

STA3000F: Advanced Theory of Statistics

Fall 2026

Lecture 6: October 16, 2026

Lecturer: Wenlong Mou

1 Covering and packing numbers

Lemma 6.1 (duality). *For any set K and metric ρ , we have*

$$M(K; \rho, 2\epsilon) \stackrel{(i)}{\leq} N(K; \rho, \epsilon) \stackrel{(ii)}{\leq} M(K; \rho, \epsilon).$$

Roughly speaking, if you do not care about constant factors in the radius, covering number and packing number are equivalent.

Proof of (i): Let $(\theta_i)_{i=1}^M$ be a 2ϵ -packing of K , and let $(\psi_i)_{i=1}^N$ be an ϵ -covering of K . We need to show that $M \leq N$. By 2ϵ -packing, we know that $\mathbb{B}_\rho(\theta_i, \epsilon) \cap \mathbb{B}_\rho(\theta_j, \epsilon) = \emptyset$ for $i \neq j$. On the other hand, since $(\psi_i)_{i=1}^N$ is an ϵ -covering, for each θ_i , there exists some ψ_j such that $\theta_i \in \mathbb{B}_\rho(\psi_j, \epsilon)$. For different θ_i and θ_ℓ , they cannot be covered by the same ψ_j , otherwise we would have $\rho(\theta_i, \theta_\ell) \leq \rho(\theta_i, \psi_j) + \rho(\psi_j, \theta_\ell) < 2\epsilon$, which contradicts the 2ϵ -packing property. So for each i , there exists j such that $\theta_i \in \mathbb{B}_\rho(\psi_j, \epsilon)$, and different i 's correspond to different j 's. This shows that $M \leq N$.

Proof of (ii): Let $(\theta_i)_{i=1}^M$ be a maximal ϵ -packing. That means we cannot add any more points to it and maintain the ϵ -packing property. Therefore, for any $\theta \in K$, there exists some θ_i such that $\rho(\theta, \theta_i) < \epsilon$, otherwise we can add θ to the packing. This shows that $(\theta_i)_{i=1}^M$ is also an ϵ -covering of K . Therefore, $N(K; \rho, \epsilon) \leq M(K; \rho, \epsilon)$.

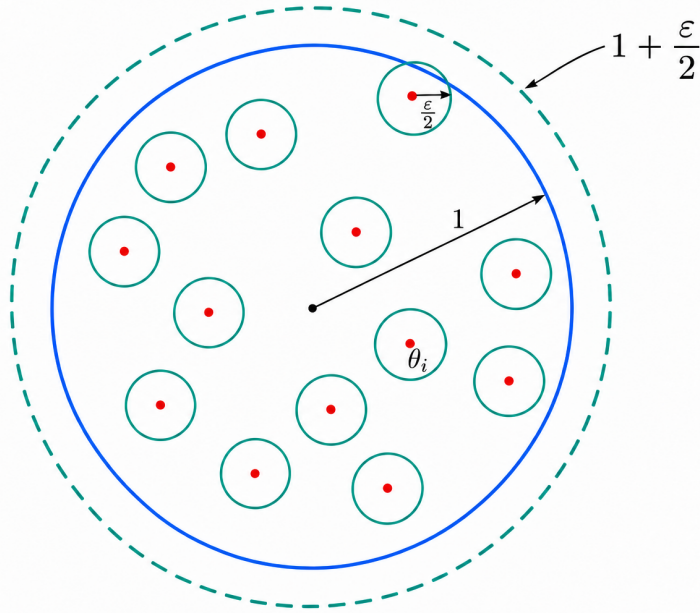
1.1 Examples

Let $K = \mathbb{B}_2^d(1)$, the unit ball in $\mathbb{R}^d$ under ℓ_2 norm. We want to compute the covering number $N(K; \|\cdot\|_2, \epsilon)$.

Upper bound: Suppose $(\theta_i)_{i=1}^M$ is an ϵ -packing of K . Then we have

$$\bigcup_{i=1}^M \mathbb{B}_2^d(\theta_i, \epsilon/2) \subseteq \mathbb{B}_2^d(1 + \epsilon/2).$$

The geometry behind this inclusion is illustrated below: the small balls are disjoint, and each is contained in the enlarged ball.



Volume comparison for an ϵ -packing. The blue circle is the unit ball; the dashed circle has radius $1 + \epsilon/2$. Each small ball has radius $\epsilon/2$ and is centered at a packing point. Schematic, not to scale.

This implies that

$$M \cdot \text{vol}(\mathbb{B}_2^d(\epsilon/2)) \leq \text{vol}(\mathbb{B}_2^d(1 + \epsilon/2)),$$

which gives us

$$M \leq \left(1 + \frac{2}{\epsilon}\right)^d,$$

and by duality, we have

$$N(K; \|\cdot\|_2, \epsilon) \leq \left(1 + \frac{2}{\epsilon}\right)^d.$$

Lower bound: Suppose $(\psi_i)_{i=1}^N$ is an ϵ -covering of K . Then we have

$$\text{vol}(\mathbb{B}_2^d(1)) \leq N \cdot \text{vol}(\mathbb{B}_2^d(\epsilon)),$$

which gives us

$$N \geq \frac{\text{vol}(\mathbb{B}_2^d(1))}{\text{vol}(\mathbb{B}_2^d(\epsilon))} = \left(\frac{1}{\epsilon}\right)^d.$$

Technically, the gap between the upper bound and lower bound is a factor of 2^d . But in statistical applications, we usually use log covering number, and the resulting difference is just a constant factor. In particular, we have

$$\log N(K; \|\cdot\|_2, \epsilon) \asymp d \log(1/\epsilon).$$

2 Application to M -estimation

Recall that we are interested in

$$\hat{\theta}_n \in \arg \min_{\theta \in \Theta} F_n(\theta), \quad \theta^* \in \arg \min_{\theta \in \Theta} F(\theta).$$

We need to show uniform convergence of $F_n(\theta)$ to $F(\theta)$.

Theorem 6.1 (Wald). *Suppose that Θ is compact (with respect to some metric ρ), and $f(x; \theta)$ is continuous in θ (with respect to ρ) for each x . Suppose furthermore that*

$$\mathbb{E} \left[\sup_{\theta \in \Theta} |f(X; \theta)| \right] < +\infty.$$

(i.e., the class has an integrable envelope function). Then we have

$$\sup_{\theta \in \Theta} |F_n(\theta) - F(\theta)| \xrightarrow{P} 0.$$

Proof. Fix $\varepsilon > 0$. Since $f(x; \theta)$ is continuous in θ and Θ is compact, we know that F is uniformly continuous on Θ . Therefore, there exists $\delta > 0$ such that for any $\theta_1, \theta_2 \in \Theta$ with $\rho(\theta_1, \theta_2) < \delta$, we have $|F(\theta_1) - F(\theta_2)| < \varepsilon$. (This is used to control the discretization error).

Let $(\theta_i)_{i=1}^N$ be a δ -covering of Θ . Since the set Θ is compact, we know that $N < +\infty$.

$$\sup_{\theta \in \Theta} |F_n(\theta) - F(\theta)| \leq \max_{i=1, \dots, N} \sup_{\theta: \rho(\theta, \theta_i) \leq \delta} \left\{ \underbrace{|F_n(\theta_i) - F(\theta_i)|}_{\max_i(\dots) \xrightarrow{P} 0, \text{ T}_1} + \underbrace{|F(\theta_i) - F(\theta)|}_{\leq \varepsilon} + \underbrace{|F_n(\theta) - F_n(\theta_i)|}_{\text{problematic term, T}_3} \right\}$$

We can upper bound the problematic term as follows:

$$\begin{aligned} \text{problematic term} &= \max_{i=1, \dots, N} \sup_{\theta: \rho(\theta, \theta_i) \leq \delta} \left| \frac{1}{n} \sum_{j=1}^n (f(X_j; \theta) - f(X_j; \theta_i)) \right| \\ &\leq \max_{i=1, \dots, N} \frac{1}{n} \sum_{j=1}^n \underbrace{\sup_{\theta: \rho(\theta, \theta_i) \leq \delta} |f(X_j; \theta) - f(X_j; \theta_i)|}_{\text{i.i.d. average of this}} \end{aligned}$$

By law of large numbers, we have

$$\begin{aligned} \max_{i=1, \dots, N} \frac{1}{n} \sum_{j=1}^n \sup_{\theta: \rho(\theta, \theta_i) \leq \delta} |f(X_j; \theta) - f(X_j; \theta_i)| &\xrightarrow{P} \max_{i=1, \dots, N} \mathbb{E} \left[\sup_{\theta: \rho(\theta, \theta_i) \leq \delta} |f(X; \theta) - f(X; \theta_i)| \right] \\ &\leq \sup_{\theta \in \Theta} \mathbb{E} \left[\sup_{\theta': \rho(\theta, \theta') \leq \delta} |f(X; \theta) - f(X; \theta')| \right], \end{aligned}$$

which is the worst-case modulus-of-continuity of $f(X; \theta)$ in θ , at scale δ .

For any $\theta \in \Theta$, define

$$M_\delta(\theta) := \sup_{\theta': \rho(\theta, \theta') \leq \delta} |f(X; \theta) - f(X; \theta')|.$$

We know that for each fixed θ , we have pointwise convergence $M_\delta(\theta) \rightarrow 0$ as $\delta \rightarrow 0$, by continuity of $f(X; \theta)$ in θ . On the other hand, we also have $M_\delta(\theta) \leq 2 \sup_{\theta \in \Theta} |f(X; \theta)|$, which is integrable by assumption. Therefore, by dominated convergence theorem, we have

$$\mathbb{E}[M_\delta(\theta)] \rightarrow 0, \quad \forall \theta \in \Theta.$$

By uniform continuity of $f(X; \cdot)$ on compact Θ and dominated convergence, $\mathbb{E}[\sup_{\theta \in \Theta} M_\delta(\theta)] \rightarrow 0$, which gives the required uniform convergence in θ .

Remarks:

1. This only needs very weak conditions, but only proves asymptotic convergence.
2. Using this result, and in combination with basic inequality, we have $F(\hat{\theta}_n) - F(\theta^*) \xrightarrow{p} 0$. Furthermore, under the same setup, F is continuous, and assuming that θ^* is unique minimizer, then we have $\hat{\theta}_n \xrightarrow{p} \theta^*$.
3. Can we turn this into a more quantitative bound?

For the term T_1 , we note that

$$\begin{aligned} & \mathbb{P}\left(\max_{i=1, \dots, N} |F_n(\theta_i) - F(\theta_i)| > \varepsilon\right) \\ & \leq \sum_{i=1}^N \mathbb{P}(|F_n(\theta_i) - F(\theta_i)| > \varepsilon) \\ & \leq N \cdot \max_{i=1, \dots, N} \mathbb{P}(|F_n(\theta_i) - F(\theta_i)| > \varepsilon). \end{aligned}$$

If $0 \leq f \leq 1$, then we can use Hoeffding's inequality to get

$$\mathbb{P}\left(\max_{i=1, \dots, N} |F_n(\theta_i) - F(\theta_i)| > \varepsilon\right) \leq 2Ne^{-2n\varepsilon^2}.$$

By taking $\varepsilon = \sqrt{\frac{\log(2N/\delta)}{2n}}$, we have the high-probability bound.

On the other hand, if f is unbounded (or even heavy-tailed), we cannot directly apply Hoeffding's inequality. Naïve application of Chebyshev's inequality gives us $\varepsilon \asymp \sqrt{N/n}$, which is not useful when N is large.

This can be overcome by “symmetrization method” and Rademacher complexity.

For the term T_3 , we can bound it as long as we have quantitative bound on the modulus-of-continuity of $f(X; \theta)$ in θ . But on the other hand, this type of one-step discretization argument is fundamentally sub-optimal. We can do better by using “chaining method”, which is a multi-step discretization argument.

3 Symmetrization and chaining methods

3.1 Symmetrization method

Introducing the notations:

$$P_n f = \frac{1}{n} \sum_{i=1}^n f(X_i), \quad P f = \mathbb{E}[f(X)], \quad (P_n - P)f = P_n f - P f.$$

For notation simplicity, we drop the parametrization by θ and write $f(X)$ instead of $f(X; \theta)$. We are interested in bounding

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} |(P_n - P)f|\right].$$

Here $\mathcal{F} = \{f(X; \theta) : \theta \in \Theta\}$ is the function class induced by the parametrization $\theta \in \Theta$.

The “symmetrized version” of the above quantity is

$$\mathcal{R}_n(\mathcal{F}) := \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right], \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \text{Rademacher}(\pm 1),$$

which is called the Rademacher complexity of the function class $\mathcal{F}$. The Rademacher complexity is a measure of the richness of the function class $\mathcal{F}$, and it is often easier to bound than the original quantity, because we can bound it by conditioning on $X_1, \dots, X_n$ and using concentration inequalities for Rademacher random variables.

Theorem 6.2 (symmetrization). *We have*

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} |(P_n - P)f| \right] \leq 2\mathcal{R}_n(\mathcal{F}).$$

Proof: Let $X'_1, \dots, X'_n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}$ be an independent copy of $X_1, \dots, X_n$. Then we have

$$\begin{aligned} \mathbb{E} \left[\sup_{f \in \mathcal{F}} |(P_n - P)f| \right] &= \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n (f(X_i) - \mathbb{E}[f(X'_i)]) \right| \right] \\ &\stackrel{(\text{Jensen})}{\leq} \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n (f(X_i) - f(X'_i)) \right| \right] \end{aligned}$$

We know that by introducing Rademacher random variables $\varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \text{Rademacher}(\pm 1)$, we have

$$(f(X_i) - f(X'_i))_{i=1}^n \stackrel{d}{=} (\varepsilon_i (f(X_i) - f(X'_i)))_{i=1}^n,$$

so we have

$$\text{above} = \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i (f(X_i) - f(X'_i)) \right| \right] \leq 2\mathcal{R}_n(\mathcal{F}).$$

Application of symmetrization method: We can reason conditionally on $X_1, \dots, X_n$, and Rademacher random variables are bounded and nicely behaved in terms of concentration inequalities.

In particular, for any $A \subseteq \mathbb{R}^n$, we define the Rademacher complexity of A as

$$R_n(A) := \mathbb{E} \left[\sup_{a \in A} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i a_i \right| \right], \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \text{Rademacher}(\pm 1).$$

This has nothing to do with the data or statistical problem, and is purely a combinatorial measure of the complexity of the set A .

Following the arguments in the previous section, we can bound it using concentration inequality.

For finite A : for any $\lambda > 0$, by Jensen's inequality,

$$\begin{aligned}
 R_n(A) &\leq \frac{1}{\lambda} \log \mathbb{E} \left[\max_{a \in A \cup (-A)} \exp \left(\lambda \frac{1}{n} \sum_{i=1}^n \varepsilon_i a_i \right) \right] \\
 &\leq \frac{1}{\lambda} \log \left\{ \sum_{a \in A \cup (-A)} \mathbb{E} \left[\exp \left(\lambda \frac{1}{n} \sum_{i=1}^n \varepsilon_i a_i \right) \right] \right\} \\
 &\leq \frac{\log(2|A|)}{\lambda} + \frac{\lambda}{2n^2} \max_{a \in A} \sum_{i=1}^n a_i^2 \\
 &\stackrel{(\text{choosing optimal } \lambda)}{=} \sqrt{\frac{2 \log(2|A|)}{n}} \cdot \max_{a \in A} \sqrt{\frac{\sum_{i=1}^n a_i^2}{n}}.
 \end{aligned}$$

3.2 Discretization and chaining

Following the same idea as before, we bound $R_n(A)$ for a bounded set $A \subseteq \mathbb{R}^n$ containing 0.

One-step discretization: Let $(a^{(i)})_{i=1}^N$ be an ε -covering of A with respect to the normalized metric

$$\|a\|_n = \sqrt{\frac{1}{n} \sum_{i=1}^n a_i^2}.$$

Denote $\pi(a) := \arg \min_{a^{(i)}} \|a - a^{(i)}\|_n$ as the projection of a onto the covering set. Then we have

$$\begin{aligned}
 R_n(A) &\leq \underbrace{\mathbb{E} \left[\sup_{a \in A} n^{-1} |\varepsilon^\top (a - \pi(a))| \right]}_{(\text{C.-S.}) \leq \sup_{a \in A} \|a - \pi(a)\|_n} + \underbrace{\mathbb{E} \left[\sup_{a \in A} n^{-1} |\varepsilon^\top \pi(a)| \right]}_{\leq \sqrt{\frac{2 \log(2N)}{n}} \cdot \text{diam}(A)} \\
 &\leq \varepsilon + \sqrt{\frac{2 \log(2N(\varepsilon))}{n}} \cdot \text{diam}(A).
 \end{aligned}$$

Is this argument tight?

Consider A being a unit ball in d -dimensional subspace (this could come from using a d -dimensional linear model for $\mathcal{F}$) By choosing $\varepsilon = O(1/n)$, we have

$$R_n(A) \leq O\left(\sqrt{\frac{d \log n}{n}}\right).$$

For linear model, by doing direct calculation, we can show that $R_n(A) \asymp \sqrt{d/n}$. So the above bound is sub-optimal by a factor of $\sqrt{\log n}$. In general, this sub-optimality can go beyond log factors for nonparametric function classes.

Multi-step discretization: Potential looseness of one-step discretization: $(a - \pi(a))$ may contain some structures, but we just bound it using Cauchy–Schwarz.

Now let us consider a sequence of discrete approximations to A (not necessarily nested).

$$\{0\} = A_0, A_1, \dots, A_m, \dots$$

Denote $\pi_m(a) := \arg \min_{a^{(i)} \in A_m} \|a - a^{(i)}\|_n$ as the projection of a onto the m -th covering set.

Concretely, we choose A_m to be δ_m -covering of A with respect to the normalized metric $\|\cdot\|_n$ (with $\delta_m \downarrow 0$ to be chosen). For each $a \in A$, we have the decomposition

$$a = \sum_{m=0}^{\infty} (\pi_{m+1}(a) - \pi_m(a)).$$

So we have

$$\mathbb{E} \left[\sup_{a \in A} n^{-1} |\varepsilon^\top a| \right] \leq \sum_{m=0}^{\infty} \mathbb{E} \left[\sup_{a \in A} n^{-1} |\varepsilon^\top (\pi_{m+1}(a) - \pi_m(a))| \right].$$

Each term measures the complexity at scale δ_m , and each term takes supremum over a finite set. So we have

$$\mathbb{E} \left[\sup_{a \in A} n^{-1} |\varepsilon^\top (\pi_{m+1}(a) - \pi_m(a))| \right] \leq (\delta_m + \delta_{m+1}) \cdot \sqrt{\frac{2 \log(2|A_m| \cdot |A_{m+1}|)}{n}}$$

By manipulating this type of bounds carefully, we can arrive at a nice form

Theorem 6.3 (Dudley's chaining bound). *There exists a universal constant $c > 0$ such that*

$$R_n(A) \leq c \int_0^{\text{diam}(A)} \sqrt{\frac{\log(2N(A; \|\cdot\|_n, \epsilon))}{n}} d\epsilon.$$

Proof: Let D be the diameter of A under $\|\cdot\|_n$. Let $\delta_m = 2^{-m}D$, and we have

$$\delta_m + \delta_{m+1} = 3 \cdot 2^{-m-1}D, \quad |A_m| \cdot |A_{m+1}| \leq |A_{m+1}|^2 = N(A; \|\cdot\|_n, \delta_{m+1})^2.$$

Substituting to the bound, we have

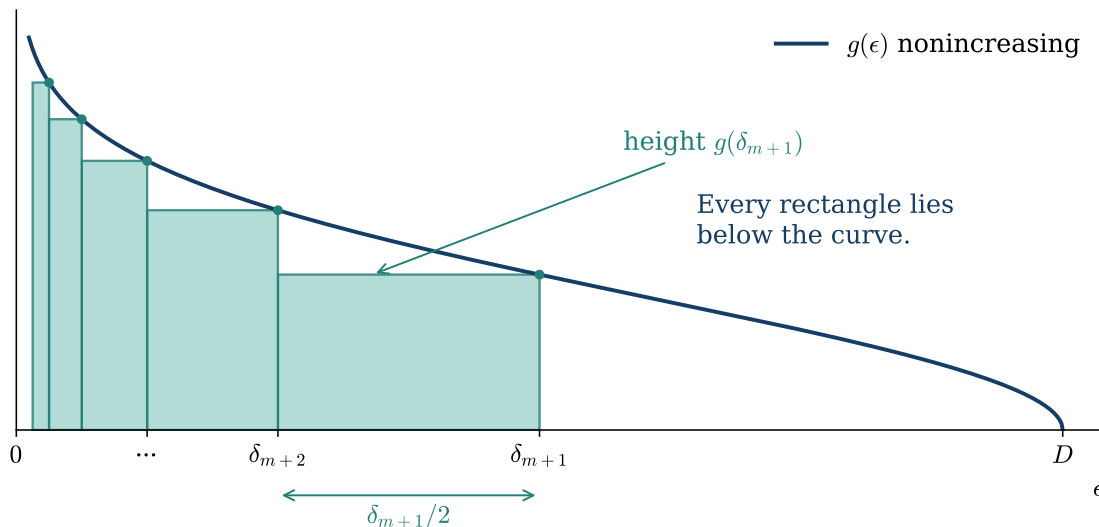
$$R_n(A) \leq \sum_{m=0}^{\infty} 3 \cdot 2^{-m-1}D \cdot \sqrt{\frac{4 \log(2N(A; \|\cdot\|_n, \delta_{m+1}))}{n}}$$

To compare this dyadic sum with an integral, put

$$g(\epsilon) = \sqrt{\frac{\log(2N(A; \|\cdot\|_n, \epsilon))}{n}}.$$

The covering number, and hence g , is nonincreasing in ϵ . On $[\delta_{m+2}, \delta_{m+1}]$, we therefore have $g(\epsilon) \geq g(\delta_{m+1})$. Since $\delta_{m+2} = \delta_{m+1}/2$, the rectangle shown below gives

$$\delta_{m+1}g(\delta_{m+1}) \leq 2 \int_{\delta_{m+2}}^{\delta_{m+1}} g(\epsilon) d\epsilon. \quad (6.1)$$



Right-endpoint rectangles lie below a nonincreasing curve. The displayed curve is illustrative; the same area comparison applies to the entropy function g above.

The preceding chaining sum is $6 \sum_{m \geq 0} \delta_{m+1} g(\delta_{m+1})$. Summing (6.1) over disjoint adjacent intervals yields

$$\begin{aligned} 6 \sum_{m=0}^{\infty} \delta_{m+1} g(\delta_{m+1}) &\leq 12 \int_0^{D/2} g(\epsilon) d\epsilon \\ &\leq 12 \int_0^D g(\epsilon) d\epsilon. \end{aligned}$$

This proves the sum-to-integral comparison, including the constant factor. For an infinite sum, apply the comparison first to finite partial sums and then use monotone convergence; if the integral diverges, the inequality remains valid but is uninformative.

Remarks:

- Dudley integral may diverge in certain case (when the log-covering number grows too fast). In this case, you may consider a truncated version of the integral. Using a finite collection of scales, and bounding the rest using Cauchy–Schwarz, we can still get a finite bound.

$$R_n(A) \leq 4\delta + c \int_{\delta}^{\text{diam}(A)} \sqrt{\frac{\log(2N(A; \|\cdot\|_n, \epsilon))}{n}} d\epsilon.$$

You can further optimize over δ to get the best bound.

- Dudley’s chaining bound is tight for many function classes, but not always. There are more advanced chaining techniques (generic chaining) that can give tight bounds for all function classes.

3.3 Back to function classes

Recall that we want to bound $\mathbb{E} \left[\sup_{f \in \mathcal{F}} |(P_n - P)f| \right]$. We assume (as in Wald’s theorem) that $\mathcal{F}$ has an envelope function G , i.e., $|f(X)| \leq G(X)$ for all $f \in \mathcal{F}$. But here we assume that $\mathbb{E}[G^2(X)] < +\infty$.

Theorem 6.4. *Under above setup, we have*

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} |(P_n - P)f| \right] \leq c \sqrt{\frac{\mathbb{E}[G^2(X)]}{n}} \int_0^1 \sqrt{\log(2 \sup_Q N(\mathcal{F}; \mathbb{L}^2(Q), \|G\|_{\mathbb{L}^2(Q)} \epsilon))} d\epsilon,$$

Chaining under the worst-case distribution Q .

Proof: Applying chaining to the symmetrized process (adjoining zero if necessary), we have the conditional expectation bound:

$$\mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right| \middle| X_1, \dots, X_n \right] \leq \frac{c}{\sqrt{n}} \int_0^{\sqrt{P_n G^2}} \sqrt{\log(2N(\mathcal{F}; \mathbb{L}^2(P_n), \epsilon))} d\epsilon.$$

The metric and integral upper limit are data-dependent. We use change-of-variable to pull this integral upper limit outside, and then replace P_n using the worst-case distribution Q . Averaging over the data and applying symmetrization gives the final bound. For the envelope function, we use the bound

$$\mathbb{E} [\sqrt{P_n G^2}] \leq \sqrt{\mathbb{E}[P_n G^2]} = \sqrt{\mathbb{E}[G^2(X)]}.$$