

## Lecture 4: Uniform law and Rademacher contraction

1 / 16

### Announcements

- Homework 1 out last Friday, due next week Wednesday
- You can ask clarifying questions in moodle
- Link to papers will be posted end of today, project sign-ups in about 2 weeks (still have time after the homework)

2 / 16

# Plans for today

- Recap Rademacher complexity
- Proof of uniform law with symmetrization
- Application of Rademacher complexity: Finite function classes
  - Massart's lemma and its proof
  - Margin bound

3 / 16

## Recap: Uniform tail bound via Rademacher complexity

- Let  $\epsilon_i$  be i.i.d. Rademacher R.V.
- $Z = \{Z_i\}_{i=1}^n$  are training points  $\stackrel{iid}{\sim} \mathbb{P}$

### Definition (Rademacher complexity)

Given a function class  $\mathcal{H}$  and distribution  $\mathbb{P}$  on its domain  $\mathcal{Z}$ , we define the Rademacher complexity as

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{\epsilon, Z} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i h(z_i)$$

### Theorem (Uniform law for the risk, MW Thm 4.10.)

For  $b$ -unif. bounded  $\mathcal{H}$ , with prob. over training data,

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} [\mathbb{E}h - \frac{1}{n} \sum_{i=1}^n h(Z_i)] \geq 2\mathcal{R}_n(\mathcal{H}) + t\right) \leq e^{-\frac{nt^2}{2b^2}}$$

4 / 16

## Recap: Intuition for Rademacher complexity

1. For prediction loss  $\ell$ , how does the empirical Rademacher complexity

$$\tilde{\mathcal{R}}_n(\mathcal{H}) = \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \ell(f(x_i), y_i)$$

$$\text{with } \mathcal{H} = \{h : h(\cdot) = \ell(\cdot; f) \quad \forall f \in \mathcal{F}\}$$

depend on the factors  $\mathcal{F}$ ,  $n$  to control excess risk? What is the connection between R.C. and VC dimension?

- If  $\mathcal{F}$  larger  $\rightarrow \mathcal{H}$  larger  $\rightarrow \tilde{\mathcal{R}}_n(\mathcal{H})$  larger. Intuition through noise-fitting, similar to VC dimension
  - As  $n$  grows, harder to fit  $\rightarrow \tilde{\mathcal{R}}_n(\mathcal{H})$  smaller
2. (New today): How does it depend on the loss function?
- If  $\ell$  varies little with first argument  $f(x)$  (smaller Lipschitz constant)  $\rightarrow$  harder to fit noise  $\rightarrow \tilde{\mathcal{R}}_n(\mathcal{H})$  is smaller (see Rademacher contraction inequality next)

5 / 16

## Empirical Rademacher complexity - notation

Note the empirical Rademacher complexity can be viewed as a measure of how the “size” of the following set of points in  $\mathbb{R}^n$ :

- A fixed  $f \in \mathcal{F}$  defines a labeling from domain  $\mathcal{X} \rightarrow \{-1, +1\}$ . For a given set  $Z^n = \{Z_i = (x_i, y_i)\}_{i=1}^n$ , the function space  $\mathcal{F}$  induces a set in  $\{-1, 1\}^n$  that reads  $\mathcal{F}(Z^n) = \{(f(Z_1), \dots, f(Z_n)) : f \in \mathcal{F}\}$
- We again use notation  $h(z) = \ell(z, f)$  and define

$$\mathcal{H}(Z^n) = \{(\ell(Z_1; f), \dots, \ell(Z_n; f)) : f \in \mathcal{F}\} \subset \mathbb{R}^n$$

Notice that  $|\mathcal{F}(Z^n)| = |\mathcal{H}(Z^n)|$

More generally, we will consider the empirical Rademacher complexity of sets  $\mathbb{T} \subset \mathbb{R}^n$  defined as

$$\tilde{\mathcal{R}}_n(\mathbb{T}) = \mathbb{E} \sup_{\theta \in \mathbb{T}} \sum_{i=1}^n \epsilon_i \theta_i.$$

Note that hence we can write  $\tilde{\mathcal{R}}_n(\mathcal{H}(Z^n))$  for  $\tilde{\mathcal{R}}_n(\mathcal{H})$  (with slight abuse of notation).

6 / 16

## Empirical Rademacher complexity of compositions

In the case of classification, we often have a loss of the form (again, slightly abusing notation)  $\ell(Z_i, f) = \ell(Y_i f(X_i))$  depending on only one argument, and can define  $\tilde{f}(Z_i) = Y_i f(X_i)$ .

The following lemma can connect the empirical Rademacher comp. of a function class  $\tilde{\mathcal{F}}$  to the empirical Rademacher comp. of a specific loss  $\ell : \mathbb{R} \rightarrow \mathbb{R}$  acting on a function class, specifically when  $\mathcal{H} = \ell \circ \tilde{\mathcal{F}}$

First note that for  $\mathbb{T} = \tilde{\mathcal{F}}(Z^n)$  we can write the empirical Rademacher complexity in two ways (abusing notation)

$$\tilde{\mathcal{R}}_n(\ell \circ \tilde{\mathcal{F}}) = \mathbb{E} \sup_{\tilde{f} \in \tilde{\mathcal{F}}} \sum_{i=1}^n \epsilon_i \ell(\tilde{f}(Z_i)) = \mathbb{E} \sup_{\theta \in \mathbb{T}} \sum_{i=1}^n \epsilon_i \ell(\theta_i) = \tilde{\mathcal{R}}_n(\ell \circ \mathbb{T})$$

7 / 16

## Rademacher contraction

The following lemma holds for general losses  $\ell : \mathbb{R}^n \rightarrow \mathbb{R}^n$  (again, abuse of notation) where the loss may differ for each element, with  $\ell(\theta) = (\ell_1(\theta_1), \dots, \ell_n(\theta_n))$  with  $L$ -Lipschitz  $\ell_j : \mathbb{R} \rightarrow \mathbb{R}$ , i.e.

$$|\ell_j(a) - \ell_j(b)| \leq L|a - b| \text{ for all } a, b \in \mathbb{R}.$$

### Lemma (Rademacher contraction, SS Lemma 26.9)

For any  $\mathbb{T} \subset \mathbb{R}^n$  and  $\ell : \mathbb{R}^n \rightarrow \mathbb{R}^n$  with univariate  $L$ -Lipschitz functions it holds that

$$\tilde{\mathcal{R}}_n(\ell \circ \mathbb{T}) \leq L \tilde{\mathcal{R}}_n(\mathbb{T})$$

8 / 16

## Proof ingredients

Let  $\epsilon$  be the vector of  $n$  i.i.d. Rademacher r.v. and define the shorthand  $\epsilon_{2:n} = (\epsilon_2, \dots, \epsilon_n)$  and same for  $\theta$ .

The following holds for all  $n$

- Key 1: de-symmetrize using the tower property: For any  $g$  we have  $\mathbb{E}_\epsilon g(\epsilon) = \mathbb{E}_{\epsilon_1} [\mathbb{E}[g(\epsilon)|\epsilon_1]] = \frac{1}{2}\mathbb{E}[g(\epsilon)|\epsilon_1 = 1] + \frac{1}{2}\mathbb{E}[g(\epsilon)|\epsilon_1 = -1]$
- Key 2: Lipschitz property  $\ell_i(\theta_i) - \ell_i(\tilde{\theta}_i) \leq L|\theta_i - \tilde{\theta}_i|$  for all  $i$
- Key 3: For each  $\epsilon$  we can define  $v(\theta_{2:n}) = \sum_{i=2}^n \epsilon_i \ell_i(\theta_i)$ . Then

$$\begin{aligned} & \sup_{\theta, \tilde{\theta} \in \mathbb{T}} |\theta_1 - \tilde{\theta}_1| + v(\theta_{2:n}) + v(\tilde{\theta}_{2:n}) = \\ & \sup_{s \in \{\pm 1\}} \sup_{\theta, \tilde{\theta} \in \mathbb{T}} s\theta_1 - s\tilde{\theta}_1 + v(\theta_{2:n}) + v(\tilde{\theta}_{2:n}) = \\ & \sup_{s \in \{\pm 1\}} \left\{ \sup_{\theta \in \mathbb{T}} s\theta_1 + h(\theta_{2:n}) + \sup_{\tilde{\theta} \in \mathbb{T}} -s\tilde{\theta}_1 + v(\tilde{\theta}_{2:n}) \right\} = \\ & \sup_{\theta \in \mathbb{T}} \theta_1 + v(\theta_{2:n}) + \sup_{\tilde{\theta} \in \mathbb{T}} -\tilde{\theta}_1 + v(\tilde{\theta}_{2:n}) \end{aligned}$$

where the last equality holds due to equality for the terms with  $s = +1$  or  $s = -1$

9 / 16

## R.C. contraction proof

$$\begin{aligned} n\tilde{\mathcal{R}}_n(\ell \circ \mathbb{T}) &= \mathbb{E}_\epsilon \sup_{\theta \in \mathbb{T}} \sum_{i=1}^n \epsilon_i \ell_i(\theta_i) \\ &\stackrel{1.}{=} \frac{1}{2} \left[ \mathbb{E}_{\epsilon_{2:n}} \sup_{\theta \in \mathbb{T}} \ell_1(\theta_1) + \sum_{i=2}^n \epsilon_i \ell_i(\theta_i) + \sup_{\tilde{\theta} \in \mathbb{T}} -\ell_1(\tilde{\theta}_1) + \sum_{i=2}^n \epsilon_i \ell_i(\tilde{\theta}_i) \right] \\ &= \frac{1}{2} \left[ \mathbb{E}_{\epsilon_{2:n}} \sup_{\theta, \tilde{\theta} \in \mathbb{T}} \ell_1(\theta_1) - \ell_1(\tilde{\theta}_1) + \sum_{i=2}^n \epsilon_i \ell_i(\theta_i) + \sum_{i=2}^n \epsilon_i \ell_i(\tilde{\theta}_i) \right] \\ &\stackrel{2.}{\leq} \frac{1}{2} \left[ \mathbb{E}_{\epsilon_{2:n}} \sup_{\theta, \tilde{\theta} \in \mathbb{T}} L|\theta_1 - \tilde{\theta}_1| + \sum_{i=2}^n \epsilon_i \ell_i(\theta_i) + \sum_{i=2}^n \epsilon_i \ell_i(\tilde{\theta}_i) \right] \\ &\stackrel{3.}{=} \frac{1}{2} \left[ \mathbb{E}_{\epsilon_{2:n}} \sup_{\theta \in \mathbb{T}} L\theta_1 + \sum_{i=2}^n \epsilon_i \ell_i(\theta_i) + \sup_{\tilde{\theta} \in \mathbb{T}} (-L\tilde{\theta}_1) + \sum_{i=2}^n \epsilon_i \ell_i(\tilde{\theta}_i) \right] \\ &\stackrel{1.}{=} \mathbb{E}_\epsilon \sup_{\theta \in \mathbb{T}} L\epsilon_1\theta_1 + \sum_{i=2}^n \epsilon_i \ell_i(\theta_i) \end{aligned}$$

Use the same argument for the RHS inductively on each coordinate.  $\square$  10 / 16

## Proof of the uniform law

First recall that we already established using McDiarmid / Azuma-Hoeffding:

### Theorem (Uniform tail bound)

For  $b$ -unif. bounded  $\ell$ , it holds that

$$\mathbb{P}(\sup_{f \in \mathcal{F}} R(f) - R_n(f) \geq \mathbb{E}[\sup_{f \in \mathcal{F}} R(f) - R_n(f)] + t) \leq e^{-\frac{nt^2}{2b^2}}$$

where the probability is over the training data.

In particular, by the uniform tail bound, if we can prove that  $\mathbb{E}[\sup_{f \in \mathcal{F}} R(f) - R_n(f)] \leq 2\mathcal{R}_n(\mathcal{H})$  then it immediately follows that

$$\begin{aligned} & \mathbb{P}(\sup_{h \in \mathcal{H}} \mathbb{E}h(Z) - \frac{1}{n} \sum_{i=1}^n h(Z_i) \geq 2\mathcal{R}_n(\mathcal{H}) + t) \\ & \leq \mathbb{P}(\sup_{f \in \mathcal{F}} R(f) - R_n(f) \geq \mathbb{E}[\sup_{f \in \mathcal{F}} R(f) - R_n(f)] + t) \leq e^{-\frac{nt^2}{2b^2}} \end{aligned}$$

This proof step is called symmetrization

11 / 16

## Proof of uniform law - Step II: Symmetrization

- (i) For any  $H$ ,  $\sup_H \mathbb{E}H(Z) \leq \mathbb{E} \sup_H H(Z)$  (Exercise)
- (ii)  $h(Z_i) - h(\tilde{Z}_i)$  is symmetric  $\rightarrow$  multiplying by  $\epsilon_i$  preserves distr.

$$\begin{aligned} \mathbb{E}_Z g_n(Z) &= \mathbb{E}_Z \sup_{h \in \mathcal{H}} \mathbb{E}h - \frac{1}{n} \sum_i h(Z_i) \\ &= \mathbb{E}_Z \sup_{h \in \mathcal{H}} \mathbb{E}_{\tilde{Z}} \frac{1}{n} \sum_{i=1}^n h(\tilde{Z}_i) - \frac{1}{n} \sum_{i=1}^n h(Z_i) \\ &\stackrel{(i)}{\leq} \mathbb{E}_{Z, \tilde{Z}} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n [h(Z_i) - h(\tilde{Z}_i)] \\ &\stackrel{(ii)}{=} \mathbb{E}_{Z, \tilde{Z}, \epsilon} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i [h(Z_i) - h(\tilde{Z}_i)] \\ &\leq 2\mathbb{E}_{Z, \epsilon} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i h(Z_i) =: 2\mathcal{R}_n(\mathcal{H}) \square \end{aligned}$$

- Tight:  $\frac{\mathcal{R}_n(\mathcal{H})}{2} \leq \mathbb{E} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_i h - \mathbb{E}h \leq 2\mathcal{R}_n(\mathcal{H})$  (MW Prop 4.11.)

12 / 16

## Caveats of the uniform law

- Requires boundedness of  $\ell$  (for bounded differences)
  - for regression you also bound suprema of empirical processes, can use Gaussian complexity and Lipschitz-of-Gaussians rule (see MW 3)
  - or argue that  $\ell$  bounded with high probability, cause  $X$  and hence  $f(X)$  bounded for continuous  $f$
- Super loose bound  $\rightarrow \mathcal{F}$  needs to be algorithm / data dependent
  - we will see for regularized optimizers
  - structural risk minimization
- in the second half of lectures we'll discuss a different way to bound the excess risk for regression  $\rightarrow$  however even there, we will control suprema of empirical processes

We now use the uniform law to obtain simple excess risk bounds for classification

13 / 16

## Classification setup & Massart's Lemma

For of all, for any set of labeling functions  $\mathcal{H}$ , we have:

### Lemma (Massart)

For  $n$  points  $Z^n := \{Z_1, \dots, Z_n\}$ , let all  $h : \mathcal{Z} \rightarrow \{0, 1\}$  and  $\mathcal{H}(Z^n) := \{(h(Z_1), \dots, h(Z_n)) : h \in \mathcal{H}\}$  with cardinality  $|\mathcal{H}(Z^n)|$ .

$$\tilde{\mathcal{R}}_n(\mathcal{H}(Z^n)) := \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i h(Z_i) \leq \sqrt{\frac{2 \log |\mathcal{H}(Z^n)|}{n}}$$

We now look at specific labeling functions

- Labels  $y \in \{-1, +1\}$  and given  $f$ , we predict the label of some  $x$  using  $\hat{y} = \text{sign}(f(x))$
- Evaluation metric:  $\ell((x, y); f) = \mathbb{1}_{\{yf(x) < 0\}}$  and hence population risk:  $R(f) = \mathbb{E}\ell((x, y); f) = \mathbb{P}(y \neq \text{sign}(f(x)))$
- We again use notation  $h(z) = \ell(z, f)$  and define

$$\mathcal{H}(Z^n) = \{(\ell(Z_1; f), \dots, \ell(Z_n; f)) : f \in \mathcal{F}\} \subset \mathbb{R}^n$$

Notice that  $|\mathcal{F}(Z^n)| = |\mathcal{H}(Z^n)|$

14 / 16

## VC bound

We now use Massart to upper bound the generalization gap  $R(f) - R_n(f)$  for function classes of finite VC dimension, where  $|\mathcal{H}(Z^n)|$  does not grow exponentially in  $n$  for any  $Z^n$ .

Recap **definition VC dimension** for binary classification:

### Definition (VC dimension of $\mathcal{H}$ )

Biggest  $n \in \mathbb{N}$  s.t. there exists  $Z^n \in \mathcal{Z}^n$  with  $\mathcal{H}(Z^n) = \{0, 1\}^n$

Function classes  $\mathcal{F}$  with finite VC dimension can make  $\mathcal{H} = \ell \circ \tilde{\mathcal{F}}$  Glivenko-Cantelli, i.e.  $\mathcal{R}_n(\mathcal{H}) = o(1)$ . More specifically:

### Theorem (uniform VC bound)

If  $\mathcal{H}$  has VC dimension  $d_{VC}$ , w/ prob  $\geq 1 - \delta$  for any estimator  $f \in \mathcal{F}$

$$\mathbb{P}(yf(X) < 0) \leq \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{y_i f(x_i) < 0} + 4\sqrt{\frac{d_{VC} \log(n+1)}{n}} + \sqrt{\frac{2 \log(1/\delta)}{n}}$$

15 / 16

## References

Uniform law

- MW Chapter 4
- “Understanding machine learning” by Shalev-Shwartz, Ben-David, Chapter 26

16 / 16