# STA 250: Theoretical Foundations for Machine Learning

## Lecture 3: Rademacher Complexity

Dogyoon Song

Spring 2025, UC Davis

# Last time...

Asymptotic analysis: $R(\hat{\theta}) - R^* \leq \frac{c}{n} + o\left(\frac{1}{n}\right)$

Non-asymptotic analysis: Generalization bound via uniform convergence

- Uniform convergence: $\Pr\left(\sup_{\theta \in \Theta} \left|\hat{R}(\theta) - R(\theta)\right| \leq \epsilon\right) \geq 1 - \delta$
- If $|\Theta| < \infty$ and $\ell\big(f_\theta(x), y\big) \in [0, B]$, then with probability at least $1 - \delta$,

$$\sup_{\theta \in \Theta} \left|\hat{R}(\theta) - \hat{R}(\theta)\right| \leq \underbrace{B\sqrt{\frac{\log(2|\Theta|)}{2n}}}_{\text{overhead for uniform control}} + B\sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)}$$

- If $\Theta$ is compact, $\ell\big(f_\theta(x), y\big) \in [0, B]$, and $\ell$ is L-Lipschitz w.r.t. $\theta$, then for any $\epsilon > 0$,

$$\sup_{\theta \in \Theta} \left|\hat{R}(\theta) - \hat{R}(\theta)\right| \leq 2L\epsilon + B\sqrt{\frac{\log(2N(\Theta, \epsilon))}{2n}} + B\sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)}$$

**Motivating question:** Is the cardinality $|\Theta|$ an appropriate notion of complexity?

# Agenda

- Rademacher complexity
- Generalization bound based on Rademacher complexity
- Examples

# Rademacher complexity

## Definition

Let $n \in \mathbb{N}$. The **Rademacher complexity** of a function class $\mathcal{G} = \{g : \mathcal{Z} \to \mathbb{R}\}$ is

$$\mathrm{Rad}_n(\mathcal{G}) := \mathbb{E}_{\varepsilon,\mathcal{D}_n}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i g(z_i)\right]$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ is a Rademacher random vector[a] and $\mathcal{D}_n = \{z_1, \ldots, z_n\} \sim \mu$ is an i.i.d. sample drawn from $\mathcal{Z}$

[a] $\varepsilon_i$ being i.i.d. Rademacher random variables; $\varepsilon_i = \pm 1$ with probability $\frac{1}{2}$ each

- Geometric interpretation as a width $\to$ Verify the properties in [Bac24, Exercise 4.9]
- Connection to generalization:
  - $z = (x, y)$
  - $g(z) = \ell(f(x), y)$

# Relating Rademacher complexity to uniform deviation

Rademacher complexity yields an upper bound on uniform deviation

## Symmetrization

For any $\mathcal{G}$, $\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n} g(z_i) - \mathbb{E}[g(z)]\right\}\right] \leq 2\mathrm{Rad}_n(\mathcal{G})$

**Proof**[1]**.** Let $\mathcal{D}' = \{z_1', \ldots, z_n'\}$ be an independent copy of data $\mathcal{D}$.

$$
\begin{aligned}
\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n} g(z_i) - \mathbb{E}[g(z)]\right\}\right] &= \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[g(z_i) - g(z_i') \,\middle|\, \mathcal{D}_n\right]\right\}\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}\Big(g(z_i) - g(z_i')\Big)\right\} \,\middle|\, \mathcal{D}_n\right]\right] \\
&= \mathbb{E}_{\mathcal{D},\mathcal{D}'}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}\Big(g(z_i) - g(z_i')\Big)\right\}\right]
\end{aligned}
$$

[1]Similarly, we can show $\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\mathbb{E}[g(z)] - \frac{1}{n}\sum_{i=1}^{n} g(z_i)\right\}\right] \leq 2\mathrm{Rad}_n(\mathcal{G})$

# Proof of symmetrization (cont'd)

By the symmetry in the laws of $\varepsilon_i$ and of $g(z_i) - g(z_i')$,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D},\mathcal{D}'}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}\Big(g(z_i)-g(z_i')\Big)\right\}\right] &= \mathbb{E}_{\mathcal{D},\mathcal{D}',\varepsilon}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Big(g(z_i)-g(z_i')\Big)\right\}\right] \\
&\leq \mathbb{E}_{\mathcal{D},\varepsilon}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(z_i)\right\}\right] \\
&\quad + \mathbb{E}_{\mathcal{D}',\varepsilon}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}-\varepsilon_i g(z_i')\right\}\right] \\
&= 2\mathrm{Rad}_n(\mathcal{G})
\end{aligned}$$

# Resulting high-probability bound

Rademacher complexity provides a control on the expectation of uniform deviation

Can we obtain high-probability bounds?

- Apply concentration inequalities

If $g(z) \in [0, B]$ for all $(g, z) \in \mathcal{G} \times \mathcal{Z}$, then with probability at least $1 - \delta$,

$$\sup_{g \in \mathcal{G}} \left[ \frac{1}{n} \sum_{i=1}^{n} g(z_i) - \mathbb{E}[g(z)] \right] \leq 2\mathrm{Rad}_n(\mathcal{G}) + B\sqrt{\frac{\log(2/\delta)}{2n}}$$

Note that $\mathrm{Rad}_n(\mathcal{G})$ is averaged over all possible $\mathcal{D}_n$

# Empirical Rademacher complexity

An empirical version can be defined, which does not take expectation with respect to $\mathcal{D}_n$:

$$\widehat{\mathrm{Rad}}_{\mathcal{D}_n}(\mathcal{G}) := \mathbb{E}_\varepsilon \left[ \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{z_i \in \mathcal{D}_n} \varepsilon_i g(z_i) \right]$$

Note that $\widehat{\mathrm{Rad}}_{\mathcal{D}_n}(\mathcal{G})$ is dependent on both function class $\mathcal{G}$ and data $\mathcal{D}_n$

As the name suggests, $\mathbb{E}_{\mathcal{D}_n}\left[\widehat{\mathrm{Rad}}_{\mathcal{D}_n}(\mathcal{G})\right] = \mathrm{Rad}_n(\mathcal{G})$

If $g(z) \in [0, B]$ for all $(g, z) \in \mathcal{G} \times \mathcal{Z}$, then with probability at least $1 - \delta$,

$$\sup_{g \in \mathcal{G}} \left[ \frac{1}{n} \sum_{i=1}^{n} g(z_i) - \mathbb{E}[g(z)] \right] \leq 2\widehat{\mathrm{Rad}}_{\mathcal{D}_n}(\mathcal{G}) + 3B\sqrt{\frac{\log(2/\delta)}{2n}}$$

# Taming Rademacher complexity

**Question:** How to prove an upper bound for Rademacher complexity?

Approach 1: General bounds based on covering number

- For computing $\widehat{\mathrm{Rad}}_{\mathcal{D}}$, we care about $f$ only through the lens of $f(z_1), \ldots, f(z_n)$, where $\mathcal{D} = \{z_1, \ldots, z_n\}$
- $\epsilon$-net and chaining

Approach 2: Tailored bounds to specific settings

- Linear models
- 2-layer neural networks (Homework)

# Finite function class

**Proposition (Massart's lemma)**

*Fix $\mathcal{D} = (z_1, \ldots, z_n)$, and let $\mathcal{G}_{\mathcal{D}} := \{(g(z_1), \ldots, g(z_n)) : g \in \mathcal{G}\}$. If $\frac{1}{n}\|v\|_2^2 \leq B^2$ for all $v \in \mathcal{G}_{\mathcal{D}}$, then*

$$\widehat{\mathrm{Rad}}_{\mathcal{D}}(\mathcal{G}) \leq B\sqrt{\frac{2\log|\mathcal{G}_{\mathcal{D}}|}{n}}.$$

Using Massart's lemma, we can also bound the Rademacher complexity in terms of $\mathcal{G}$:

$$\tfrac{1}{n}\textstyle\sum_{i=1}^n g_j(z_i)^2 \leq B^2 \text{ almost surely for all } g \in \mathcal{G} \implies \mathrm{Rad}_n(\mathcal{G}) \leq B\sqrt{\tfrac{2\log|\mathcal{G}|}{n}}$$

Therefore, with probability at least $1-\delta$,

$$\sup_{g\in\mathcal{G}}\left[\frac{1}{n}\sum_{i=1}^n g(z_i) - \mathbb{E}[g(z)]\right] \leq 2\mathrm{Rad}_n(\mathcal{G}) + B\sqrt{\frac{\log(2/\delta)}{2n}} \leq 2B\sqrt{\frac{2\log(|\mathcal{G}|)}{n}} + B\sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}$$

# General bound using $\epsilon$-net

When $\mathcal{G}_\mathcal{D}$ is infinite, we may discretize $\mathcal{G}_\mathcal{D}$ w.r.t. $d(v, v') = \frac{1}{\sqrt{n}}\|v - v'\|_2$

**Proposition**

*Let $\mathcal{G}$ be a family of functions from $\mathcal{Z}$ to $[-1, 1]$ and $\mathcal{D} = (z_1, \ldots, z_n)$. Then*

$$\widehat{\mathrm{Rad}}_\mathcal{D}(\mathcal{G}) \leq \inf_{\epsilon > 0} \left( \epsilon + \sqrt{\frac{2 \log N(\mathcal{G}_\mathcal{D}, \epsilon, d)}{n}} \right)$$

We can obtain the following (stronger) result using the chaining argument:

**Theorem (Dudley's theorem)**

*Let $\mathcal{G}$ be a family of functions from $\mathcal{Z}$ to $\mathbb{R}$ and $\mathcal{D} = (z_1, \ldots, z_n)$. Then*

$$\widehat{\mathrm{Rad}}_\mathcal{D}(\mathcal{G}) \leq 12 \int_0^\infty \sqrt{\frac{2 \log N(\mathcal{G}_\mathcal{D}, \epsilon, d)}{n}} d\epsilon$$

# Lipschitz continuous loss

**Proposition (Talagrand's contraction principle)**

*Let $a_i : \Theta \to \mathbb{R}$, $i \in [n]$ and $b : \Theta \to \mathbb{R}$ be arbitrary functions. Let $\varphi_i : \mathbb{R} \to \mathbb{R}$ be a $L$-Lipschitz function for all $i \in [n]$. Then*

$$\mathbb{E}_\varepsilon \left[ \sup_{\theta \in \Theta} \left\{ b(\theta) + \sum_{i=1}^n \varepsilon_i \cdot \varphi_i\big(a_i(\theta)\big) \right\} \right] \leq L \cdot \mathbb{E}_\varepsilon \left[ \sup_{\theta \in \Theta} \left\{ b(\theta) + \sum_{i=1}^n \varepsilon_i \cdot a_i(\theta) \right\} \right]$$

*where $\varepsilon$ is a random vector with independent Rademacher entries.*

Apply this contraction principle to the supervised learning situation, conditioned on $\mathcal{D}_n$:

- Suppose a map that $\varphi : u_i \mapsto \ell(u_i, y_i)$ is $L$-Lipschitz for all $i \in [n]$ a.s.
- Let $\Theta = \{\big(f(x_1), \ldots, f(x_n)\big) : f \in \mathcal{F}\}$
- $a_i(\theta) = \theta_i$, $b = 0$, $\varphi_i(u) = \ell(u, y_i)$

This implies that $\widehat{\mathrm{Rad}}_{\mathcal{D}}(\mathcal{G}) \leq L \cdot \widehat{\mathrm{Rad}}_{\mathcal{D}}(\mathcal{F}) \implies$ Rademacher complexity of the *class of prediction functions* controls the uniform deviations

# Norm-constrained linear predictions

Suppose that $\mathcal{F} = \{f_\theta(x) : \langle \theta, \varphi(x) \rangle, \ \|\theta\| \le D\}$

Letting $\Phi = \begin{bmatrix} \varphi(x_1) & \dots & \varphi(x_n) \end{bmatrix}^\top$, observe that

$$\begin{aligned}
\mathrm{Rad}_n(\mathcal{F}) &= \mathbb{E}\left[ \sup_{\|\theta\| \le D} \left\{ \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \langle \theta, \varphi(x_i) \rangle \right\} \right] \\
&= \mathbb{E}\left[ \sup_{\|\theta\| \le D} \frac{1}{n} \varepsilon^\top \Phi \theta \right] \\
&= \frac{D}{n} \mathbb{E}\left[ \|\Phi^\top \varepsilon\|_* \right]
\end{aligned}$$

where $\|\cdot\|_*$ is the dual norm[2] of $\|\cdot\|$

[2] $\|w\|_* := \sup_{\|v\| \le 1} \langle v, w \rangle$

# Norm-constrained linear predictions: Examples

**Example 1**: Let $\mathcal{F} = \{f_\theta(x) = \langle \theta, \varphi(x) \rangle, \|\theta\|_2 \leq D\}$ and suppose $\mathbb{E}\left[\|\varphi(x_i)\|_2^2\right] \leq R^2$

$$\mathbb{E}\left[\|\Phi^\top \varepsilon\|_2\right] \leq \sqrt{\mathbb{E}\left[\|\Phi^\top \varepsilon\|_2^2\right]} = \sqrt{\mathbb{E}\left[\mathrm{Tr}\left(\Phi^\top \varepsilon\varepsilon^\top \Phi\right)\right]}$$

$$= \sqrt{\mathbb{E}\left[\mathrm{Tr}\left(\Phi^\top \Phi\right)\right]} = \sqrt{\mathbb{E}\left[\sum_{i=1}^n \|\varphi(x_i)\|_2^2\right]} = \sqrt{n} \cdot \sqrt{\mathbb{E}\left[\|\varphi(x_i)\|_2^2\right]}$$

$$\implies \mathrm{Rad}_n(\mathcal{F}) = \frac{D}{n}\mathbb{E}\left[\|\Phi^\top \varepsilon\|_2\right] \leq \frac{RD}{\sqrt{n}}$$

**Example 2**: Let $\mathcal{F} = \{f_\theta(x) = \langle \theta, \varphi(x) \rangle, \|\theta\|_1 \leq D\}$ and suppose $\|\varphi(x_i)\|_\infty \leq R$ a.s.

$$\implies \mathrm{Rad}_n(\mathcal{F}) = \frac{D}{n}\mathbb{E}\left[\|\Phi^\top \varepsilon\|_\infty\right] \leq \frac{RD}{\sqrt{n}}\sqrt{2\log(2d)}$$

**Example 3**: Let $p > 1$ and $q$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Let $\mathcal{F} = \{f_\theta(x) = \langle \theta, \varphi(x) \rangle, \|\theta\|_p \leq D\}$ and suppose $\|\varphi(x_i)\|_q \leq R$ a.s.

$$\implies \mathrm{Rad}_n(\mathcal{F}) = \frac{D}{n}\mathbb{E}\left[\|\Phi^\top \varepsilon\|_\infty\right] \leq \frac{RD}{\sqrt{n}}\frac{1}{\sqrt{p-1}}$$

# References

Francis Bach.
*Learning Theory from First Principles*.
MIT press, 2024.