IEMS 402: Statistical Learning **Winter 2024-2025**

Lecture 8: Uniform Convergence

Lecturer: Yiping Lu *Scribes: Dongyun Kim*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

This lecture and the following scribe notes have references from

- [Ref-1]: Lecture Notes for Machine learning theory (CS229M/STATS214) taught by Tengyu Ma
- [Ref-2]: Scribe notes for Information-theoretic methods in high-dimensional statistics (ECE598) taught by Yihong Wu
- [Ref-3]: Scribe notes for Statistical learning theory (CS281B/Stat241B) taught by Peter Bartlett

8.1 (Recap) Uniform bounds

Recall the following notations:

- Θ: a parameter space
- $\mathcal{H}$: hypothesis class parameterized by $\theta \in \Theta$ (e.g. $\mathcal{H} = \{h : h_\theta(x) = \theta^T x, \theta \in \Theta\}$)
(Note, $|H|= |\Theta|$ and we refer to the hypothesis space as Θ or $\mathcal{H}$ interchangeably throughout this scribe)
- $L(\theta)$: the *expected risk* (over the population P) of the function in $\mathcal{H}$ that uses parameter $\theta \in \Theta$
- $\hat{L}(\theta)$ the *empirical risk* (over a sample data $\hat{P}_n$) of the function in $\mathcal{H}$ that uses parameter $\theta \in \Theta$

Our goal is to bound the *excess risk* $L(\hat{\theta}) - L(\theta^*)$ of the *empirical risk minimization* (ERM) estimator $\hat{\theta} := \arg\min_{\theta\in\Theta} \hat{L}(\theta)$, where $\theta^* := \arg\min_{\theta\in\Theta} L(\theta)$ is the 'best' estimator possible.

The excess risk can be decomposed as

$$\underbrace{L(\hat{\theta}) - L(\theta^*)}_{\text{excess risk}} = \underbrace{L(\hat{\theta}) - \hat{L}(\hat{\theta})}_{\textbf{generalization error(1)}} + \underbrace{\hat{L}(\hat{\theta}) - \hat{L}(\theta^*)}_{\text{optimization error}} + \underbrace{\hat{L}(\theta^*) - \inf_{\theta\in\Theta} L(\theta)}_{\textbf{generalization error(2)}} + \underbrace{\inf_{\theta\in\Theta} L(\theta) - L(\theta^*)}_{\text{approximation error}}$$

Observe that the optimization error term is non-negative, so it can be ignored in terms of deriving an upper bound for the excess risk. Also, throughout this lecture, we ignore the approximation error. Recall that in Lecture 1, we had a uniform bound $2\sup_{\theta\in\Theta} |L(\theta) - \hat{L}(\theta)|$ for the two generalization error terms above. Our goal is to estimate this uniform bound.

Instead of dealing with this uniform bound, it is tempting to deal with the generalization error terms directly. However, since estimator $\hat{\theta}$ is trained from the sample data, and the empirical risk $\hat{L}(\cdot)$ is also computed by taking the expectation over the sample data, $\hat{\theta}$ and $\hat{L}(\cdot)$ are correlated. This may cause concentration inequalities to not work. For example, the generalization error(2) term can be bounded by

$\tilde{O}(\frac{1}{\sqrt{n}})$ via Hoeffding's inequality. However, we cannot apply Hoeffding's inequality for the generalization error(1) term because the independent condition of the Hoeffding's inequality does not hold (See Section 4.1 of [Ref-1]).

Uniform convergence is one way we can handle this issue. Since $\mathbf{Pr}[\cup A_i] \leqslant \sum_i \mathbf{Pr}[A_i]$, we can create the general bound as

$$\mathbf{Pr}\left[\exists \theta \in \Theta \text{ s.t. } |\hat{L}(\theta) - L(\theta)| \geqslant \varepsilon\right] \leqslant \sum_{\theta \in \Theta} \mathbf{Pr}\left[|\hat{L}(\theta) - L(\theta)| \geqslant \varepsilon\right]. \tag{8.1}$$

provided that we can derive a bound for the summands and on the right-hand-side term above. We can then use Hoeffding's inequality to deal with the summands as the θ there is no longer data-dependent.

8.2 Finite hypothesis class

In this section, we assume that $\mathcal{H}$ is finite. The following theorem gives a bound for the excess risk $L(\hat{h}) - L(h^*)$, where $\hat{h}$ and h^* are the minimizers of the empirical loss and population loss, respectively.

Theorem 8.1 *Suppose that our hypothesis class $\mathcal{H}$ is finite and that our loss function ℓ is bounded in $[0,1]$, i.e. $0 \leqslant \ell((x,y),h) \leqslant 1$. Then $\forall \delta$ s.t. $0 < \delta < \frac{1}{2}$, with probability at least $1-\delta$, we have*

$$|L(h) - \hat{L}(h)| \leqslant \sqrt{\frac{\ln|\mathcal{H}| + \ln(2/\delta)}{2n}} \quad \forall h \in \mathcal{H} \tag{8.2}$$

We also have

$$L(\hat{h}) - L(h^*) \leqslant \sqrt{\frac{2(\ln|\mathcal{H}| + \ln(2/\delta))}{n}} \tag{8.3}$$

Proof: *(Sketch)* Recall that *a bounded random variable is sub-Gaussian.* Then by Hoeffding's inequality,

$$\mathbf{Pr}[|\hat{L}(h) - L(h)| \geqslant \varepsilon] \leqslant 2\exp(-2n\varepsilon^2)$$

Summing over all hypotheses of $\mathcal{H}$ we get

$$\mathbf{Pr}[\exists h \in \mathcal{H} \text{ s.t. } |\hat{L}(h) - L(h)| \geqslant \varepsilon] \leqslant \sum_{h \in \mathcal{H}} \mathbf{Pr}[|\hat{L}(h) - L(h)| \geqslant \varepsilon] = 2|\mathcal{H}|\exp(-2n\varepsilon^2)$$

Taking $\delta = 2|\mathcal{H}|\exp(-2n\varepsilon^2)$ yields $\varepsilon = \sqrt{\frac{\ln|\mathcal{H}| + \ln(2/\delta)}{2n}}$. The second result can be obtained by bounding the excess risk with this uniform bound. See Theorem 4.1 of [Ref-1] for details. ■

In summary, when the hypothesis class is finite, we have a probabilistic bound of the excess risk of order

$$\sqrt{\frac{\ln(\text{number of hypotheses})}{n}}$$

8.3 Infinite Hypothesis Class

When $\mathcal{H}$ is infinite, we attempt to approximate the $\mathcal{H}$ space with a finite number of points in the $\mathcal{H}$ space. Similar discussions have exist in the name of *rate-distortion theory* in the context of information theory.

This section and the next lecture will discuss in the following order: The *covering number* of $\mathcal{H}$ quantifies the size of $\mathcal{H}$, which is used in the *discretization theorem* and *Dudley's theorem* to bound the *empirical Rademacher complexity* of $\mathcal{H}$, which is used to obtain a bound for the excess risk, our ultimate goal.

8.3.1 Coverings and Packings

Definition 8.2 *(ε-covering, ε-net). Let $(V, \|\cdot\|)$ be a normed space, and $\Theta \subset V$. Set $\{V_1, \dots, V_N\}$ is an ε-covering of Θ if*

$$\Theta \subset \cup_{i=1}^{N} B(V_i, \varepsilon),$$

or equivalently, $\forall \theta \in \Theta$, $\exists i$ such that $\|\theta - V_i\| \leqslant \varepsilon$.

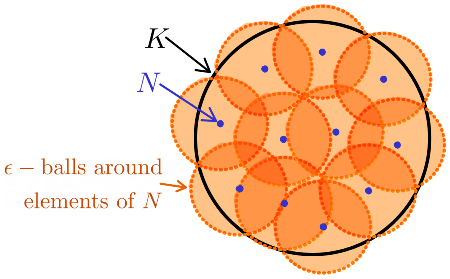

Figure 8.1: Illustration of an ε-covering of K, the area contained within the black circle. The blue points are the N points of $\{V_1, \dots, V_N\}$, and the orange circles are ε-balls around each element of $\{V_1, \dots, V_N\}$. This is a valid ε-covering, as all of K is covered by the orange circles.

Definition 8.3 *(ε-packing). Let $(V, \|\cdot\|)$ be a normed space, and $\Theta \subset V$. Set $\{\theta_1, \dots, \theta_M\}$ is an ε-packing of Θ if $\min_{i \neq j} \|\theta_i - \theta_j\| > \varepsilon$ (notice the inequality is strict), or equivalently, the $\epsilon/2$ balls $B(\theta_1, \epsilon/2), \dots, B(\theta_M, \epsilon/2)$ are pairwise disjoint:*

$$B(\theta_i, \varepsilon/2) \cap B(\theta_j, \varepsilon/2) = \emptyset \quad \forall i, j \in \{1, \dots, M\}, i \neq j.$$

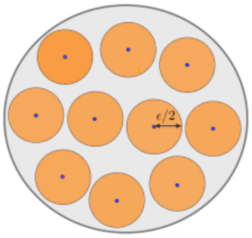

Figure 8.2: Illustration of an ε-packing of the area contained within the black circle. The blue points are the M points of $\{\theta_1, \dots, \theta_M\}$, and the orange circles are $\varepsilon/2$-balls around each element of $\{\theta_1, \dots, \theta_M\}$. This is a valid ε-packing, as the orange circles do not overlap with each other.

As a measure of how large the hypothesis class, one can naturally ask what is the minimal number of ε-balls one needs in order to cover Θ, and what is the maximal number of $\varepsilon/2$-balls one can pack in Θ. Those numbers are defined as covering and packing numbers.

Definition 8.4 *(Covering number). $N(\Theta, \|\cdot\|, \varepsilon) := \min\{n : \exists\varepsilon\text{-covering over } \Theta \text{ of size } n\}$. In words, the covering number N is the size of the smallest possible ε-covering of Θ.*

Definition 8.5 *(Packing number). $M(\Theta, \|\cdot\|, \varepsilon) := \max\{m : \exists\varepsilon\text{-packing of } \Theta \text{ of size } m\}$. In words, the packing number M is the size of the largest possible ε-packing of Θ.*

Lemma 8.6 *If $\{\theta_1, \dots, \theta_m\}$ is a maximal ε-packing of Θ ,i.e., it is not possible to add a point to $\{\theta_1, \dots, \theta_m\}$ and have it remain as an ε-packing, it is also an ε-covering for Θ.*

Proof: $\forall x \in \Theta$, let θ' be the closest member of $\{\theta_1, \dots, \theta_m\}$ from x. We must have $d(x, \theta') \leqslant \varepsilon$, or else $\{\theta_1, \dots, \theta_m\}$ would not be maximal: we could add x to $\{\theta_1, \dots, \theta_m\}$, making a larger ε-packing. Thus, $\{\theta_1, \dots, \theta_m\}$ satisfies the definition of an ε-covering. ■

Lemma 8.7 $M(\Theta, \|\cdot\|, 2\varepsilon) \overset{(1)}{\leqslant} N(\Theta, \|\cdot\|, \varepsilon) \overset{(2)}{\leqslant} M(\Theta, \|\cdot\|, \varepsilon)$.

Proof: First, (2) follows by applying Lemma 8.6 to the largest(maximum) ε-packing : the largest ε-packing of Θ must be maximal, so it is also an ε-covering. So, the smallest ε-covering is at most the size of the largest ε-packing.

To prove (1), we consider the largest 2ε-packing of Θ, denoted P, and the smallest ε-covering of Θ, denoted N. By the definition of an ε-covering, each $p \in P$ must be in an ε-ball of some $x \in N$. Also, by the definition of a 2ε-packing, no $p, q \in P$, $p \neq q$, are in the same ε-ball. This is illustrated in Fig 8.3.

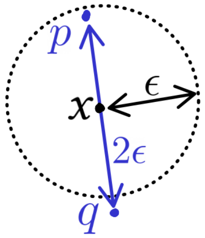

Figure 8.3: Illustration showing that no p, q ($p \neq q$) in a 2ε-packing of Θ can be in the same ε-ball.

Therefore, the map $(p \in P) \rightarrow (x \in N)$ is injective but not necessarily surjective (one-to-one but not necessarily onto), so $|P| \leqslant |N|$. Equivalently, $P(K, 2\varepsilon) \leqslant N(K, \varepsilon)$. ■

8.3.2 Volume argument and dimension dependency

The section discusses the details (that were omitted in class) of the intuition:

> *"regardless of what the norm $\|\cdot\|$ is, a d-dimensional set Θ has covering number $N(\Theta, \|\cdot\|, \varepsilon) = \mathit{\Theta}(1/\varepsilon^d)$".*

Although $\Theta(1/\varepsilon^d)$ is a tight bound when ignoring constants (it is a *big-Theta* result!), it is actually tight only when ε is small; when ε is large, the constant term hiding in Θ is large. Therefore, this so-called *volume argument (volumetric estimate)* used to obtain a uniform bound would not work well in high-dimensional settings (because $\varepsilon \sim d/n$).

When ϵ is large, we can instead use a *a convex hull argument (the empirical method of Maurey)* to compute the covering number of $\mathcal{H}$ (see the last question of the practice midterm for an example).

Lemma 8.8 *If $K \subset \mathbb{R}^d$, $\varepsilon > 0$, under the Euclidean metric,*

$$\frac{\mathrm{vol}(K)}{\mathrm{vol}(\varepsilon B_2^d)} \overset{(1)}{\leqslant} N(K,\varepsilon) \overset{(2)}{\leqslant} M(K,\varepsilon) \overset{(3)}{\leqslant} \frac{\mathrm{vol}\left(K + \frac{\varepsilon}{2}B_2^d\right)}{\mathrm{vol}\left(\frac{\varepsilon}{2}B_2^d\right)}.$$

In the right-hand expression, $K + \frac{\varepsilon}{2}B_2^d$ is known as a Minkowski sum, and it corresponds to adding an $\frac{\varepsilon}{2}$-ball around each point in K.

Proof: We must prove each of the three inequalities above.

(1): By the definition of an ε-covering, all of K is covered by ε-balls around each point in the net. So, $\mathrm{vol}(K) \leqslant \mathrm{vol}\left(\varepsilon B_2^d\right) N(K,\varepsilon)$.

(2): This inequality is directly stated by Lemma 8.7.

(3): For ε-packing P, $\frac{\varepsilon}{2}$-balls around each $p \in P$ must be disjoint. They, however, may extend beyond the bounds of K; instead, they are contained within $K + \frac{\varepsilon}{2}B_2^d$. So, $|P|\mathrm{vol}\left(\frac{\varepsilon}{2}B_2^d\right) \leqslant \mathrm{vol}(K + \frac{\varepsilon}{2}B_2^d)$. ■

Corollary 8.9 $\left(\frac{1}{\varepsilon}\right)^d \leqslant N(B_2^d,\varepsilon) \leqslant \left(1+\frac{2}{\varepsilon}\right)^d$.

Proof: We plug in B_2^d for K in Lemma 8.8. The volume of $rB_2^d = Cr^d$, where C is a constant that only depends on d. Also, the Minkowski sum $B_2^d + \frac{\varepsilon}{2}B_2^d$ is equal to $\left(1+\frac{\varepsilon}{2}\right)B_2^d$. So, the statement of Lemma 8.8 is

$$\frac{\mathrm{vol}(B_2^d)}{\mathrm{vol}(\varepsilon B_2^d)} = \left(\frac{1}{\varepsilon}\right)^d \leqslant N(B_2^d,\varepsilon) \leqslant \frac{\mathrm{vol}\left(\left(1+\frac{\varepsilon}{2}\right)B_2^d\right)}{\mathrm{vol}\left(\frac{\varepsilon}{2}B_2^d\right)} = \frac{C\left(1+\frac{\varepsilon}{2}\right)^d}{C\left(\frac{\varepsilon}{2}\right)^d} = \left(1+\frac{2}{\varepsilon}\right)^d.$$

■

Remark 1 *For small ε, the lower bound of Corollary 8.9 is approximately a factor of 2^d away from the upper bound, so the upper bound is relatively tight.*

Remark 2 *We have not discussed here how to construct good ε-coverings, i.e., those that are close to minimal, as we do not need explicit constructions for our purposes. Construction of optimal coverings is an area of ongoing research—unfortunately, random coverings have logarithmic dependence in d, which is far from the bound in Corollary 8.9.*

Remark 3 *The proofs in this section show that $N(K,\epsilon) = \Theta(1/\varepsilon^d)$ for the Euclidean metric. Since norms on finite-dimensional real vector space are equivalent, this result holds regardless of the norm we are using.*

8.3.3 The discretization theorem

In the previous section, we have computed the covering number of $\mathcal{H}$. The discretization theorem provides a connection between the *empirical Rademacher complexity* $\hat{R}(\mathcal{H})$ (which will be introduced in the next lecture) of a hypothesis class $\mathcal{H}$ and its covering number.

Theorem 8.10 *(Discretization theorem)*

$$\hat{R}(\mathcal{H}) \leqslant \inf_{\varepsilon > 0} \left(\varepsilon + \sqrt{\frac{2 \log N\left(\mathcal{H}, L_2\left(\hat{P}_n\right), \varepsilon\right)}{n}} \right)$$

where the $L_2(\hat{P}_n)$ metric is defined as

$$L_2(\hat{P}_n)(f, f') = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(z_i) - f'(z_i))^2}$$

for empirical distribution $\hat{P}_n$ and the n data points $z_1, \ldots, z_n$ of $\hat{P}_n$.

8.3.4 Chaining and Dudley's Theorem

Can we improve the bound from the discretization theorem for the empirical Rademacher complexity $\hat{R}(\mathcal{H})$? Using the idea of chaining by Michel Talagrand (the 2024 Abel Prize laureate!), Dudley's theorem states a stronger bound by constructing a chained ϵ-covering scheme. We will discuss more about this theorem and chaining idea in the next lecture.

Theorem 8.11 *(Dudley's theorem)*

$$\hat{R}(\mathcal{H}) \leqslant \int_0^\infty 12\sqrt{\frac{2 \log N\left(\mathcal{H}, L_2\left(\hat{P}_n\right), \varepsilon\right)}{n}} d\varepsilon$$

An interesting remark (not discussed in class) is that while Theorem 8.10 assumes that every function in $\mathcal{H}$ to be bounded by $[-1, 1]$, Theorem 8.11 does not require for the functions to be bounded (see Section 4.6.1 in [Ref-1] for details). Depending on the context, this assumption may not be a big issue. For example, in the context of 0-1 loss functions, the boundness assumption for Theorem 8.10 holds.

Why do we care to bound $\hat{R}(\mathcal{H})$? In the next lecture, we will discuss a theorem that connects $\hat{R}(\mathcal{H})$ and the uniform bound, which will give a bound for the excess risk.