

As in the previous lecture, let  $\mathcal{H} = \{h : \mathcal{X} \mapsto [-1, 1]\}$  be a VC-subgraph class and  $f \in \mathcal{F} = \text{conv } \mathcal{H}$ . The classifier is  $\text{sign}(f(x))$ . The set

$$\{y \neq \text{sign}(f(x))\} = \{yf(x) \leq 0\}$$

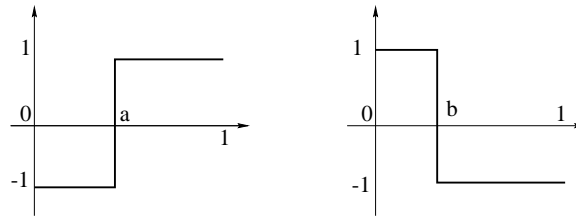
is the set of misclassified examples and  $\mathbb{P}(yf(x) \leq 0)$  is the misclassification error.

Assume the examples are labeled according to  $C_0 = \{x \in \mathcal{X} : y = 1\}$ . Let  $C = \{\text{sign}(f(x)) > 0\}$ . Then  $C_0 \Delta C$  are misclassified examples.

$$\mathbb{P}(C \Delta C_0) = \frac{1}{n} \sum_{i=1}^n I(x_i \in C \Delta C_0) + \underbrace{\mathbb{P}(C \Delta C_0) - \frac{1}{n} \sum_{i=1}^n I(x_i \in C \Delta C_0)}_{\text{small. estimate uniformly over sets } C} .$$

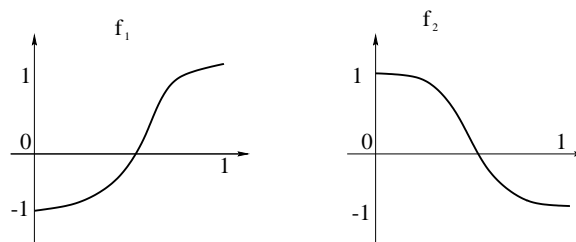
For voting classifiers, the collection of sets  $C$  can be "very large".

**Example 1.** Let  $\mathcal{H}$  be the class of simple step-up and step-down functions on the  $[0, 1]$  interval, parametrized by  $a$  and  $b$ . Then  $VC(\mathcal{H}) = 2$ . Let  $\mathcal{F} = \text{conv } \mathcal{H}$ . First, rescale the



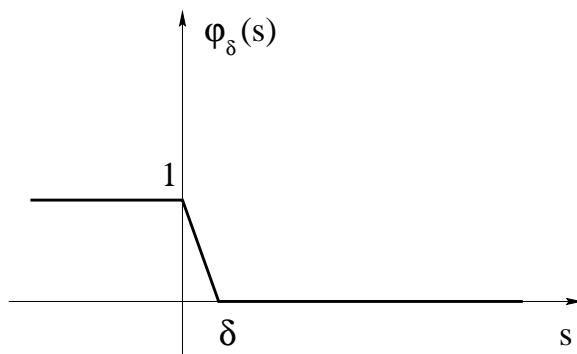
functions:  $f = \sum_{i=1}^T \lambda_i h_i = 2 \sum_{i=1}^T \lambda_i \left(\frac{h_i+1}{2}\right) - 1 = 2f' - 1$  where  $f' = \sum_{i=1}^T \lambda_i h'_i$ ,  $h'_i = \frac{h_i+1}{2}$ .

We can generate any non-decreasing function  $f'$  such that  $f'(0) = 0$  and  $f'(1) = 1$ . Similarly, we can generate any non-increasing  $f'$  such that  $f'(0) = 1$  and  $f'(1) = 0$ . Rescaling back to  $f$ , we can get any non-increasing and non-decreasing functions of the form



Any function with sum of jumps less than 1 can be written as  $f = \frac{1}{2}(f_1 + f_2)$ . Hence, we can generate basically all sets by  $\{f(x) > 0\}$ , i.e.  $\text{conv } \mathcal{H}$  is bad.

Recall that  $\mathbb{P}(yf(x) \leq 0) = \mathbb{E}I(yf(x) \leq 0)$ . Define function  $\varphi_\delta(s)$  as follows:



Then,

$$I(s \leq 0) \leq \varphi_\delta(s) \leq I(s \leq \delta).$$

Hence,

$$\begin{aligned} \mathbb{P}(yf(x) \leq 0) &\leq \mathbb{E}\varphi_\delta(yf(x)) \\ &= \frac{1}{n} \sum_{i=1}^n \varphi_\delta(y_i f(x_i)) + \left( \mathbb{E}\varphi_\delta(yf(x)) - \frac{1}{n} \sum_{i=1}^n \varphi_\delta(y_i f(x_i)) \right) \\ &\leq \frac{1}{n} \sum_{i=1}^n I(y_i f(x_i) \leq \delta) + \left( \mathbb{E}\varphi_\delta(yf(x)) - \frac{1}{n} \sum_{i=1}^n \varphi_\delta(y_i f(x_i)) \right) \end{aligned}$$

By going from  $\frac{1}{n} \sum_{i=1}^n I(y_i f(x_i) \leq 0)$  to  $\frac{1}{n} \sum_{i=1}^n I(y_i f(x_i) \leq \delta)$ , we are penalizing small confidence predictions. The margin  $yf(x)$  is a measure of the confidence of the prediction.

For the sake of simplicity, denote  $\mathbb{E}\varphi_\delta = \mathbb{E}\varphi_\delta(yf(x))$  and  $\bar{\varphi}_\delta = \frac{1}{n} \sum_{i=1}^n \varphi_\delta(y_i f(x_i))$ .

**Lemma 18.1.** Let  $\mathcal{F}_d = \text{conv}_d \mathcal{H} = \{\sum_{i=1}^d \lambda_i h_i, h_i \in \mathcal{H}\}$  and fix  $\delta \in (0, 1]$ . Then

$$\mathbb{P} \left( \forall f \in \mathcal{F}_d, \frac{\mathbb{E}\varphi_\delta - \bar{\varphi}_\delta}{\sqrt{\mathbb{E}\varphi_\delta}} \leq K \left( \sqrt{\frac{dV \log \frac{n}{\delta}}{n}} + \sqrt{\frac{t}{n}} \right) \right) \geq 1 - e^{-t}.$$

*Proof.* Denote

$$\varphi_\delta(y\mathcal{F}_d(x)) = \{\varphi_\delta(yf(x)), f \in \mathcal{F}_d\}.$$

Note that  $\varphi_\delta(yf(x)) : \mathcal{X} \times \mathcal{Y} \mapsto [0, 1]$ .

For any  $n$ , take any possible points  $(x_1, y_1), \dots, (x_n, y_n)$ . Since

$$|\varphi_\delta(s) - \varphi_\delta(t)| \leq \frac{1}{\delta} |s - t|,$$

we have

$$\begin{aligned}
 d_{x,y}(\varphi_\delta(yf(x)), \varphi_\delta(yg(x))) &= \left( \frac{1}{n} \sum_{i=1}^n (\varphi_\delta(y_i f(x_i)) - \varphi_\delta(y_i g(x_i)))^2 \right)^{1/2} \\
 &\leq \left( \frac{1}{\delta^2} \frac{1}{n} \sum_{i=1}^n (y_i f(x_i) - y_i g(x_i))^2 \right)^{1/2} \\
 &= \frac{1}{\delta} \left( \frac{1}{n} \sum_{i=1}^n (f(x_i) - g(x_i))^2 \right)^{1/2} \\
 &= \frac{1}{\delta} d_x(f, g)
 \end{aligned}$$

where  $f, g \in \mathcal{F}_d$ .

Choose  $\varepsilon \cdot \delta$ -packing of  $\mathcal{F}_d$  so that

$$d_{x,y}(\varphi_\delta(yf(x)), \varphi_\delta(yg(x))) \leq \frac{1}{\delta} d_x(f, g) \leq \varepsilon.$$

Hence,

$$\mathcal{N}(\varphi_\delta(y\mathcal{F}_d(x)), \varepsilon, d_{x,y}) \leq \mathcal{D}(\mathcal{F}_d, \varepsilon\delta, d_x)$$

and

$$\log \mathcal{N}(\varphi_\delta(y\mathcal{F}_d(x)), \varepsilon, d_{x,y}) \leq \log \mathcal{D}(\mathcal{F}_d, \varepsilon\delta, d_x) \leq KdV \log \frac{2}{\varepsilon\delta}.$$

We get

$$\log \mathcal{D}(\varphi_\delta(y\mathcal{F}_d), \varepsilon/2, d_{x,y}) \leq KdV \log \frac{2}{\varepsilon\delta}.$$

So, we can choose  $f_1, \dots, f_D$ ,  $D = \mathcal{D}(\mathcal{F}_d, \varepsilon\delta, d_x)$  such that for any  $f \in \mathcal{F}_d$  there exists  $f_i$ ,  $d_x(f, f_i) \leq \varepsilon\delta$ . Hence,

$$d_{x,y}(\varphi_\delta(yf(x)), \varphi_\delta(yf_i(x))) \leq \varepsilon$$

and  $\varphi_\delta(yf_1(x)), \dots, \varphi_\delta(yf_D(x))$  is an  $\varepsilon$ -cover of  $\varphi_\delta(y\mathcal{F}_d(x))$ . □