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 :

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

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.

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 ϵ_i is $+1$ with probability $1/2$ and -1 with probability $1/2$.

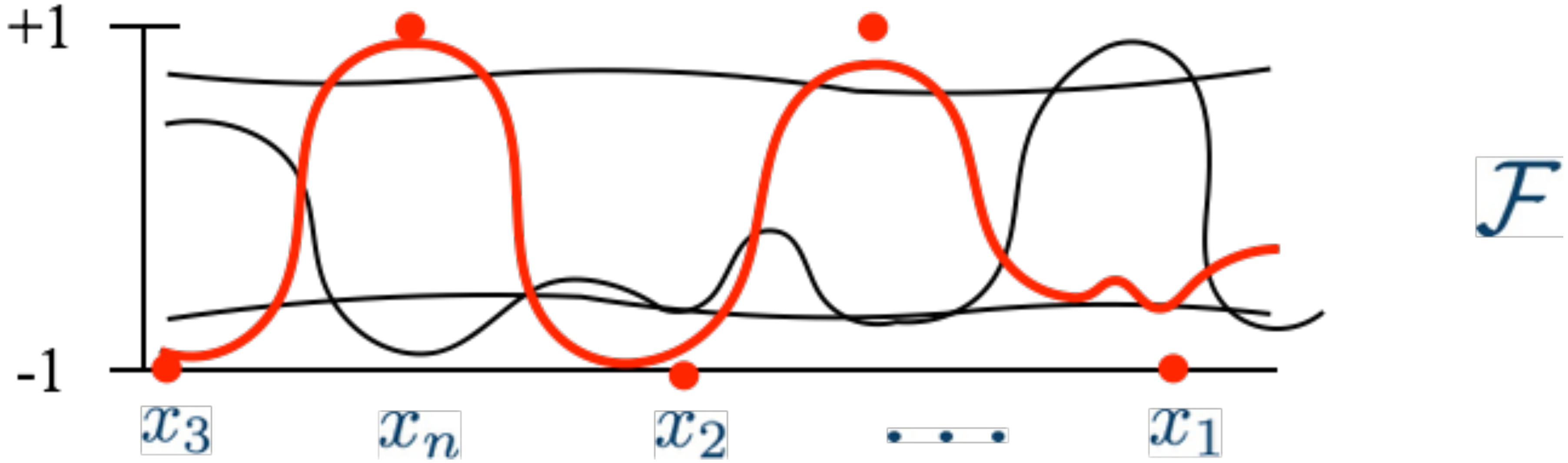
We will see that:

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

Rademacher Complexity

Rademacher Complexity

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



Proof of Symmetrization

Why is moving to Rademacher Complexity useful?

Rademacher Complexity

Given sample, define $\mathcal{F}_{|x_1,\ldots,x_n} = \{(f(x_1),\ldots,f(x_n)) : f \in \mathcal{F}\}$

$$\mathbb{E}_\epsilon\left[\max_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \epsilon_t \ell(f(x_t), y_t)\right] = \mathbb{E}_\epsilon\left[\max_{\mathbf{f}\in\mathcal{F}_{|x_1,\ldots,x_n}} \frac{1}{n}\sum_{t=1}^{n} \epsilon_t \ell(\mathbf{f}_t, y_t)\right]$$

$\mathcal{F}_{|x_1,\ldots,x_n}$

Rademacher Complexity

Given sample, define $\mathcal{F}_{|x_1,\ldots,x_n} = \{(f(x_1),\ldots,f(x_n)) : f \in \mathcal{F}\}$

$$\mathbb{E}_{\epsilon}\left[\max_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \epsilon_t \ell(f(x_t), y_t)\right] = \mathbb{E}_{\epsilon}\left[\max_{\mathbf{f}\in\mathcal{F}_{|x_1,\ldots,x_n}} \frac{1}{n}\sum_{t=1}^{n} \epsilon_t \ell(\mathbf{f}_t, y_t)\right]$$

For each f, this term is average of 0 mean terms and hence concentrates

Only cardinality of $\mathcal{F}_{|x_1,\ldots,x_n}$ **matters**

Rademacher Complexity

Lemma 4. *For any class $\mathcal{F}$ and any loss bounded by* 1,

$$\mathbb{E}_{\epsilon}\left[\max_{\mathbf{f}\in\mathcal{F}_{|x_1,\ldots,x_n}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(\mathbf{f}[t],y_t)\right|\right]\leq O\left(\sqrt{\frac{\log\left|\mathcal{F}_{|x_1,\ldots,x_n}\right|}{n}}\right)$$

Eg. Thresholds, rectangle

Growth Function and VC dimension

$$\Pi(\mathcal{F}, n) = \max_{x_1,\ldots,x_n} \left|\mathcal{F}_{|x_1,\ldots,x_n}\right|$$

Consider the case of binary classification:

Definition 1. *VC dimension of a binary function class $\mathcal{F}$ is the largest number of points $d = \mathrm{VC}(\mathcal{F})$, such that*

$$\Pi_{\mathcal{F}}(d) = 2^d$$

If no such d exists then $\mathrm{VC}(\mathcal{F}) = \infty$

Maximum number of points that can be shattered.

If VC dimension is infinite then learning is not possible!

Think of a proof strategy

Growth Function and VC dimension

Lemma 3 (VC'71/Sauer'72/Shelah'72). *For any class* $\mathcal{F} \subset \{\pm 1\}^{\mathcal{X}}$ *with* $\mathrm{VC}(\mathcal{F}) = d$*, we have that,*

$$\Pi(\mathcal{F}, n) \leq \sum_{i=0}^{d} \binom{n}{i}$$

Proof of the above lemma is done via induction on $n + d$. Also note that $\sum_{i=0}^{d} \binom{n}{i} \leq n^d$

If VC is finite, growth function has a nice bound and hence we can learn!