0$ for all $t \in (0,1)$.

- Then, we have $E[F(Y)] = \phi(1) \geq \phi(0) = E[F(X)]$.

Slepian's inequality

Corollary 5.26 (Slepian's inequality)

Let $X \in \mathbb{R}^N$ and $Y \in \mathbb{R}^N$ be zero-mean Gaussian random vectors such that

$$\begin{aligned} E[X_i X_j] &\geq E[Y_i Y_j] \qquad \textit{for all } i \neq j, \\ E[X_i^2] &= E[Y_i^2] \qquad \textit{for all } i = 1, 2, ..., N. \end{aligned}$$

Then we are guaranteed

$$E[\max_{i=1,\ldots,N} X_i] \leq E[\max_{i=1,\ldots,N} Y_i]$$

Proof of Slepian's inequality

Proof of Slepian's inequality:
We define $F_\beta(x) := \beta^{-1} log(\Sigma_{j=1}^N exp(\beta x_j))$, for each $\beta > 0$.
Then for all $\beta > 0$, we have

$$\max_{i=1,\ldots,N} X_i \le F_\beta(x) \le \max_{i=1,\ldots,N} X_i + \frac{logN}{\beta},$$

and

$$\frac{\partial^2 F}{\partial x_i \partial x_j} = -\beta \frac{exp(\beta(x_i + x_j))}{(\Sigma_{j=1}^N exp(_j))^2} \le 0.$$

Applying the above theorem, we have

$$E[\max_{i=1,\ldots,N} X_i] \le E[F_\beta(X)] \le E[F_\beta(Y)] \le E[\max_{i=1,\ldots,N} Y_i] + \frac{logN}{\beta}$$

Taking the limit $\beta \to +\infty$ yields the claim. $\square$

Sudakov–Fernique comparison

Defined the associated pseudometrics of X and Y:

$$\rho_X^2(i,j) = E(X_i - X_j)^2 \textit{ and } \rho_Y^2(i,j) = E(Y_i - Y_j)^2.$$

Theorem5.27 (Sudakov–Fernique)

Given a pair of zero-mean N-dimensional Gaussian vectors $(X_1, ..., X_N)$ *and* $(Y_1, ..., Y_N)$*, suppose that*

$$E[(X_i - X_j)^2] \le E[(Y_i - Y_j)^2] \qquad \textit{for all } (i,j) \in [N] \times [N],$$

Then

$$E[\max_{i=1,\dots,N} X_i] \le E[\max_{i=1,\dots,N} Y_i]$$

An example

Example. Suppose that X, resp., Y are centered Gaussian vectors on $\mathbb{R}^n$ with covariances C, resp., $\tilde{C}$. Show that if $\tilde{C} - C$ is positive semi-definite, then

$$E[\max_{i=1,\ldots,N} X_i] \leq E[\max_{i=1,\ldots,N} Y_i].$$

For any $(i,j) \in [N] \times [N]$, we have

$$(1,-1)(\tilde{C}\binom{i,j}{i,j} - C\binom{i,j}{i,j})(1,-1)^T \geq 0$$

Then invoking Sudakov–Fernique comparison, we get the result.

Gaussian contraction inequality

$\phi_j : \mathbb{R} \to \mathbb{R}$ is a centered 1-Lipschitz function if:

- $|\phi_j(s) - \phi_j(t)| \leq |s - t|$ for all $s, t \in \mathbb{R}$,
- $\phi_j(0) = 0$.

Proposition 5.28 (Gaussian contraction inequality)

For any set $\mathbb{T} \subseteq \mathbb{R}^d$ and any family of centered 1-Lipschitz functions $(\phi_j, j = 1, ..., d)$, we have

$$\mathcal{G}(\phi(\mathbb{T})) = E[\sup_{\theta \in \mathbb{T}} \Sigma_{j=1}^d \omega_j \phi_j(\theta_j)] \leq E[\sup_{\theta \in \mathbb{T}} \Sigma_{j=1}^d \omega_j \theta_j] = \mathcal{G}(\mathbb{T}),$$

where $\phi(\theta) := (\phi_1(\theta_1), ..., \phi_d(\theta_d)) \in \mathbb{R}^d$.

Proof of Gaussian contraction inequality

Proof of Gaussian contraction inequality:

- If $|\mathbb{T}|$ is limited.
  Define $\mathbb{T} := \{\theta_1, \theta_2, ..., \theta_N\}$.
  For $\theta_a, \theta_b \in \mathbb{T}$, we have

$$E[(\omega^T \phi(\theta_a) - \omega^T \phi(\theta_b))^2] = \Sigma_{j=1}^{d} (\phi_j(\theta_a) - \phi_j(\theta_b))^2$$

$$\leq \Sigma_{j=1}^{d} (\theta_a - \theta_b)^2 = E[(\omega^T \theta_a - \omega^T \theta_b)^2]$$

Invoking Sudakov–Fernique comparison, then we get

$$E[\max_{i=1,...,N} \Sigma_{j=1}^{d} \omega^T \phi(\theta_i)] \leq E[\max_{i=1,...,N} \Sigma_{j=1}^{d} \omega^T \theta_i].$$

Proof of Gaussian contraction inequality

- If $|\mathbb{T}|$ is countable.
  Define $\mathbb{T} := \{\theta_1, \theta_2, ...\}$.
  As above, for every $N \in \mathbb{N}_+$we can get

$$E[\max_{i=1,...,N} \Sigma_{j=1}^{d} \omega^T \phi(\theta_i)] \leq E[\max_{i=1,...,N} \Sigma_{j=1}^{d} \omega^T \theta_i]$$

  Taking the limit $N \to +\infty$ yields the claim.
- If $|\mathbb{T}|$ is uncountable.
  For$\mathbb{F} \subseteq \mathbb{R}^d$, $\mathbb{F}$ has a dense countable subset $\mathbb{G}$.
  For any $\theta \in \mathbb{F}$, we have $(\theta_i, i = 1, 2, ...) \in \mathbb{G}$ that $\lim_{i \to \infty} \theta_i = \theta$.
  Then

$$E[\omega^T \phi(\theta)] \leq E[\sup_{\theta_a \in \mathbb{G}} \omega^T \phi(\theta_a)] \leq E[\sup_{\theta_a \in \mathbb{G}} \omega^T \theta_a] \leq E[\sup_{\theta_a \in \mathbb{T}} \omega^T \theta_a].$$

$\square$

An example

Example 5.29 Given a function class $\mathbb{F}$ and a collection of design points $x_1^n \in \mathbb{R}^n$.
In various statistical problems, it is often more natural to consider the Gaussian complexity of the set

$$\mathbb{F}^2(x_1^n) := \{(f^2(x_1), f^2(x_2), \ldots, f^2(x_n)) | f \in \mathbb{F}\} \subset \mathbb{R}^n.$$

In particular, suppose that the function class is b-uniformly bounded, so that $\|f\|_\infty \leq b$ for all $f \in \mathbb{F}$. We then claim that

$$\mathcal{G}(\mathbb{F}^2(x_1^n)) \leq 2b\mathcal{G}(\mathbb{F}(x_1^n)).$$

An example

In order to establish this bound, define the function $\phi_b : \mathbb{R} \to \mathbb{R}$ via

$$\phi_b(t) = \begin{cases} t^2/(2b) & \textit{if } |t| \leq b, \\ b/2 & \textit{else} \end{cases}$$

Since $|f(x_i)| \leq b$, we have $\phi_b(f(x_i)) = f^2(x_i)/(2b)$ for all $f \in \mathbb{F}$ and i = 1, 2,..., n, and hence

$$\frac{1}{2b}\mathcal{G}(\mathbb{F}^2(x_1^n)) = E[\sup_{f \in \mathbb{F}} \Sigma_{i=1}^n \omega_i \phi_b(f(x_i))].$$

Applying Gaussian contraction inequality yields

$$E[\sup_{f \in \mathbb{F}} \Sigma_{i=1}^n \omega_i \phi_b(f(x_i))] \leq E[\sup_{f \in \mathbb{F}} \Sigma_{i=1}^n \omega_i f(x_i)] = \mathcal{G}(\mathbb{F}(x_i^n)).$$

Putting together the pieces yields the claim.

Table of Contents

Sudakov minoration

Theorem 5.30 (Sudakov minoration)

Let $\{X_\theta, \theta \in \mathbb{T}\}$ be a zero-mean Gaussian process defined on the non-empty set $\mathbb{T}$. Then

$$E[\sup_{\theta\in\mathbb{T}} X_\theta] \geq \sup_{\delta>0} \frac{\delta}{2}\sqrt{log M_X(\delta;\mathbb{T})},$$

where $M_X(\delta;\mathbb{T})$ is the -packing number of $\mathbb{T}$ in the metric $\rho_X(\theta,\tilde{\theta}) := \sqrt{E[(X_\theta X_{\tilde{\theta}})^2]}$.

Proof of Sudakov minoration

Proof of Sudakov minoration:

- For any $\delta > 0$, let $\{\theta^1, ..., \theta^M\}$ be a δ-packing of $\mathbb{T}$, and consider the sequence $\{Y_i\}_{i=1}^M$, with elements $Y_i := X_i$, we have

$$E[(Y_i - Y_j)^2] = \rho_X^2(\theta^i, \theta^j) > \delta^2 \, for \, all \, i \neq j.$$

- Define an i.i.d. sequence of Gaussian random variables $Z_i N(0, \delta^2/2)$ for i = 1,..., M, we have

$$E[(Z_i - Z_j)^2] = \delta^2 \, for \, all \, i \neq j.$$

- Invoking Sudakov–Fernique comparison, then we have

$$E[\max_{i=1,\ldots,M} Z_i] \geq E[\max_{i=1,\ldots,M} Y_i] \geq E[\sup_{\theta \in \mathbb{T}} X_\theta].$$

- We can apply standard results on i.i.d. Gaussian maxima (viz. Exercise 2.11) to obtain the lower bound $E[\max_{i=1,\ldots,M} Z_i] \geq \frac{\delta}{2}\sqrt{logM}$, thereby completing the proof. $\square$

Some examples

Example 5.31 (Gaussian complexity of l_2-ball)
We have shown previously that the Gaussian complexity $\mathcal{G}(\mathbb{B}_2^d)$ of the d-dimensional Euclidean ball is upper bounded as $\mathcal{G}(\mathbb{B}_2^d) \leq \sqrt{d}$ by Proposition 5.17.
From Example 5.9, the metric entropy of the ball $\mathbb{B}_2^d$ in l_2-norm is lower bounded as $logN_2(\delta; \mathbb{B}^d) \geq dlog(1/\delta)$. Then we have $logM_2(\delta; \mathbb{B}^d) \leq dlog(1/\delta)$.
Therefore, the Sudakov bound implies that

$$\mathcal{G}(\mathbb{B}_2^d) \geq \sup_{\delta > 0} (\frac{\delta}{2} \sqrt{dlog(1/\delta)}) \geq \frac{log4}{8} \sqrt{d},$$

where we set $\delta = 1/4$ in order to obtain the second inequality.

Some examples

Example 5.32 (Metric entropy of l_1-ball)
Let us use the Sudakov minoration to upper bound the metric entropy of the l_1-ball $\mathbb{B}_1^d = \{\theta \in \mathbb{R}^d | \Sigma_{i=1}^d |\theta_i| \leq 1\}$.
We first observe that its Gaussian complexity can be upper bounded as

$$\mathcal{G}(\mathbb{B}_1) = E[\sup_{\|\theta\|_1 \leq 1} < \omega, \theta >] = E[\|\omega\|_\infty] \leq 2logd.$$

where we standard results on Gaussian maxima (see Exercise 2.11).
Applying Sudakov's minoration, we conclude that the metric entropy of the d-dimensional ball $\mathbb{B}_1^d$ in the l_2-norm is upper bounded as

$$logN_X(\delta; \mathbb{T}, \|\cdot\|_2) \leq logM_X(\delta; \mathbb{T}, \|\cdot\|_2) \leq c(\frac{1}{\delta})^2 logd.$$

Table of Contents

ψ_q-Orlicz norm

For a given parameter $q \in [1, 2]$, consider the function $\psi_q(t) := exp(t^q) - 1$. This function can be used to define a norm on the space of random variables as follows:

Definition 5.34 (ψ_q-Orlicz norm)

The ψ_q-Orlicz norm of a zero-mean random variable X is given by

$$\|X\|_{\psi_q} := inf\{\lambda > 0 | E[\psi_q(|X|/\lambda)] \leq 1\}$$

The Orlicz norm is infinite if there is no $\lambda \in \mathbb{R}$ for which the given expectation is finite.

Remarks of ψ_q-Orlicz norm

Remarks of Orlicz norm:

- If X has a bounded Orlicz norm, it satisfies a concentration inequality specified in terms of the function ψ_q:

$$P[|X| \geq t] = P[\psi_q(|X|/\|X\|_{\psi_q}) \geq \psi_q(t/\|X\|_{\psi_q})] \leq \frac{1}{\psi_q(t/\|X\|_{\psi_q})},$$

- X has bounded ψ_1-Orlicz norm. $\iff \exists\lambda > 0 \; s.t. \; E[exp(|X|/\lambda)] \leq 2.$ $\iff$ X is sub-exponential variable.
- X has bounded ψ_2-Orlicz norm. $\iff \exists\lambda > 0 \; s.t. \; E[exp(X^2/\lambda^2)] \leq 2.$ $\iff$ X is sub-Gaussian variable.

ψ_q-process & generalized Dudley entropy integral

Definition 5.35 (ψ_q-process)

A zero-mean stochastic process $\{X_\theta, \theta \in \mathbb{T}\}$ is a ψ_q-process with respect to a metric ρ if

$$\|X_\theta - X_{\tilde{\theta}}\|_{\psi_q} \leq \rho(\theta, \tilde{\theta}) \textit{ for all } \theta, \tilde{\theta} \in \mathbb{T}.$$

Definition (generalized Dudley entropy)

The generalized Dudley entropy integral is

$$\mathcal{J}_q(\delta; D) := \int_\delta^D \psi_q^{-1}(N(u; \mathbb{T}, \rho))du,$$

where $\psi_q^{-1}(u) := [log(1+u)]^{1/q}$ is the inverse function of ψ_q, and $D = \sup_{\theta, \tilde{\theta} \in \mathbb{T}} \rho(\theta, \tilde{\theta})$ is the diameter of the set $\mathbb{T}$ under ρ.

We can find that$\mathcal{J}(\delta; D) = \mathcal{J}_2(\delta; D)$.

Orlicz norm

Definition(Young function)

A Young function is a convex function such that

$$\frac{\psi(x)}{x} \to \infty, \ as\ x \to \infty,$$

$$\frac{\psi(x)}{x} \to 0, \ as\ x \to 0.$$

Definition(Orlicz norm)

Orlicz norm of a zero-mean random variable X is

$$\|X\|_\psi := inf\{\lambda > 0 | E[\psi(|X|/\lambda)] \leq 1\},$$

where ψ is a Young function.

Orlicz space

Definition(Orlicz space)

Orlicz space L_ψ is the space of all random variables X that $\|X\|_\psi < \infty$.

The properties of Orlicz space:

- Orlicz spaces generalize L_p spaces
- The Orlicz space is a Banach space.

Some more about the random variables on Orlicz spaces are in
Ledoux, Michel; Talagrand, Michel, Probability in Banach Spaces.

Orlicz space

For examples:

- (Theorem 6.21, The upper bound of ψ_q-Orlicz norm in some cases)
  For $1 \leq q \leq 2$, there is a constant K_q, depending on q only, such that for all finite sequences (X_i) of independent mean zero random variables in L_ψ
  $$\|\Sigma X_i\|_{\psi_q} \leq K_q(\|\Sigma X_i\|_1 + (\Sigma\|X_i\|_{\psi_q}^p)^{\frac{1}{p}}),$$
  where $1/p + 1/q = 1$.
- (Theorem 11.1, Promote Theorem 5.22 in some places)
  Let $\{X_\theta, \theta \in \mathbb{T}\}$ be a random ψ-process in L_ψ, then if
  $$\int_0^D \psi^{-1}(N(u; \mathbb{T}, \rho))du < \infty.$$
  X is almost surely bounded and we actually have
  $$E[\sup_{\theta, \tilde{\theta} \in \mathbb{T}} |X_\theta - X_{\tilde{\theta}}|] \leq 8 \int_0^D \psi^{-1}(N(u; \mathbb{T}, \rho))du.$$

A concentration inequality of ψ_q-process

Theorem 5.36

Let $\{X_\theta, \theta \in \mathbb{T}\}$ be a ψ_q-process with respect to ρ. Then there is a universal constant c_1 such that

$$P[\sup_{\theta,\tilde{\theta}\in\mathbb{T}} |X_\theta - X_{\tilde{\theta}}| \geq c_1(\mathcal{J}_q(0;D) + t)] \leq 2e^{-\frac{t^q}{D^q}} \text{ for all } t > 0.$$

The theorem should be understood as generalizing Theorem 5.22 in two ways:

- It applies to general Orlicz processes for $q \in [1, 2]$, with the sub-Gaussian setting corresponding to the special case q = 2.
- It provides a tail bound on the random variable, as opposed to a bound only on its expectation.

Lemma used to prove Theorem 5.36

We define $E_A[Y] := \int_A YdP$.
Note that we have $E_A[Y] = E[Y|Y \in A]P[A]$ by construction.

Lemma 5.37

Suppose that $Y_1, ..., Y_N$ are non-negative random variables such that $\|Y_i\|\psi_q \leq 1$. Then for any measurable set A, we have

$$E_A[Y_i][A]\psi_q^{-1}(\frac{1}{P(A)}) \text{ for all } i = 1, ..., N,$$

and moreover

$$E_A[\max_{i=1,...,N} Y_i] \leq P[A]\psi_q^{-1}(\frac{N}{P(A)}).$$

Proof of lemma 5.37

Proof of lemma 5.37:

- Invoking Jensen's inequality and consider the fact that $E_A[\psi_q(Y)] \leq E[\psi_q(Y)] \leq 1$, we have

$$E_A[Y] = P[A]\frac{1}{P[A]}E_A[\psi_q^{-1}(\psi_q(Y))]$$

$$\leq P[A]\psi_q^{-1}(\frac{1}{P[A]}E_A[\psi_q(Y)]) \leq P[A]\psi_q^{-1}(\frac{1}{P[A]}).$$

- Any measurable set A can be partitioned into a disjoint union of sets A_i, i = 1, 2,..., N, such that $Y_i = \max_{j=1,\dots,N} Y_j$ on A_i.
  Invoking Jensen's inequality, we have

$$E_A[\max_{i=1,\dots,N} Y_i] = \Sigma_{i=1}^N E_{A_i}[Y_i]$$

$$\leq P[A]\Sigma_{i=1}^N \frac{P[A_i]}{P[A]}\psi_q^{-1}(\frac{1}{P[A]}) \leq P[A]\psi_q^{-1}(\frac{N}{P(A)}).$$

Lemma used to prove Theorem 5.36

Lemma

The supremum $Z := \sup_{\theta,\tilde{\theta}\in\mathbb{T}} |X_\theta - X_{\tilde{\theta}}|$ *satisfies the inequality*

$$E_A[Z] \leq 8P[A] \int_0^D \psi_q^{-1}\left(\frac{N(u;\mathbb{T},\rho)}{P[A]}\right)du.$$

Proof of the Lemma:

- Our problem was reduced to bounding the quantity $E[\sup_{\theta,\tilde{\theta}\in\mathbb{U}} |X_| thetaX_{\tilde{\theta}}|]$, where $U = \{\theta 1, ..., \theta_N\}$ was a δ-cover of the original set.

Proof of the Lemma

- For each m = 1, 2,..., L, let U_m be a minimal $D2^{-m}$-cover of U in the metric $_X$.
  We choose L that satisfy $U_L = U$.
  Then, the set U_m has $N_m = N_X(D2^{-m}; U)$ elements.
  Define the mapping $\pi_m : U \to U_m$ via $\pi_m(\theta) = argmin_\gamma \rho_X(\theta, \gamma)$.
- Using this notation, we derived the chaining upper bound

$$E_A[\max_{\theta,\tilde{\theta}\in\mathbb{U}} |X_\theta - X_{\tilde{\theta}}|] \leq 2\Sigma_{m=1}^{L} E_A[\max_{\gamma\in\mathbb{U}_m} |X_\gamma - X_{\pi_{m-1}(\gamma)}|].$$

- For each $\gamma \in \mathbb{U}_m$ we are guaranteed that

$$\|X_\gamma - X_{\pi_{m-1}(\gamma)}\|_{\psi_q} \leq \rho_X(\gamma, \pi_{m-1}(\gamma)) \leq D2^{-(m-1)}.$$

Proof of the Lemma

Lemma 5.37 implies that

$$E_A[\max_{\gamma\in\mathbb{U}_m}|X_\gamma - X_{\pi_{m-1}(\gamma)}|] \leq P[A]D2^{-(m-1)}\psi_q^{-1}(\frac{N(D2^{-m})}{P(A)}).$$

We have

$$E_A[\max_{\theta,\tilde{\theta}\in\mathbb{U}}|X_\theta - X_{\tilde{\theta}}|] \leq 2P[A]\Sigma_{m=1}^{L}D2^{-(m-1)}\psi_q^{-1}(\frac{N(D2^{-m})}{P(A)})$$

$$\leq cP[A]\int_0^D \psi_q^{-1}(\frac{N(u;\mathbb{U})}{P[A]})du$$

Taking the limit $\delta \to 0$ yields the claim. $\square$

Proof of Theorem 5.36

Proof of the theorem:
Let$Z := \sup_{\theta,\tilde{\theta}\in\mathbb{T}} |X_\theta - X_{\tilde{\theta}}|$.
We choose $A = \{Z \geq t\}$. Invoking Markov's inequality and Lemma 2, we have

$$P[Z \geq t] \leq \frac{E_A[Z]}{t} \leq 8\frac{P[Z \geq t]}{t}\int_0^D \psi_q^{-1}(\frac{N(u;\mathbb{T},\rho)}{P[Z \geq t]})du$$

Using the inequality$\psi^{-1}(st) \leq c(\psi^{-1}(s) + \psi^{-1}(t))$, we obtain

$$t \leq 8c(\int_0^D \psi_q^{-1}(N(u;\mathbb{T},\rho))du + D\psi_q^{-1}(\frac{1}{P[Z \geq t]})).$$

Proof of Theorem 5.36

Let $\delta > 0$ be arbitrary, and set $t = 8c(\mathcal{J}_q(D) + \delta)$.
Then we can get

$$\delta \leq D\psi_q^1(\frac{1}{P[Z \geq t]})$$

$$i.e.\ P[Z \geq 8c(\mathcal{J}_q(0; D) + \delta)] \leq \frac{1}{\psi_q(\delta/D)}.$$

For $\frac{1}{exp(x)-1} \leq 2exp(-x)$, we have

$$P[Z \geq 8c(\mathcal{J}_q(0; D) + \delta)] \leq 2e^{-\frac{\delta^q}{D^q}}$$

$\square$