# Mathematical Foundations of ML (CS 4785/5783)
# Lecture 3

Uniform Convergence, Symmetrization and Rademacher Complexity

http://www.cs.cornell.edu/Courses/cs4783/2022sp/notes03.pdf

# Statistical Learning Framework

$$\mathbf{D} \text{ is a distribution on } \mathcal{X} \times \mathcal{Y}$$

**D** captures the idea of this set **U**

$$\text{Training sample } S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \quad \text{Each } (x_t, y_t) \sim \mathbf{D}$$

$$\text{Risk of a model } g \text{ defined as } L_{\mathbf{D}}(g) = \mathbb{E}_{(x,y)\sim\mathbf{D}}\left[\ell(g(x), y)\right]$$

(Future instances drawn from **D**)

Excess risk of model $g$ w.r.t. model class $\mathcal{F}$ defined as

$$L_{\mathbf{D}}(g) - \min_{f \in \mathcal{F}} L_{\mathbf{D}}(f)$$

**Goal: provide an algorithm for which excess risk is small**

# Empirical Risk Minimization

**Pick a model in class that minimizes training error**

$$\hat{f}_{\mathrm{ERM}} \in \arg\min_{f \in \mathcal{F}} \widehat{L}_S(f)$$

- When does this succeed?
  - When model class is too complex, we already saw this can fail
  - When model class is say just one function, it succeeds due to law of large numbers (concentration)
  - In general how well does this algorithm do?

# ERM Over Finite Class

If losses are bounded by 1 (in absolute) and $|\mathcal{F}| < \infty$, then, for any $\delta > 0$ with probability at least $1 - \delta$,

$$L_{\mathbf{D}}(\hat{f}_{\mathrm{ERM}}) - \min_{f \in \mathcal{F}} L_{\mathbf{D}}(f) \leq \sqrt{\frac{8 \log\left(2|\mathcal{F}|/\delta\right)}{n}}$$

Hoeffding Inequality: Let $Z_1, \ldots, Z_n$ be a sequence of $n$ random variables bounded by $1$, drawn iid from a fixed distribution. Then:

$$P\left(\left|\frac{1}{n}\sum_{t=1}^{n} Z_t - \mathbb{E}Z\right| > \epsilon\right) \leq 2\exp(-\frac{n\epsilon^2}{2})$$

# ERM Over Finite Class

Hoeffding Inequality: Let $Z_1, \ldots, Z_n$ be a sequence of $n$ random variables bounded by $1$, drawn iid from a fixed distribution. Then:

$$P\left(\left|\frac{1}{n}\sum_{t=1}^{n} Z_t - \mathbb{E}Z\right| > \epsilon\right) \leq 2\exp(-\frac{n\epsilon^2}{2})$$

**Proof idea:**

$$\text{For each } f \in \mathcal{F} \text{ define } Z_t^f = \ell(f(x_t), y_t)$$

**Apply Hoeffding for each f individually**

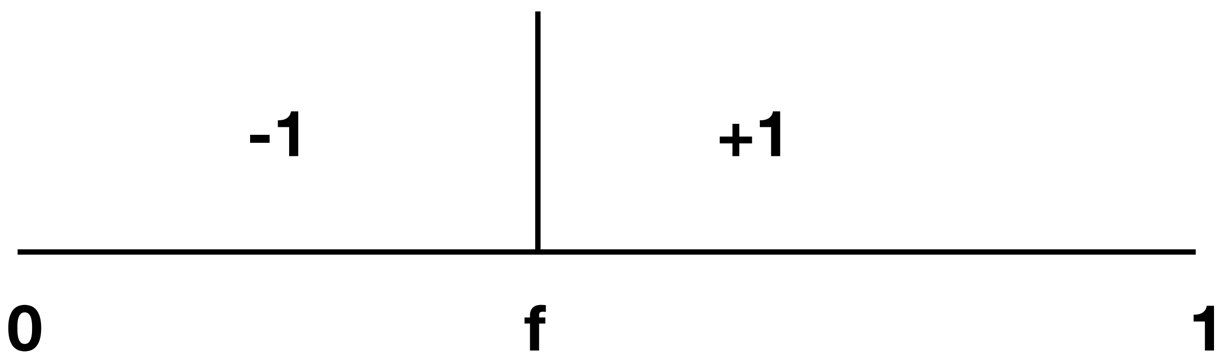
**Use union bound to move to uniform deviation**

# Beyond Finite Model Class

- Idea 1: Find a finite set $\mathcal{F}'$ such that for any $f \in \mathcal{F}$ there exists an. $f' \in \mathcal{F}'$ s.t.

$$\forall x, y, \quad |\ell(f'(x), y) - \ell(f(x), y)| < \Delta$$

- But this may not always work, consider the example of learning thresholds:




$$\mathcal{X} = [0, 1] \qquad f(x) = \text{sign}(x - f)$$

$$\mathcal{F} \text{ indexed by set } [0, 1]$$

For any $\Delta < 1/2$, this class cannot be approximated by a finite set.

# Uniform Convergence

We have shown that for any $\epsilon > 0$,

$$P\left(L_{\mathbf{D}}(\hat{f}_{\text{ERM}}) - \min_{f \in \mathcal{F}} L_{\mathbf{D}}(f) > 2\epsilon\right) \leq P\left(\max_{f \in \mathcal{F}} \left|\hat{L}_S(f) - L_{\mathbf{D}}(f)\right| > \epsilon\right)$$

Next, we will see that $\max_{f \in \mathcal{F}} \left|\hat{L}_S(f) - L_{\mathbf{D}}(f)\right|$ is concentrated near its expectation.

# McDiarmid's Inequality

Let $Z_1, \ldots, Z_n \in \mathcal{Z}$ be a sequence of $n$ random variables drawn iid from a fixed distribution. Assume that $\Phi : \mathcal{Z}^n \mapsto \mathbb{R}$ is a function satisfying the condition that: For any $i \in [n]$, and any $z_1, \ldots, z_n \in \mathcal{Z}$ and any $z_i' \in \mathcal{Z}$,

$$\left|\Phi(z_1, \ldots, z_i, \ldots, z_n) - \Phi(z_1, \ldots, z_i', \ldots, z_n)\right| \leq \frac{C}{n}$$

Then we have the following concentration result :

$$P\left(|\phi(Z_1, \ldots, Z_n) - \mathbb{E}\left[\phi(Z_1, \ldots, Z_n)\right]| > \epsilon\right) \leq 2\exp(-\frac{2n\epsilon^2}{C^2})$$

# Uniform Convergence

Eg: The function $\phi((x_1, y_1), \ldots, (x_n, y_n)) = \max_{f \in \mathcal{F}} \left|\hat{L}_S(f) - L_{\mathbf{D}}(f)\right|$ satisfies the condition with $C = 2$ when loss is bounded by $1$.

# Uniform Convergence

Eg: The function $\phi((x_1, y_1), \ldots, (x_n, y_n)) = \max_{f \in \mathcal{F}} \left|\hat{L}_S(f) - L_{\mathbf{D}}(f)\right|$ satisfies the condition with $C = 2$ when loss is bounded by $1$.

Hence we have that for any $\delta > 0$, with probability at least $1 - \delta$,

$$\max_{f \in \mathcal{F}} \left|\hat{L}_S(f) - L_{\mathbf{D}}(f)\right| \leq 2\, \mathbb{E}\left[\max_{f \in \mathcal{F}} \left|\hat{L}_S(f) - L_{\mathbf{D}}(f)\right|\right] + O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

**Complexity Measure**

# Symmetrization and Rademacher Complexity

Let $\epsilon_1, \ldots, \epsilon_n \in \{\pm 1\}$ be Rademacher random variables where each $\epsilon_i$ is $+1$ with probability $1/2$ and $-1$ with probability $1/2$.

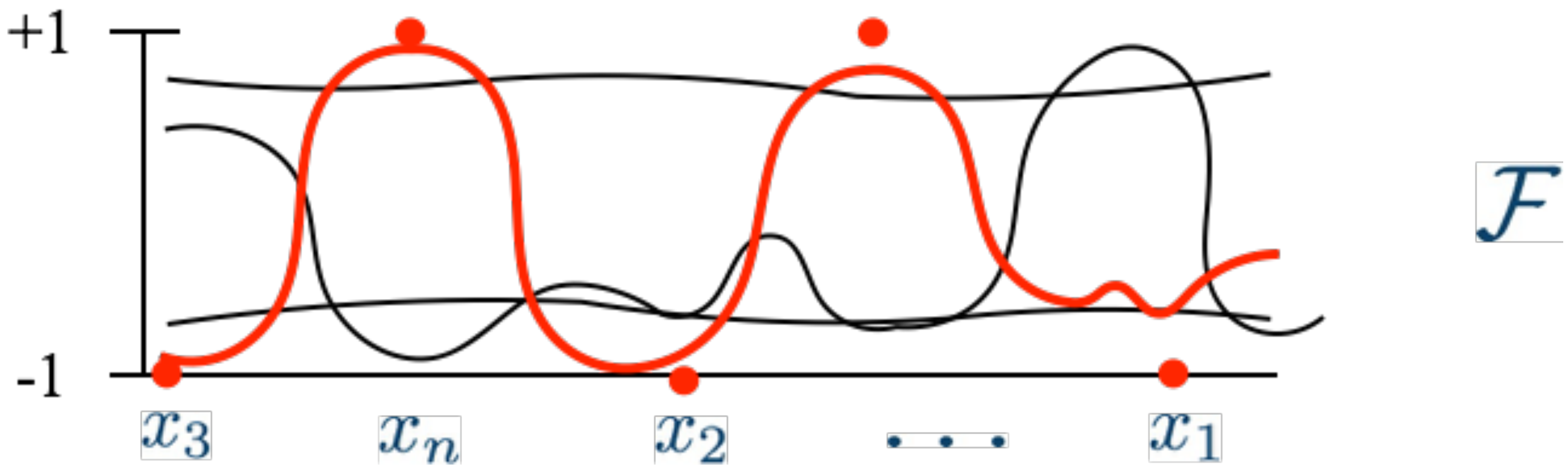
We will see that:

$$\mathbb{E}\left[\max_{f \in \mathcal{F}} \left|\hat{L}_S(f) - L_{\mathbf{D}}(f)\right|\right] \leq \frac{2}{n} \mathbb{E}_S\left[\mathbb{E}_\epsilon\left[\max_{f \in \mathcal{F}} \left|\sum_{t=1}^{n} \epsilon_t \ell(f(x_t), y_t)\right|\right]\right]$$

**Rademacher Complexity**

# Rademacher Complexity

Example : $\mathcal{X} = [0,1],\ \mathcal{Y} = [-1,1]$



# Proof of the Result

# Why is this useful?