CS 839 Probability and Learning in High Dimension Lecture 13 - 03/09/2022

Lecture 13: Random Process and Metric Entropy

Lecturer: Yudong Chen *Scribe: Puqian Wang*

In this lecture,[1] we will introduce the concept of ϵ-net, covering number, metric entropy, and random processes with sub-Gaussian increment. We will then present the Sudakov's minorization inequality, which provides a lower bound of the supremum $\mathbb{E}[\sup_{\theta\in T} Z_\theta]$ of a Gaussian random process $(Z_\theta)_{\theta\in T}$ in terms of the metric entropy of the set T.

1 ϵ-net, Covering Number and Metric Entropy

Below T is a abstract set equipped with a (pseudo-)metric ρ.

Definition 1 (ϵ-net)**.** *$T_\epsilon \subseteq T$ is called an ϵ-net of T w.r.t. ρ if:*

$$\forall u \in T, \exists u_0 \in T_\epsilon : \rho(u, u_0) \le \epsilon. \tag{1}$$

Definition 2 (Covering Number)**.** *The smallest cardinality of an ϵ-net of set T w.r.t. metric ρ is called the covering number of T, denoted by $\mathcal{N}(\epsilon; T, \rho)$.*

Definition 3 (Metric Entropy)**.** *The metric entropy of a set T is defined as the logarithm of its covering number:* $\log\left(\mathcal{N}(\epsilon; T, \rho)\right)$.

Note that the metric entropy reduces to Shannon entropy in the discrete, equiprobability case, hence the name.

We now provide an example of the covering number of ℓ_2 ball and sphere.

Example 1. Recall that $\mathbb{B}_2^d \triangleq \{u \in \mathbb{R}^d : \|u\|_2 \le 1\}$ and $\mathbb{S}_2^{d-1} \triangleq \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$ are the unit ℓ_2 ball and sphere, respectively. One can show that

$$\mathcal{N}(\epsilon; \mathbb{S}_2^{d-1}, \|\cdot\|_2) \le \mathcal{N}\left(\frac{\epsilon}{2}; \mathbb{B}_2^d, \|\cdot\|\right) \le \left(\frac{2}{\epsilon} + 1\right)^d, \tag{2}$$

$$\mathcal{N}\left(\frac{\epsilon}{2}; \mathbb{B}_2^d, \|\cdot\|\right) \ge \left(\frac{2}{\epsilon}\right)^d. \tag{3}$$

where the first inequality comes from the monotonicity of covering number, and the next two inequalities can be derived using volume calculation (see Example 12). It is worth noticing that the bounds are quite tight, as the upper and lower bounds nearly match.

The covering number of a set T is the minimum number of ϵ balls that covers T. Symmetrically, we can define the packing number of T, which is the maximum number of ϵ balls that can be contained in set T.

Definition 4 (ϵ-packing)**.** *$T_\epsilon \subset T$ is an ϵ-packing of T w.r.t. metric ρ if:*

$$\rho(\theta, \theta') > \epsilon, \ \forall \theta, \theta' \in T_\epsilon.$$

[1] *Reading:*

1. Section 4.2, 7.1, 7.4 in *High -Dimensional Probability: An Introduction with Applications in Data Science*, Roman Vershynin, Cambridge University Press, 2018.
2. Section 5.1, 5.2, 5.5 in *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Martin J. Wainwright, Cambridge University Press, 2019.

Definition 5 (Packing Number)**.** *The largest cardinality of ϵ-packing of set T is called the packing number of T, denoted by $\mathcal{M}(\epsilon; T, \rho)$.*

Actually, the packing number of a set T and the covering number are almost equivalent, in the sense that one can be bounded by the other as shown in the following lemma. Therefore, we do not differentiate these to notions in the rest of the lecture.

Lemma 1. *For all $\epsilon > 0$, it holds that*

$$\mathcal{M}(2\epsilon; T, \rho) \leq \mathcal{N}(\epsilon; T, \rho) \leq \mathcal{M}(\epsilon; T, \rho). \tag{4}$$

We leave the proof of Eq. (4) as an exercise. *Hint: See Vershynin Ex 7.4.2*

2 Random Process

As introduced in the previous lecture, the excess risk of the empirical risk minimizer $\hat{f}_n$, defined as $L(\hat{f}_n) - L(f_0)$, can be bounded by the supremum of some random process X_f indexed by functions f from a function class $\mathcal{F}$. Hence the task is to bound $\sup X_f$ properly. Now we consider a more general setting involving a random process $(Z_\theta)_{\theta \in T}$ indexed by a general set T.

Definition 6 (Random Process)**.** *A random process is a collection of random variables $(Z_\theta)_{\theta \in T}$, defined on the same probability space, that are indexed by elements θ of some set T.*

Here, we exhibit first some important random processes.
Example 2 (Processes indexed by integers)**.** Consider discrete time $T = \{1, 2, \cdots, n\}$, then $(Z_1, Z_2, \cdots, Z_n)$ is a random vector in $\mathbb{R}^n$.

Example 3 (Processes indexed by vectors)**.** Consider $T \in \mathbb{R}^d$.

1. Rademacher Process: $Z_\theta = \langle \epsilon, \theta \rangle = \sum_{i=1}^d \epsilon_i \theta_i$, where $\epsilon \overset{iid}{\sim} unif\{\pm 1\}$.
2. Gaussian Process: $\forall$finite $T_0 \subset T, (Z_\theta)_{\theta \in T_0}$ is jointly Gaussian.
3. Canonical Gaussian Process: $Z_\theta = \langle g, \theta \rangle = \sum_{i=1}^d g_i \theta_i$, where $g_i \overset{iid}{\sim} N(0, 1)$.

Example 4 (Processes indexed by functions)**.** The index set T can also be a function class. Consider $T = \mathcal{F}$ being the class of function mapping from $\mathcal{X} \to \mathbb{R}$. Let $\{X_i\}$ be independent and identically distributed random variables. For each $f \in \mathcal{F}$, define

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

The process $(Z_f)_{f \in \mathcal{F}}$ is called an **empirical process**, as it is defined via an empirical average.

In general, we are interested in finding the upper or lower bounds on these processes. We now introduce the concept of sub-Gaussian increment, which can be viewed as a generalization of Gaussian process.

Definition 7 (Sub-Gaussian Increments)**.** *$(Z_\theta)_{\theta \in T}$ has sub-Gaussian increment with respect to metric ρ on T if:*

$$\mathbb{E}\left[\exp(\lambda(Z_\theta - Z_{\theta'})\right] \leq \exp\left(\frac{1}{2}\lambda^2 \rho(\theta, \theta')^2\right), \forall \theta, \theta' \in T, \lambda \in \mathbb{R}. \tag{5}$$

i.e., $Z_\theta - Z_{\theta'}$ is sub-Gaussian with parameter $\rho(\theta, \theta')^2$.

We provide some examples.
Example 5 (Processes with sub-Gaussian Increments)**.**

1. Rademacher Process: By Hoeffding we know that $Z_\theta - Z_{\theta'}$ is $\|\theta - \theta'\|_2^2$ sub-Gaussian,

$$\implies (Z_\theta) \text{ has sub-Gaussian increments w.r.t. } \rho(\theta, \theta') = \|\theta - \theta'\|_2 .$$

2. Zero-mean Gaussian Process: We have $Z_\theta - Z_{\theta'} \sim N\left(0, \mathbb{E}\left[(Z_\theta - Z_{\theta'})^2\right]\right)$

$$\implies (Z_\theta) \text{ has sub-Gaussian increments w.r.t. } \rho(\theta, \theta') \triangleq \sqrt{\mathbb{E}\left[(Z_\theta - Z_{\theta'})^2\right]}.$$

3. Canonical Gaussian Process: We have $Z_\theta - Z_{\theta'} = \langle g, \theta - \theta' \rangle \sim N\left(0, \|\theta - \theta'\|_2^2\right)$

$$\implies (Z_\theta) \text{ has sub-Gaussian increments w.r.t. } \rho(\theta, \theta') = \|\theta - \theta'\|_2 .$$

In subsequent lectures, we will develop powerful techniques for proving upper bounds on processes with sub-Gaussian increments. Today we will focus on Gaussian process and lower bounds.

3 Sudakov's Lower Bound

Now, we turn to the lower bound on the supreme of the random process:

$$\mathbb{E} \sup_{\theta \in T} Z_\theta.$$

We introduce the Sudakov's minorization inequality, which provides a lower bound for the quantity above for Gaussian processes.

Theorem 6 (Sudakov's Minorization Inequality)**.** *Let $(Z_\theta)_{\theta \in T}$ be a zero-mean Gaussian process. Then,*

$$\mathbb{E}\left[\sup_{\theta \in T} Z_\theta\right] \geq \frac{\epsilon}{2} \sqrt{\log \mathcal{N}(\epsilon; T, \rho)}, \ \forall \epsilon \geq 0, \tag{6}$$

where the metric is $\rho(\theta, \theta') = \sqrt{\mathbb{E}\left[(Z_\theta - Z_{\theta'})^2\right]}$.

To prove this theorem, we need Lemma 1 and the following Lemma 2.

Lemma 2 (Finite Gaussian Maxima, Lower Bound)**.** *Let $X_i \overset{iid}{\sim} N(0, \sigma^2)$. Then,*

$$\mathbb{E}\left[\max_{i=1,2,\cdots,N} X_i\right] \gtrsim \sigma \sqrt{\log N}. \tag{7}$$

The proof of this lemma can be find at: http://www.gautamkamath.com/writings/gaussian_max.pdf.

Now we are ready for the proof of the main theorem.

Proof [Sudakov's minorization inequality] Let T_ϵ be an maximal ϵ-packing of T. Then by Eq. (4) in Lemma 1, we have:

$$|T_\epsilon| = \mathcal{M}(\epsilon; , T, \rho) \geq \mathcal{N}(\epsilon; T, \rho).$$

We also have

$$\mathbb{E} \sup_{\theta \in T} Z_\theta \geq \mathbb{E} \sup_{\theta \in T_\epsilon} Z_\theta.$$

We now compare $(Z_\theta)_{\theta \in T_\epsilon}$ with another process $(Y_\theta)_{\theta \in T_\epsilon}$, where $Y_\theta \overset{iid}{\sim} N(0, \frac{\epsilon^2}{2}), \theta \in T_\epsilon$. We can check that for all $\theta, \theta' \in T_\epsilon$, it holds that

$$\begin{aligned}
\mathbb{E}\left[(Z_\theta - Z_{\theta'})^2\right] &= \rho(\theta, \theta')^2 && \text{definition of } \rho \\
&> \epsilon^2 && T_\epsilon \text{ is } \epsilon\text{-packing} \\
&= \mathbb{E}\left[(Y_\theta - Y_{\theta'})^2\right] && \text{by the definition of } Y_\theta.
\end{aligned}$$

Then, by Sudakov-Fernique from the last lecture, we obtain

$$\begin{aligned}
\mathbb{E}\sup_{\theta\in T_\epsilon} Z_\theta &\geq \mathbb{E}\sup_{\theta\in T_\epsilon} Y_\theta \\
&\gtrsim \frac{\epsilon}{\sqrt{2}}\sqrt{\log|T_\epsilon|} && \text{by Lemma 2} \\
&\gtrsim \epsilon\sqrt{\log\mathcal{N}(\epsilon;T,\rho)}
\end{aligned}$$

Note that the above argument is valid for any $\epsilon > 0$. □

4 Applications of Sudakov's Inequality

We discuss several applications of Sudakov's Inequality in Theorem 6.

4.1 Lower Bounding the Supremum

In the first two example, we use Sudakov's inequality to lower bound the supremum of some random processes.

Example 7 (Gaussian complexity of unit ℓ_2 ball $\mathbb{B}^d$.)**.** It is easy to obtain the upper bound as:

$$\mathbb{E}\sup_{\theta\in\mathbb{B}^d}\langle\theta,g\rangle \leq \mathbb{E}\|g\|_2 \leq \sqrt{\mathbb{E}\|g\|_2^2} \leq \sqrt{d}.$$

As for the lower bound, apply the Sudakov's inequaliy: $\forall\epsilon > 0$

$$\begin{aligned}
\mathbb{E}\sup_{\theta\in\mathbb{B}^d}\langle\theta,g\rangle &\gtrsim \epsilon\sqrt{\log\mathcal{N}(\epsilon;\mathbb{B}^d,\|\cdot\|_2)} \\
&\geq \epsilon\sqrt{\log\left(\frac{1}{\epsilon}\right)^d} && \text{from Example 1} \\
&\gtrsim \sqrt{d}. && \text{take } \epsilon = \frac{1}{e}
\end{aligned}$$

Note that the upper and lower bounds match up to a constant.

Example 8 (Max singular value of Gaussian matrix)**.** Let $X \in \mathbb{R}^{n\times n}$, where $X_{ij} \overset{iid}{\sim} N(0,1)$. Then,

$$\begin{aligned}
\mathbb{E}\|X\|_{op} &= \mathbb{E}\sup_{u,v\in\mathbb{S}^{n-1}}\langle X, uv^T\rangle \\
&\gtrsim \epsilon\sqrt{\log\mathcal{N}(\epsilon;\mathbb{S}^{n-1}\times\mathbb{S}^{n-1},\|uv^T-\tilde{u}\tilde{v}^T\|_F)} \quad \forall\epsilon>0 \\
&\gtrsim \sqrt{n}. && \text{(exercise)}
\end{aligned}$$

The above matches (up to a constant) the upper bound $\mathbb{E}\|X\|_{op} \leq 2\sqrt{n}$ from last lecture.

4.2 Upper Bounding the Metric Entropy

Sudakov's inequality can also be also used inversely to upper bound the covering number and the metric entropy.

To proceed, we first state an upper bound on Gaussian maxima, complementing the lower bound in Lemma 2. We note that g_i's need not to be independent below, unlike in the lower bound.

Lemma 3 (Finite Gaussian Maxima, Upper Bound). *If $g_i \sim N(0,1), i = 1, 2, \cdots, d$, then:*

$$\mathbb{E} \max_{i=1,2,\cdots,d} |g_i| \lesssim \sqrt{\log d}. \tag{8}$$

Proof For each fixed $\beta > 0$, we have

$$\begin{aligned}
\mathbb{E} \max_{i=1,2,\cdots,d} |g_i| &= \frac{1}{\beta} \mathbb{E} \log e^{\beta \max |g_i|} && |g_i| = \max\{g_i, -g_i\} \\
&\le \frac{1}{\beta} \mathbb{E} \log \left(\sum_i e^{\beta g_i} + \sum_i e^{-\beta g_i} \right) && \max \le \sum \\
&\le \frac{1}{\beta} \log \mathbb{E} \left(\sum_i e^{\beta g_i} + \sum_i e^{-\beta g_i} \right) && \text{Jensen's} \\
&= \frac{1}{\beta} \log \left(2d \, \mathbb{E} \, e^{\beta g_1} \right) \\
&= \frac{1}{\beta} \log \left(2d e^{\beta^2/2} \right) \\
&\lesssim \sqrt{\log d} && \text{take } \beta = \sqrt{\log d}.
\end{aligned}$$

□

We now present two examples for bounding the covering number and metric entropy.

Example 9 (Covering number of ℓ_1 ball in ℓ_2 norm). Let $\mathbb{B}_1^d := \{\theta \in \mathbb{R}^d : \|\theta\|_1 \le 1\}$. For all $\epsilon > 0$, by Sudakov we have

$$\begin{aligned}
&\epsilon \sqrt{\log \mathcal{N}(\epsilon; \mathbb{B}_1^d, \|\cdot\|_2)} \\
&\lesssim \mathbb{E} \sup_{\theta \in \mathbb{B}_1^d} \langle \theta, g \rangle && g_i \overset{iid}{\sim} N(0,1), i = 1, 2, \cdots, d. \\
&\lesssim \mathbb{E} \|g\|_\infty .
\end{aligned}$$

Using Lemma 3 to bound $\mathbb{E} \|g\|_\infty$, we obtain that

$$\log \mathcal{N}(\epsilon; \mathbb{B}_1^d, \|\cdot\|_2) \lesssim \frac{1}{\epsilon^2} \log d.$$

Note that the last RHS is logarithmic in d. Compare with the metric entropy of the ℓ_2 ball given in Example 1:

$$\log \mathcal{N}(\epsilon; \mathbb{B}_2^d, \|\cdot\|_2) \lesssim d \log(1 + \frac{4}{\epsilon}),$$

which is linear in d. We observe that ℓ_1 ball is much smaller than ℓ_2 ball when the dimension d is large. This fact is one of the reasons why ℓ_1 regularization methods, like Lasso, works in high dimension scenarios.

Example 10 (Covering number of polytopes). Suppose $P \subset \mathbb{R}^d$ a polytope with m vertices and radius < 1, i.e.,

$$\max_{\theta \in P} \|\theta\|_2 \le 1.$$

Say $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(m)}$ are the vertices. Then, for all $\epsilon > 0$, by Sudakov we have

$$\begin{aligned}
&\epsilon\sqrt{\log \mathcal{N}(\epsilon; P, \|\cdot\|_2)} \\
&\lesssim \mathbb{E} \sup_{\theta \in P} \langle \theta, g \rangle \\
&= \mathbb{E} \sup_{i=1,2,\cdots,m} \left\langle g, \theta^{(i)} \right\rangle && \text{note that linear functions are maximized at vertices} \\
&= \mathbb{E} \max_{i=1,2,\cdots,m} w_i, && \text{where } w_i = \left\langle g, \theta^{(i)} \right\rangle \sim N(0, \left\|\theta^{(i)}\right\|_2) \\
&\lesssim \sqrt{\log m} && \text{by Lemma 3.}
\end{aligned}$$

It follows that

$$\log \mathcal{N}(\epsilon; P, \|\cdot\|_2) \leq \frac{1}{\epsilon^2} \log m.$$

Note that the right hand side of the equation above is independent of dimension d. This dimension dependence is much better than the naive bound

$$\log \mathcal{N}(\epsilon; P, \|\cdot\|_2) \leq \log \mathcal{N}(\epsilon; \mathbb{B}_2^d, \|\cdot\|_2) \leq d \log(1 + \frac{4}{\epsilon}).$$

5 Volume-based Formula for Covering Number

In this section, we present a general, volume-based approach for estimating the covering number of sets in $\mathbb{R}^d$ with respect to some norm.

Theorem 11 (Volume and Covering Number)**.** *Let $T \subset \mathbb{R}^d$ and $\|\cdot\|$ be any norm. Let $\mathrm{vol}(\cdot)$ denote the volume (Lesbegue measure), $+$ denote the Minkowski sum, $\mathbb{B}(\epsilon) := \{\theta \in \mathbb{R}^d : \|\theta\| \leq \epsilon\}$ be the ball of radius ϵ w.r.t. the norm $\|\cdot\|$, and finally, $\mathbb{B} := \mathbb{B}(1)$ be the unit ball. We have*

$$\left(\frac{1}{\epsilon}\right)^d \frac{\mathrm{vol}(T)}{\mathrm{vol}(\mathbb{B})} \leq \mathcal{N}(\epsilon; T, \|\cdot\|) \leq \mathcal{M}(\epsilon; T, \|\cdot\|) \leq \frac{\mathrm{vol}(T + \frac{\epsilon}{2}\mathbb{B})}{\mathrm{vol}(\frac{\epsilon}{2}\mathbb{B})} \overset{(a)}{\leq} \frac{\mathrm{vol}(\frac{3}{2}T)}{\mathrm{vol}(\frac{\epsilon}{2}\mathbb{B})} = \left(\frac{3}{\epsilon}\right)^d \frac{\mathrm{vol}(T)}{\mathrm{vol}(\mathbb{B})}, \tag{9}$$

where equality (a) holds when T is convex and $\epsilon\mathbb{B} \subset T$.

The proof of this theorem can be found at http://www.stat.yale.edu/~yw562/teaching/598/lec14.pdf (Theorem 14.2).

Example 12 (Covering ℓ_∞ and ℓ_2 balls in $\|\cdot\|_2$ norm)**.** For $p \in [1, \infty]$, let $\mathbb{B}_p^d := \{\theta \in \mathbb{R}^d : \|\theta\|_p \leq 1\}$ denote the unit ℓ_p norm ball in $\mathbb{R}^d$.

First consider covering the d-dimensional ℓ_2 unit ball $\mathbb{B}_2^d$. Taking $T = \mathbb{B} = \mathbb{B}_2^d$ in Eq. (9), we would get:

$$\left(\frac{1}{\epsilon}\right)^d \leq \mathcal{N}(\epsilon; \mathbb{B}_2^d, \|\cdot\|_2) \leq \frac{\mathrm{vol}\left(\left(1 + \frac{\epsilon}{2}\right)\mathbb{B}_2^d\right)}{\mathrm{vol}(\frac{\epsilon}{2}\mathbb{B}_2^d)} = \left(\frac{2}{\epsilon} + 1\right)^d.$$

Thus, the lower bound and the upper bound of the ℓ_2 ball metric entropy are:

$$d \log\left(\frac{1}{\epsilon}\right) \leq \log \mathcal{N}(\epsilon; \mathbb{B}_2, \|\cdot\|_2) \leq d \log\left(\frac{2}{\epsilon} + 1\right). \tag{10}$$

Note that in d-dimensional Euclidean space, the volume of a unit ℓ_2 ball is $\mathrm{vol}(\mathbb{B}_2^d) = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$, where Γ denotes the gamma function, whereas the volume of a unit ℓ_∞ ball is $\mathrm{vol}(\mathbb{B}_\infty^d) = 2^d$. Hence, if we take $T = \mathbb{B}_\infty^d$ and $\mathbb{B} = \mathbb{B}_2^d$ in Eq. (9), we obtain

$$\mathcal{N}(\epsilon; \mathbb{B}_\infty^d, \|\cdot\|_2) \geq \left(\frac{1}{\epsilon}\right)^d \frac{2^d}{\pi^{d/2}} \Gamma\left(\frac{d}{2}+1\right) \overset{\text{Stirling's}}{\simeq} \left(\frac{1}{\epsilon}\right)^d \sqrt{\pi d} \left(\frac{2d}{\pi e}\right)^{\frac{d}{2}} \geq \left(\frac{1}{\epsilon}\right)^d \left(\frac{d}{4}\right)^{\frac{d}{2}},$$

and

$$\mathcal{N}(\epsilon; \mathbb{B}_\infty^d, \|\cdot\|_2) \leq \left(\frac{3}{\epsilon}\right)^d \frac{2^d}{\pi^{d/2}} \Gamma\left(\frac{d}{2}+1\right) \overset{\text{Stirling's}}{\simeq} \left(\frac{3}{\epsilon}\right)^d \sqrt{\pi d} \left(\frac{2d}{\pi e}\right)^{\frac{d}{2}} \leq \left(\frac{3}{\epsilon}\right)^d d^{\frac{d}{2}},$$

where we use the Stirling's formula to approximate the gamma function. Combining, we get:

$$d \log\left(\frac{1}{2\epsilon}\sqrt{d}\right) \leq \log \mathcal{N}(\epsilon; \mathbb{B}_\infty^d, \|\cdot\|_2) \leq d \log\left(\frac{3}{\epsilon}\sqrt{d}\right) \tag{11}$$

Comparing Eq. (10) and Eq. (11), we see that the metric entropy of ℓ_∞ ball is larger than that of ℓ_2 ball by a logarithmic factor in the dimension d. (The covering number is larger by a $(\sqrt{d})^d$ factor.)