5 Covering with ε-nets

The main objective in this lecture is to relax the strong and impractical assumption of Theorem 18, namely the finite function class condition. This assumption is valid only if the practitioner is allowed to choose among a finite number of functions.

In the majority of machine learning methodology, we train our models (weights) by assuming a parametric structure on the function class and the parameter space is infinitely rich (more like uncountably rich).

Model Setup: Suppose that we have a family of functions parametrized in the following sense

$$\mathcal{F} = \{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$$

and a loss function $\ell((x, y), f_\theta) \coloneqq \ell((x, y), \theta)$. In the sequel, we assume that the parameter space is a d-dimensional ball with radius R, i.e. $\Theta = \{\theta : ||\theta|| \leq R\}$, while the loss function ℓ is bounded by B and is L-Lipschitz continuous in θ. It is worth noting that Θ can be any set that is compact, which can be confined inside a ball with some radius, so all the arguments we will soon be making still are valid in general.

Definition 20 (Lipschitz continuous)**.** *A function f is L-Lipschitz continuous if*

$$\forall \theta, \theta', \quad |f(\theta) - f(\theta')| \leq L\|\theta - \theta'\|.$$

Functions that satisfy this condition have stable fluctuations, i.e., if θ and θ' are points that are close to each other, the function values f evaluated at θ and θ' should also be close to each other. For differentiable functions, the above assumption is equivalent to having a uniformly bounded gradient $\|\nabla f(\theta)\| \leq L$. We also notice that it enforces function to have at most linear growth, i.e., let $\theta' = 0$. This rules out the options such as quadratic functions like $f(\theta) = \theta^2$. Functions that are strongly convex cannot be Lipschitz continuous.

5.1 ε-covers of sets in $\mathbb{R}^d$

Remember in the proof of Theorem 18, we have used a union bound over the finite set of functions. This is the main obstacle in our new setup where we have an uncountable set of parameter space Θ. We simply cannot apply union bound over an uncountable sets! But what we can do is to discretize this uncountable set in a way so that it is a *good* representation of the original set, but we can apply union bound. We introduce the following notion of set covers for this task.

Definition 21 (ε-Net)**.** *For $\varepsilon > 0$, $\mathcal{N}_\varepsilon$ is an ε-net (or an ε-cover) over the set $\Theta \subseteq \mathbb{R}^d$ if for all $\theta \in \Theta$, there exists $\theta' \in \mathcal{N}_\varepsilon$ such that $\|\theta - \theta'\| \leq \varepsilon$. That is,*

$$\forall \theta \in \Theta, \ \exists \theta' \in \mathcal{N}_\varepsilon \quad \text{such that} \quad \|\theta - \theta'\| \leq \varepsilon.$$

The size of the ε-net with smallest size $|\mathcal{N}_\varepsilon|$ is called the covering number.

In our applications, we are ideally looking for ε-nets over our parameter space Θ, that have small number of points. But it is worth noting that we are not looking for the optimal ε-net. The following examples will demonstrate the level of optimality we require for our purposes.

Example: Suppose $\Theta = [0, 1]$. To find the ε-net of Θ, one can divide the interval into shorter intervals, each with length 2ε. By defining the set N_ϵ to include all the endpoints of each small interval, for any point θ in the $[0, 1]$ interval, we can always find a point $\theta' \in \mathcal{N}_\varepsilon$ such that $|\theta - \theta'| \leq \varepsilon$. This is demonstrated in Figure 1, and it gives us a valid $\mathcal{N}_\varepsilon$ with $|\mathcal{N}_\varepsilon| = \frac{1}{2\varepsilon} + 1$.

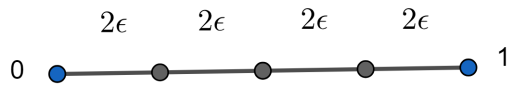

Figure 1: ε-Net for $\Theta = [0, 1]$

Example: Suppose $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\| \leq R\}$. We can do something similar by dividing the ball into grids of size a. Since the number of points required in 1 dimension is $\frac{2R}{a} + 1$, the total number of points required in the d-dimensional space is $(\frac{2R}{a} + 1)^d$. Within each d-dimensional cube of edge length a, the largest distance between the interior points and the vertices comes from the center of the cube, which is $\frac{a\sqrt{d}}{2}$. This is demonstrated in Figure 2. Therefore, to guarantee a full cover of all the points in the ball, the largest grid size should satisfy $\varepsilon = \frac{a\sqrt{d}}{2}$. This leads to the upper bound of the size of an ϵ-net for θ, which is

$$|\mathcal{N}_\varepsilon| \leq \left(\frac{R\sqrt{d}}{\varepsilon} + 1\right)^d \leq \left(\frac{2R\sqrt{d}}{\varepsilon}\right)^d. \tag{5.1}$$

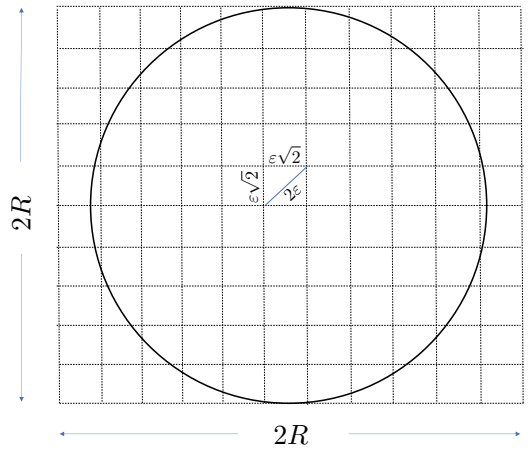

Figure 2: ϵ-Net for $\Theta = \{\theta \in \mathbb{R}^2 : \|\theta\| \leq R\}$

.

The above dependence $\mathcal{O}(d^d)$ is not looking good. But we will next see that the exponential decay in Hoeffding's inequality will be able to compensate for this.

5.2 Generalization for parametrized function classes

We state the following generalization bound for parametrized function classes.

Theorem 22 (Generalization by covering)**.** *Assume that the loss function is bounded by B and L-Lipschitz continuous in its second argument θ. For the parametric function class $\mathcal{F} = \{f_\theta : \|\theta\| \leq R\}$, and the corresponding empirical and population risk minimizers*

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\Theta} \hat{R}(\theta) \quad \textit{and} \quad \theta_* = \operatorname*{argmin}_{\theta\in\Theta} R(\theta),$$

we have with probability at least $1 - 2e^{-d/2}$

$$R(\hat{\theta}) - R(\theta_*) \le c\sqrt{\frac{d\log(n)}{n}} \quad \textit{where} \quad c = 2(B \vee 8RL),$$

whenever $n \ge 16$.

Remark. We make a few remarks before proving the above theorem.

- Convergence rate is $\sqrt{\frac{d\log(n)}{n}}$. This is slower than the previous result by a factor of $\log(n)$, which is due to the covering argument we are about to make.
- Note that the function class is parametrized over a ball of radius R. This is not at all needed and our proof would still follow for any compact set Θ by simply replacing R with $\mathrm{diam}(\Theta)/2$.
- The above probability is decaying exponentially fast with dimension. The constants are arbitrary and can be improved with a more careful treatment.

Proof. The proof will be similar to the finite function class case, with an additional step where we discretize the uncountably rich parameter space Θ.

1. **Concentration**: Since loss is bounded by B, for a non-random θ, by the Hoeffding's inequality applied on $\hat{R}(\theta) - R(\theta)$, we obtain

$$\mathbb{P}\big(|\hat{R}(\theta) - R(\theta)| \ge \epsilon/4\big) \le 2 \cdot \exp\left\{ -\frac{n\epsilon^2}{8B^2} \right\}.$$

2. **Discretization**: In order to apply union bound, we first discretize our uncountable parameter space using an ε-net argument. Before we introduce the ε-net, we first derive a few useful inequalities using the Lipschitz continuity of loss. By the definition of $R(\theta)$ and $\hat{R}(\theta)$, if $l((x,y),\theta)$ is L-Lipschitz, then both $R(\theta)$ and $\hat{R}(\theta)$ would also be L-Lipschitz.

   First, we notice that since ℓ is L-Lipschitz, R and $\hat{R}$ are both L-Lipschitz continuous. Next, by the triangle inequality,

$$\begin{aligned}
|\hat{R}(\theta) - R(\theta)| =& |\hat{R}(\theta') - R(\theta') + \hat{R}(\theta) - \hat{R}(\theta') - R(\theta) + R(\theta')| \\
\le& |\hat{R}(\theta') - R(\theta')| + |\hat{R}(\theta) - \hat{R}(\theta')| + |R(\theta) - R(\theta')| \\
\le& |\hat{R}(\theta') - R(\theta')| + 2L\|\theta - \theta'\|.
\end{aligned}$$

   Now, let $\mathcal{N}_\Delta$ be a Δ-net over $\Theta \subseteq \mathbb{R}^d$. For any $\theta \in \Theta$, there exists $\theta' \in \mathcal{N}_\Delta$ such that $\|\theta - \theta'\| \le \Delta$. Using this and together with the previous inequality, we obtain that

$$|\hat{R}(\theta) - R(\theta)| \le |\hat{R}(\theta') - R(\theta')| + 2L\Delta.$$

   By first taking maximum over the Δ-net over the right hand side, and next taking a supremum on the left hand side, we obtain

$$\sup_{\theta\in\Theta} |\hat{R}(\theta) - R(\theta)| \le \max_{\theta\in\mathcal{N}_\Delta} |\hat{R}(\theta) - R(\theta)| + 2L\Delta.$$

3. **Union bound:** Now that we discretized the parameter space, we can apply the union bound. Using the previous display, we write

$$\begin{aligned}\mathbb{P}\bigg(\sup_{\theta\in\Theta}|\hat{R}(\theta)-R(\theta)|\geq\epsilon/2\bigg)\leq&\mathbb{P}\bigg(\max_{\theta\in\mathcal{N}_\Delta}|\hat{R}(\theta)-R(\theta)|\geq\epsilon/2-2L\Delta\bigg)\\ \leq&\mathbb{P}\bigg(\max_{\theta\in\mathcal{N}_\Delta}|\hat{R}(\theta)-R(\theta)|\geq\epsilon/4\bigg)=(*)\end{aligned}$$

where we let $\Delta=\epsilon/8L$ in the last step. By the union bound, we get

$$\begin{aligned}(*)=\mathbb{P}\left(\bigcup_{\theta\in\mathcal{N}_\Delta}\{|\hat{R}(\theta)-R(\theta)|\geq\epsilon/4\}\right)\leq&\sum_{\theta\in\mathcal{N}_\Delta}\mathbb{P}\Big(|\hat{R}(\theta')-R(\theta')|\geq\epsilon/4\Big),\\ \leq&2|\mathcal{N}_\Delta|\exp\left\{-\frac{n\epsilon^2}{8B^2}\right\}.\end{aligned}$$

The bound on the right hand side is quite explicit and depends on the covering number of the parameter space Θ. But we already have a bound on this covering number from the previous example as given in (5.1), that is, $|\mathcal{N}_\Delta|\leq(\frac{2R\sqrt{d}}{\Delta})^d$ where $\Delta=\epsilon/8L$. Hence, the above inequality suggests that

$$\mathbb{P}\bigg(\sup_{\theta\in\Theta}|\hat{R}(\theta)-R(\theta)|\geq\epsilon/2\bigg)\leq2\left(\frac{16RL\sqrt{d}}{\epsilon}\right)^d\exp\bigg\{-\frac{n\epsilon^2}{8B^2}\bigg\}=(**)$$

By moving all terms inside the exponent, we get

$$(**)\leq2\exp\bigg\{-\frac{n\epsilon^2}{8B^2}+d\log(16RL\sqrt{d})+d\log(\epsilon^{-1})\bigg\}.$$

At this point, we are ready to identify the convergence rate. We start by trying out $\epsilon=c\sqrt{\frac{d}{n}}$, which yields

$$(**)\leq2\exp\bigg\{-\frac{d}{8B^2}+d\log(16RL\sqrt{d})-d\log(c\sqrt{d})+d\log(c\sqrt{n})\bigg\}.$$

Notice that the second and third terms can cancel each other with a right choice of c, but the first and last terms cannot, as one is decaying with d and other is growing with $d\log(n)$. In fact, any rate slower that $\sqrt{n}$ would work here. But we can also get away with only losing a log factor.

By choosing $\epsilon=c\sqrt{\frac{d\log(n)}{n}}$, we have

$$\begin{aligned}(**)\leq&2\exp\bigg\{-\frac{c^2d\log(n)}{8B^2}+d\log(16RL\sqrt{d})-\frac{d}{2}\log\log(n)+\frac{d}{2}\log(n)-d\log(c\sqrt{d})\bigg\},\\ \leq&2\exp\{-d/2\},\end{aligned}$$

where we let $c=2(B\vee8RL)$ and $n\geq16$.

4. **Uniform convergence $\implies$ generalization:** The last step is to convert the bound on the empirical process to a bound on the excess risk. We have

$$\mathbb{P}\big(R(\hat{\theta}) - R(\theta_*) \geq \epsilon\big) \leq \mathbb{P}\left(\sup_{\theta \in \Theta} |\hat{R}(\theta) - R(\theta)| \geq \epsilon/2\right),$$

combining this with the previous result, we obtain

$$\mathbb{P}\left(R(\hat{\theta}) - R(\theta_*) \geq c\sqrt{\frac{d\log(n)}{n}}\right) \leq 2\exp\{-d/2\},$$

whenever $c = 2(B \vee 8RL)$ and $n \geq 16$.

□