High Dimensional Probability and Statistics **2nd Semester, 2021-2022**

Lecture 4: Expectation of Suprema: Finite Approximation

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2022/04/10)*

Recap and Motivation: As already mentioned previously, estimating the suprema of the form

$$\sup_{t\in T} X_t \tag{4.1}$$

arises in a wide range of contexts. Two representative examples are:

- The generalization error analysis in empirical risk minimization finally reduces to

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{k=1}^{n} f(X_k) - \mathbb{E}\left[f(X)\right] \right|,$$

  where $\mathcal{F}$ is a set of functions. This is typically referred to as *Uniform Law of Large Numbers*.

- The spectral norm of a random matrix $W \in \mathbb{R}^{m\times n}$ can be expressed as

$$\|W\|_2 = \sup_{\|u\|_2=1, \|v\|_2=1} u^\top W v.$$

While the concentration of (4.1) around its mean for typical applications can be established by the concentration inequalities discussed in the last three lectures, computing the expectation

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \tag{4.2}$$

is by no means easy. This will be the focus in the next few lectures. We will first study the general form (4.2), and then give a particular treatment of this problem for uniform law of large numbers. When there is no conflict from the context, it is always assumed that $\mathbb{E}\left[X_t\right] = 0$.

Agenda:

- Finite maxima
- Gaussian complexity and Rademacher complexity
- Covering and packing
- Finite approximation bound

4.1 Finite Maxima

The problem here is to bound

$$\mathbb{E}\left[\max_{k=1,\cdots,n} X_k\right],$$

where X_k is σ^2-sub-Gaussian. Maybe the most naive approach is to bound the supremum by a sum,

$$\mathbb{E}\left[\max_{k=1,\cdots,n} X_k\right] \leq \mathbb{E}\left[\sum_{k=1}^{n} |X_k|\right] \leq n \max_{k=1,\cdots,n} \mathbb{E}\left[|X_k|\right] \lesssim n\sigma,$$

where the last inequality follows from the σ^2-sub-Gaussian property of each X_k. Of course, bounding a maximum by a sum is an exceedingly crude idea. We may consider a transform of $\max_{k=1,\cdots,n} X_k$ such that the gap between supreme and sum is seemingly not so large after the transform. Next we attempt to provide a bound based on the higher order moments,

$$\begin{aligned}
\mathbb{E}\left[\max_{k=1,\cdots,n} X_k\right] &\leq \left(\left(\mathbb{E}\left[\max_{k=1,\cdots,n} |X_k|\right]\right)^p\right)^{1/p} \\
&\leq \left(\mathbb{E}\left[\max_{k=1,\cdots,n} |X_k|^p\right]\right)^{1/p} \\
&\leq \left(n \max_{k=1,\cdots,n} \mathbb{E}\left[|X_k|^p\right]\right)^{1/p} \\
&\lesssim n^{1/p}\sigma\sqrt{p}.
\end{aligned}$$

Minimizing the righthand side with respect to p yields that

$$\mathbb{E}\left[\max_{k=1,\cdots,n} X_k\right] \lesssim \sigma\sqrt{\log n}.$$

Of course we can apply the moment generating function to estimate the maximum as in the development of the tail bound by the Chernoff method. More precisely, we have the following lemma.

Lemma 4.1 *Let $\{X_k\}_{k=1}^n$ be σ^2-sub-Gaussian random variables. Then*

$$\mathbb{E}\left[\max_{k=1,\cdots,n} X_k\right] \leq \sigma\sqrt{2\log n}.$$

Proof: We have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda \max_k X_k\right)\right] = \mathbb{E}\left[\max_k \exp\left(\lambda X_k\right)\right] &\leq \sum_k \mathbb{E}\left[\exp\left(\lambda X_k\right)\right] \\
&\leq n\exp(\sigma^2\lambda^2/2) = \exp(\log(n) + \sigma^2\lambda^2/2).
\end{aligned}$$

Thus, the application of Jensen's inequality yields that

$$\exp\left(\mathbb{E}\left[\lambda \max_k X_k\right]\right) \leq \mathbb{E}\left[\exp\left(\lambda \max_k X_k\right)\right] \leq \exp(\log(n) + \sigma^2\lambda^2/2),$$

which leads to

$$\mathbb{E}\left[\max_k X_k\right] \leq \frac{\log(n)}{\lambda} + \frac{\sigma^2\lambda}{2}.$$

Taking $\lambda = \sqrt{2\log(n)}/\sigma$ concludes the proof. ∎

It is evident that we cannot obtain a general low bound, for example, letting $X_1 = \cdots = X_n$. Nevertheless, the upper bound in the last lemma is indeed tight for independent Gaussian random variables. More details on lower bound will be provided in the sequel for suprema of random Gaussian processes.

Lemma 4.2 *Let $\{X_k\}_{k=1}^n$ be i.i.d $\mathcal{N}(0,\sigma^2)$ random variables. Then there exists a small absolute constant $c > 0$ such that*

$$\mathbb{E}\left[\max_{k=1,\cdots,n} X_k\right] \geq c \cdot \sigma\sqrt{\log n}.$$

Proof: First we have

$$\begin{aligned}
\mathbb{E}\left[\max_{k=1,\cdots,n} X_k\right] &= \mathbb{E}\left[\max\left\{\max_{k=1,\cdots,n} X_k, 0\right\}\right] + \mathbb{E}\left[\min\left\{\max_{k=1,\cdots,n} X_k, 0\right\}\right] \\
&\geq \mathbb{E}\left[\max\left\{\max_{k=1,\cdots,n} X_k, 0\right\}\right] + \mathbb{E}\left[\min\left\{X_1, 0\right\}\right] \\
&= \int_0^\infty \mathbb{P}\left[\max_{k=1,\cdots,n} X_k > t\right] dt + \mathbb{E}\left[\min\left\{X_1, 0\right\}\right] \\
&\geq \delta \cdot \mathbb{P}\left[\max_{k=1,\cdots,n} X_k > \delta\right] - \mathbb{E}\left[|X_1|\right] \\
&= \delta\left(1 - \left(\mathbb{P}\left[X_1 \leq \delta\right]\right)^n\right) - \mathbb{E}\left[|X_1|\right] \\
&= \delta\left(1 - \left(1 - \mathbb{P}\left[X_1 > \delta\right]\right)^n\right) - \mathbb{E}\left[|X_1|\right].
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{P}\left[X_1 > \delta\right] &= \frac{1}{\sqrt{2\pi}\sigma}\int_\delta^\infty e^{-\frac{x^2}{2\sigma^2}} dx \\
&= \frac{1}{\sqrt{2\pi}\sigma}\int_\delta^\infty e^{-\frac{(y+\delta)^2}{2\sigma^2}} dy \\
&\geq \frac{e^{-\delta^2/\sigma^2}}{c_1}
\end{aligned}$$

for some numerical constant $c_1 > 0$. Choosing $\delta = \sigma\sqrt{\log(n/c_1)}$ yields that $\mathbb{P}\left[X_1 > \delta\right] \geq 1/n$. Consequently,

$$\begin{aligned}
\mathbb{E}\left[\max_{k=1,\cdots,n} X_k\right] &\geq \sigma\sqrt{\log(n/c_1)}(1 - (1-1/n)^n) - \sigma \\
&\geq (1-1/e)\sqrt{\log(n/c_1)}\sigma - \sigma,
\end{aligned}$$

which concludes the proof for sufficiently large n. ∎

4.2 Rademacher Complexity and Gaussian Complexity

This section studies $\mathbb{E}\left[\sup_{t\in T} X_t\right]$ associated with Rademacher complexity and Gaussian complexity. Recall that give a set $T \subset \mathbb{R}^d$, the Rademacher complexity is defined as

$$\mathcal{R}(T) = \mathbb{E}\left[\sup_{t\in T} \langle \varepsilon, t\rangle\right], \quad \varepsilon = (\varepsilon_1, \cdots, \varepsilon_d).$$

while the Gaussian complexity of T is defined as

$$\mathcal{G}(T) = \mathbb{E}\left[\sup_{t\in T} \langle g, t\rangle\right], \quad g \sim \mathcal{N}(0, I_d),$$

Lemma 4.3 *We have*

$$\mathcal{R}(T) \lesssim \mathcal{G}(T) \lesssim \mathcal{R}(T)\sqrt{\log d}.$$

Proof: Lower bound. First we have

$$\begin{aligned}
\mathcal{R}(T) &= \mathbb{E}_\varepsilon\left[\sup_{t\in T}\sum_{k=1}^d \varepsilon_k t_k\right] = \sqrt{\frac{\pi}{2}}\mathbb{E}_\varepsilon\left[\sup_{t\in T}\mathbb{E}_g\left[\sum_{k=1}^d \varepsilon_k |g_k| t_k\right]\right] \\
&\le \sqrt{\frac{\pi}{2}}\mathbb{E}_\varepsilon\left[\mathbb{E}_g\left[\sup_{t\in T}\sum_{k=1}^d \varepsilon_k |g_k| t_k\right]\right] \\
&= \sqrt{\frac{\pi}{2}}\mathbb{E}\left[\sup_{t\in T}\sum_{k=1}^n g_k t_k\right] \\
&= \sqrt{\frac{\pi}{2}}\mathcal{G}(T),
\end{aligned}$$

where the third follows from the fact that $\varepsilon_k|g_k|$ has the same distribution with g_k.

Upper bound. To prove the upper bound, first note that the function

$$f(a_1, \cdots, a_d) := \mathbb{E}\left[\sup_{t\in T}\sum_{k=1}^d \varepsilon_k t_k a_k\right]$$

is a convex function. Thus, the maximum of f over the region $\{(a_1, \cdots, a_d) : \ |a_k| \le 1, k = 1, \cdots, d\}$ must be achieved at the boundary. Then it follows that

$$\mathbb{E}\left[\sup_{t\in T}\sum_{k=1}^d \varepsilon_k t_k a_k\right] \le \mathbb{E}\left[\sup_{t\in T}\sum_{k=1}^d \varepsilon_k t_k\right] \quad \text{for all } |a_k| \le 1, k = 1, \cdots, d.$$

Consequently,

$$\begin{aligned}
\mathcal{G}(T) &= \mathbb{E}_g\left[\mathbb{E}_\varepsilon\left[\sup_{t\in T}\sum_{k=1}^d \varepsilon_k |g_k| t_k\right]\right] \\
&= \mathbb{E}_g\left[\max_k |g_k| \cdot \mathbb{E}_\varepsilon\left[\sup_{t\in T}\sum_{k=1}^d \varepsilon_k \frac{|g_k|}{\max_k |g_k|} t_k\right]\right]
\end{aligned}$$

$$\leq \mathbb{E}_g \left[\max_k |g_k| \cdot \mathcal{R}(T) \right]$$
$$\asymp \mathcal{R}(T)\sqrt{\log d},$$

as claimed. ■

Example 4.4 (Unit 2-norm ball $\mathbb{B}_2^d$) *Let $\mathbb{B}_2^d = \{t \in \mathbb{R}^d : \ \|t\|_2 \leq 1\}$. Then,*

$$\mathcal{R}(T) = \mathbb{E}\left[\sup_{\|t\|_2 \leq 1} \langle \varepsilon, t \rangle \right] = \mathbb{E}\left[\|\varepsilon\|_2 \right] = \sqrt{d}.$$

The same argument shows that

$$\mathcal{G}(T) = \mathbb{E}\left[\|g\|_2 \right] \leq \sqrt{\mathbb{E}\left[\|g\|_2^2 \right]} = \sqrt{d}.$$

Together with Lemma 4.3, we can conclude that $\mathcal{G}(\mathbb{B}_2^d) \asymp \sqrt{d}$. Therefore, the Rademacher complexity and the Gaussian complexity of $\mathbb{B}_2^d$ are essentially the same.

Example 4.5 (Unit 1-norm ball $\mathbb{B}_1^d$) *Let $\mathbb{B}_1^d = \{t \in \mathbb{R}^d : \ \|t\|_1 \leq 1\}$. Then,*

$$\mathcal{R}(T) = \mathbb{E}\left[\sup_{\|t\|_1 \leq 1} \langle \varepsilon, t \rangle \right] = \mathbb{E}\left[\|\varepsilon\|_\infty \right] = 1,$$

while

$$\mathcal{G}(T) = \mathbb{E}\left[\sup_{\|t\|_1 \leq 1} \langle g, t \rangle \right] = \mathbb{E}\left[\|g\|_\infty \right] \asymp \sqrt{\log d}.$$

Therefore, in this case, the Rademacher complexity and Gaussian complexity differ by the order $\sqrt{\log d}$. By Lemma 4.3, this difference turns out to be the worst possible.

4.3 Covering and Packing

For general random variables X_t and infinite number of elements in T, the first step can be made by approximating the supremum with a maximum of finite number of random variables. It should be not surprising that overall the bound for (4.2) should rely on the complexity or richness of the index set T. This section studies two closely related ways to measure the complexity of T, which indeed shows how to approximate T with a set of finite number of elements to achieve certain accuracy.

Definition 4.6 (ε-net and covering number) *Let (T, d) be a metric space. A set $N \subset T$ is called an ε-net of (T, d) if for every $t \in T$, there exists a $\pi(t) \in N$ such that $d(t, \pi(t)) \leq \varepsilon$. The covering number of (T, d), denoted $N(T, d, \varepsilon)$, is the smallest possible cardinality of an ε-net of (T, d). That is,*

$$N(T, d, \varepsilon) = \inf\{|N| : \ N \text{ is a } \varepsilon\text{-net of } (T, d)\}.$$

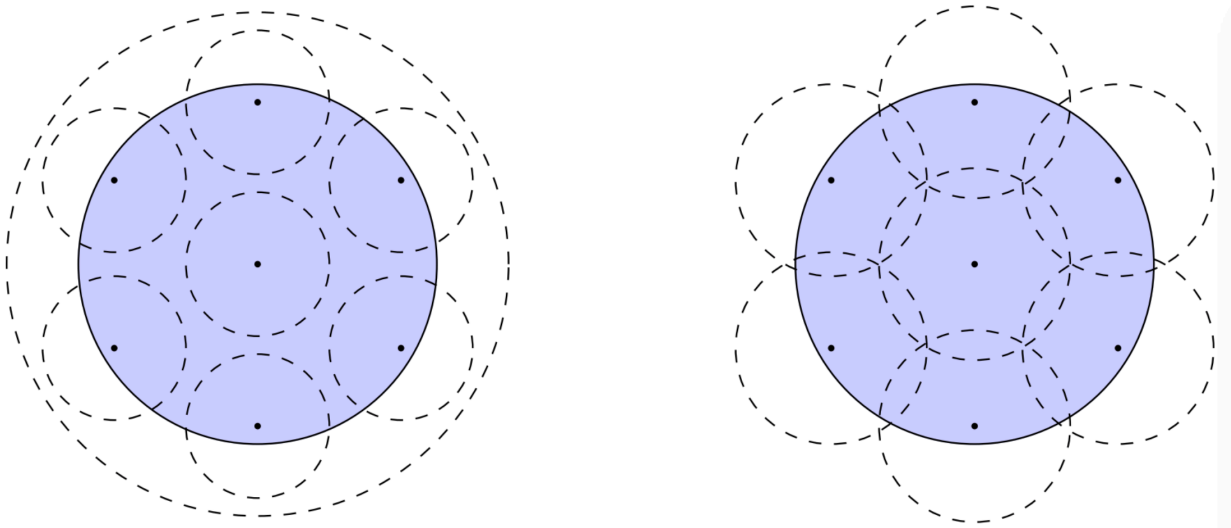

Figure 4.1: Covering (right) and packing (left) [2].

Note that N is a ε-net of T if and only if (see right of Figure 4.1)

$$T \subset \bigcup_{t \in N} B(t, \varepsilon), \quad \text{where } B(t, \varepsilon) = \{s \in T : \ d(t, s) \leq \varepsilon\}.$$

The covering number $N(T, d, \varepsilon)$ can be viewed as a measure of the complexity of T at the scale ε: The more complex T is, the more number of points we need to approximate it up to a certain precision. In addition, the logarithm of the covering number $\log N(T, d, \varepsilon)$ is often called the *metric entropy* of T as it is equivalent to the number of bits needed to encode every points of T up to a prescribed precision ε.

Example 4.7 *Consider the interval $T = [-1, 1]$ with the metric $d(t, t') = |t - t'|$. If we let*

$$N = \{t_k = -1 + 2(k-1)\varepsilon, \ k = 1, \cdots, k_{\max}\}$$

for the $k_{\max}$ such that $t_{k_{\max}} \leq 1$. it is not hard to see that N is an ε-net of T. Thus, we have

$$N(T, d, \varepsilon) \leq \frac{1}{\varepsilon} + 1.$$

Exercise 4.8 *Generalize the above result to the d-dimensional cube $T = [-1, 1]^d$ with $d(t, t') = \|t - t'\|_\infty$ and show that $N(T, d, \varepsilon) \leq \left(\frac{1}{\varepsilon} + 1\right)^d$.*

Definition 4.9 (ε-packing and packing number) *Let (T, d) be a metric space. A set $P \subset T$ is called an ε-packing of (T, d) if for every $t, t' \in P$ and $t \neq t'$, we have $d(t, t') > \varepsilon$. The packing number of (T, d), denoted $P(T, d, \varepsilon)$, is the largest possible cardinality of an ε-packing of (T, d). That is,*

$$P(T, d, \varepsilon) = \sup\{|P| : \ P \text{ is a } \varepsilon\text{-packing of } (T, d)\}.$$

The key idea, which was already hinted at above, is that the notion of packing is dual to the notion of covering (i.e., the typical primal-dual relationship between inf and sup), as given in the following lemma. This means that we can use covering and packing interchangeably. It is often the case that estimating one of them is easier than estimating the other in the applications.

Lemma 4.10 (Dual or equivalence between covering and packing) *For any $\varepsilon > 0$,*

$$P(T, d, 2\varepsilon) \leq N(T, d, \varepsilon) \leq P(T, d, \varepsilon).$$

Proof: **Upper bound.** Let P be a maximal ε-packing of (T, d). Then it is not hard to see that P is also a ε-net of (T, d); otherwise it will violates the maximality. The upper bound follows immediately.

Lower bound. Let $P = \{x_i\}$ be an 2ε-packing of (T, d) and let $N = \{y_j\}$ be a ε-net of (T, d). It can be argued that each closed $B(y_j, \varepsilon)$ ball can only contain one x_i due to the 2ε-separability of $\{x_i\}$. Since each x_i must be contained in one $B(y_j, \varepsilon)$, we must have $|P| \leq |N|$. The lower bound follows due to the arbitrary of P and N. ■

The following lemma studies the covering of unit-norm balls in $\mathbb{R}^d$. The proof is based on a clever technique known as a volume argument.

Lemma 4.11 *Let $\|\cdot\|$ be a norm defined in $\mathbb{R}^d$ (e.g., 1-norm, 2-norm or infinity-norm). Let $\mathbb{B}^d$ be a unit $\|\cdot\|$ ball, i.e., $\mathbb{B}^d = \{t \in \mathbb{R}^d : \|t\| \leq 1\}$. Then*

$$\left(\frac{1}{\varepsilon}\right)^d \leq N(\mathbb{B}^d, \|\cdot\|, \varepsilon) \leq \left(1 + \frac{2}{\varepsilon}\right)^d.$$

Proof: **Lower bound.** Let $N = N(\mathbb{B}^d, \|\cdot\|, \varepsilon)$. We know that B can be covered with N balls of radius ε (see right of Figure 4.1 for an illustration under the 2-norm). Thus,

$$\mathrm{vol}(\mathbb{B}^d) \leq N \mathrm{vol}(\varepsilon \mathbb{B}^d).$$

Noting that $\mathrm{vol}(\varepsilon \mathbb{B}^d) = \varepsilon^d \mathrm{vol}(\mathbb{B}^d)$, the lower bound follows.

Upper bound. For the upper bound we consider $P = P(\mathbb{B}^d, \|\cdot\|, \varepsilon)$. Let $\{x_i\} \subset \mathbb{B}^d$ be the ε-packing of $\mathbb{B}^d$. Construct P balls $B(x_i, \varepsilon/2)$. Then we have

$$\bigcup_{x_i} B(x_i, \varepsilon/2) \subset \mathbb{B}^d + \frac{\varepsilon}{2}\mathbb{B}^d = \left(1 + \frac{\varepsilon}{2}\right)\mathbb{B}^d,$$

see the left of Figure 4.1. Thus, it follows that

$$P \cdot \mathrm{vol}\left(\frac{\varepsilon}{2}\mathbb{B}^d\right) \leq \mathrm{vol}\left(\left(1 + \frac{\varepsilon}{2}\right)\mathbb{B}^d\right).$$

It follows that $P \leq \left(1 + \frac{2}{\varepsilon}\right)^d$, and the upper bound follows by noting Lemma 4.10. ■

The result in Lemma 4.11 for the unit 2-norm ball $\mathbb{B}_2^d$ will be very useful for studying the spectral norm of a random matrix. It follows that the metric entropy of $\mathbb{B}_2^d$ satisfies

$$\log N(\mathbb{B}_2^d, \|\cdot\|_2, \varepsilon) \sim d \log\left(\frac{1}{\varepsilon}\right),$$

which *scales linearly with respect to d*. For some spaces (of functions), the metric entropy scales exponentially with respect to d, hence suffering from the *curse of dimensionality*.

Remark 4.12 *Result and argument in Lemma 4.11 can be generalized to any set T in $\mathbb{R}^d$, see [3].*

So far, we have studied the covering number of various subsets of $\mathbb{R}^d$. Next, we turn to the metric entropy of Lipschitz functions. Consider the function class $\mathcal{F} = \{f \in \mathrm{Lip}\,([0,1], |\cdot|) : \ 0 \le f \le 1\}$. We have the following result.

Lemma 4.13 *There is a numerical constant $c > 0$ such that*

$$N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \le e^{c/\varepsilon} \text{ for } \varepsilon < \frac{1}{2} \quad \text{and} \quad N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) = 1 \text{ for } \varepsilon \ge \frac{1}{2}.$$

Proof: The claim $N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) = 1$ for $\varepsilon \ge \frac{1}{2}$ is trivial since $\left\|f - \frac{1}{2}\right\|_\infty \le \frac{1}{2}$ for each $f \in \mathcal{F}$. To prove the first one, we partition the horizontal axis into consecutive nonoverlapping intervals $I_1, \cdots, I_{\lceil 2/\varepsilon \rceil}$ of length $\varepsilon/2$, and partition the vertical axis into consecutive nonoverlapping intervals $J_1, \cdots, J_{\lceil 1/\varepsilon \rceil}$ of length ε, see Figure 4.2. For each $f \in \mathcal{F}$, we define $\pi(f)$ as follows:

$$\pi(f)(x) = \frac{\max J_\ell + \min J_\ell}{2} \quad \text{for } x \in I_k, \text{ where } J_\ell \text{ is the interval containing } f(\min I_k).$$

Let $N = \{\pi(f) : \ f \in \mathcal{F}\}$. We first show that N is an ε-net of $\mathcal{F}$. This follows from that $\forall x \in I_k$, we have

$$\begin{aligned} |f(x) - \pi(f)(x)| &\le |f(x) - f(\min I_k)| + \left| f(\min I_k) - \frac{\max J_\ell + \min J_\ell}{2} \right| \\ &\le |x - \min I_k| + \left| f(\min I_k) - \frac{\max J_\ell + \min J_\ell}{2} \right| \\ &\le \varepsilon. \end{aligned}$$

It remains to bound $|N|$. A trivial bound for $|N|$ is $|N| \le \lceil 1/\varepsilon \rceil^{\lceil 2/\varepsilon \rceil}$. If we explore the Lipschitz continuity of f more carefully, we can achieve the bound in the lemma, see [2] for details. ■

4.4 Finite Approximation Bound

At last, we consider the general case $\mathbb{E}\left[\sup_{t\in T} X_t\right]$ and present a simple bound via one step of finite approximation. More precisely, the idea is approximating $\mathbb{E}\left[\sup_{t\in T} X_t\right]$ by a finite maximum over a ε-net of T, together with the approximation error.

Theorem 4.14 *Assume X_t is σ^2-sub-Gaussian for every $t \in T$. Then,*

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le \inf_{\varepsilon > 0} \left\{ \mathbb{E}\left[\sup_{t\in T} \left(X_t - X_{\pi(t)}\right)\right] + \sqrt{2\sigma^2 \log N(T, d, \varepsilon)} \right\}.$$

Proof: Let $\varepsilon > 0$ and N be a ε-net of (T, d). For any t, let $\pi(t)$ be the point in N such that $d(t, \pi(t)) \le \varepsilon$. Then,

$$\sup_{t\in T} X_t = \sup_{t\in T} \left(X_t - X_{\pi(t)} + X_{\pi(t)}\right) \le \sup_{t\in T} \left(X_t - X_{\pi(t)}\right) + \sup_{t\in T} X_{\pi(t)}.$$

Taking the expectation on both sides and using the upper bound for the supremum of a finite number of sub-Gaussian random variables (Lemma 4.4 of Lecture 4) yields

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le \mathbb{E}\left[\sup_{t\in T} \left(X_t - X_{\pi(t)}\right)\right] + \sqrt{2\sigma^2 \log N(T, d, \varepsilon)}.$$

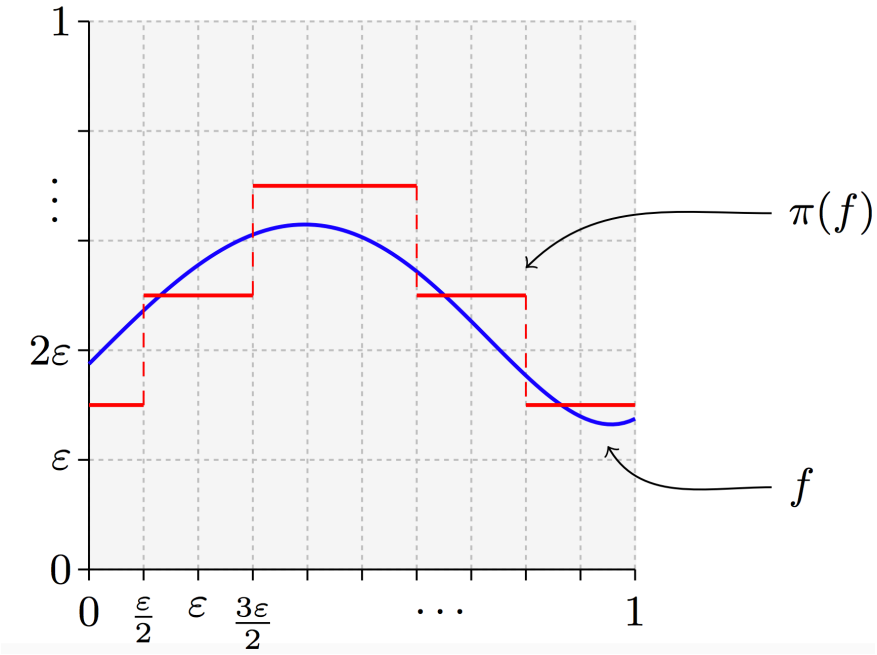

Figure 4.2: Illustration for Lemma 4.13 [2].

Noting that the first term indeed relies on ε since $d(t, \pi(t)) \leq \varepsilon$, taking the infimum over all $\varepsilon > 0$ concludes the proof. ■

There is a trade-off in the bound of Theorem 4.14. When ε decreases, the first term will potentially become smaller since X_t becomes closer to $X_{\pi(t)}$, but the second term increases as the covering number increases under a decreasing precision.

Example 4.15 (Maximum singular value of sub-Gaussian random matrix) *Let $W \in \mathbb{R}^{m \times n}$ be a random matrix with i.i.d σ^2-sub-Gaussian entries. We would like to estimate*

$$\mathbb{E}\left[\|W\|_2\right].$$

By the variational formula of the spectral norm,

$$\|W\|_2 = \sup_{u \in \mathbb{B}_2^m, v \in \mathbb{B}_2^n} u^T W v,$$

*it is equivalent to bound $\mathbb{E}\left[\sup_{u \in \mathbb{B}_2^m, v \in \mathbb{B}_2^n} u^T W v\right]$. Firstly note that $u^T W v$ is σ^2-sub-Gaussian for every $u \in \mathbb{B}_2^m$ and $v \in \mathbb{B}_2^n$ (**verify this!**). Let M be an ε-net of $\mathbb{B}_2^m$ and N be an ε-net of $\mathbb{B}_2^n$. By Theorem 4.14, we have*

$$\mathbb{E}\left[\|W\|_2\right] \leq \mathbb{E}\left[\sup_{u \in \mathbb{B}_2^m, v \in \mathbb{B}_2^n} \left(u^T W v - \pi(u)^T W \pi(v)\right)\right] + \sqrt{2\sigma^2 \log |M||N|}, \quad \forall \varepsilon > 0.$$

- *By Lemma 4.11, we can choose $M \leq (1 + 2/\varepsilon)^m$ and $N \leq (1 + 2/\varepsilon)^n$.*

- *Because for any* $u \in \mathbb{B}_2^m, v \in \mathbb{B}_2^n$,

$$\begin{aligned} |u^T W v - \pi(u)^T W \pi(v)| &\le |(u - \pi(u))^T W v| + |\pi(u)^T W (v - \pi(v))| \\ &\le 2\varepsilon \|W\|_2, \end{aligned}$$

we have $\mathbb{E}\left[\sup_{u\in\mathbb{B}_2^m, v\in\mathbb{B}_2^n} \left(u^T W v - \pi(u)^T W \pi(v)\right)\right] \le 2\varepsilon \mathbb{E}\left[\|W\|_2\right]$.

Combining all them together yields

$$\mathbb{E}\left[\|W\|_2\right] \le \frac{1}{1-2\varepsilon}\sqrt{2\sigma^2(m+n)\log(3/\varepsilon)}, \quad \forall \varepsilon > 0.$$

Taking ε *to be a small constant (e.g.,* $\varepsilon = 1/4$*) yields that*

$$\mathbb{E}\left[\|W\|_2\right] \lesssim \sigma(\sqrt{m} + \sqrt{n}).$$

As can be seen in the next lecture, this crude bound already captures the correct (tight) order of magnitude of the matrix norm.

A careful reader may find out that what have used in the above analysis is essentially the result

$$\|W\|_2 = \sup_{u\in\mathbb{B}_2^m, v\in\mathbb{B}_2^n} u^T W v \le \frac{1}{1-2\varepsilon} \sup_{u\in M, v\in N} u^T W v.$$

Moreover, because the remaining term is of the same order with the target to bound but with a smaller factor, i.e.,

$$\mathbb{E}\left[\sup_{u\in\mathbb{B}_2^m, v\in\mathbb{B}_2^n} \left(u^T W v - \pi(u)^T W \pi(v)\right)\right] \le 2\varepsilon \mathbb{E}\left[\|W\|_2\right],$$

optimal bound (in order) can be achieved for this case. But sometimes, the bound for the remaining term obtained via invoking the Lipschitz property is inefficient, so the finite approximate scheme only yields sub-optimal bound.

Example 4.16 (Uniform law of large numbers over Lipschitz functions (sub-optimal bound)) *Consider the metric space* $([0,1], |\cdot|)$ *and a probability measure* $\mathbb{P}$ *defined on it. Let* $\mathcal{F} = \{f \in \mathrm{Lip}\left([0,1], |\cdot|\right) : \ 0 \le f \le 1\}$. *Let* $X_1, \cdots, X_n \sim \mathbb{P}$ *be i.i.d samples. A key step when studying the uniform law of large numbers is to bound*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{k=1}^n f(X_k) - \mathbb{E}\left[f(X)\right] \right| \right].$$

For each of notation, let $Z_f = \frac{1}{n}\sum_{k=1}^n f(X_k) - \mathbb{E}\left[f(X)\right]$. *Without loss of generality (**why we can make the simplification?**), we consider*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} Z_f\right].$$

*First note that Z_f is $1/4n$-sub-Gaussian since $f \in [0,1]$ (**check this!**). Let N be the ε-net of $(\mathcal{F}, \|\cdot\|_\infty)$, by Lemma 4.13, we have $N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \leq e^{c/\varepsilon}$, for $\varepsilon < 1/2$. Thus, the application of Theorem 4.14 yields that*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} Z_f\right] \leq \inf_{0<\varepsilon<1/2}\left\{\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(Z_f - Z_{\pi(f)}\right)\right] + \sqrt{\frac{c}{2n\varepsilon}}\right\}$$

Moreover, we have

$$\begin{aligned} Z_f - Z_{\pi(f)} &= \left(\frac{1}{n}\sum_{k=1}^n \left(f(X_k) - \pi(f)(X_k)\right)\right) + \mathbb{E}\left[\pi(f)(X) - f(X)\right] \\ &\leq \frac{2}{n}\sum_{k=1}^n \|f - \pi(f)\|_\infty \\ &\leq 2\varepsilon. \end{aligned}$$

It follows that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} Z_f\right] \leq \inf_{0<\varepsilon<1/2}\left\{2\varepsilon + \sqrt{\frac{c}{2n\varepsilon}}\right\} \asymp n^{-1/3}.$$

*However, this rate is **not** tight. Note that for a single function,*

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{k=1}^n f(X_k) - \mathbb{E}\left[f(X)\right]\right|\right] \leq \left(\mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^n f(X_k) - \mathbb{E}\left[f(X)\right]\right)^2\right]\right)^{1/2} \lesssim n^{-1/2}.$$

*Thus, it would be more desirable if we can still get the same $n^{-1/2}$ rate for $\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^n f(X_k) - \mathbb{E}\left[f(X)\right]\right|\right]$. In the next section, we will show that this is **indeed true**.*

Reading Materials

[1] Martin Wainwright, *High Dimensional Statistics – A non-asymptotic viewpoint*, Chapters 5.1, 5.2 and 5.3.

[2] Ramon van Handel, *Probability in High Dimension*, Chapter 5.1, 5.2.

[3] Roman Vershynin, *High-Dimensional Probability: An introduction with applications in data science*, Chapter 4.2, 4.4.