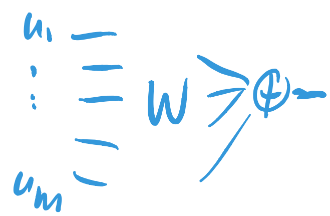
- Lecture began with discussion of Theorem 8.4 and its application to get bound for AdaBoost (Thm. 8.6)

Second part discussed Sect. 8.5 beginning after first two pages.
i.e. beginning with $\mathcal{F}_0, \mathcal{F}_1, \ldots, \mathcal{F}_\ell$ (multiple layer neural nets)

Remainder (see below)

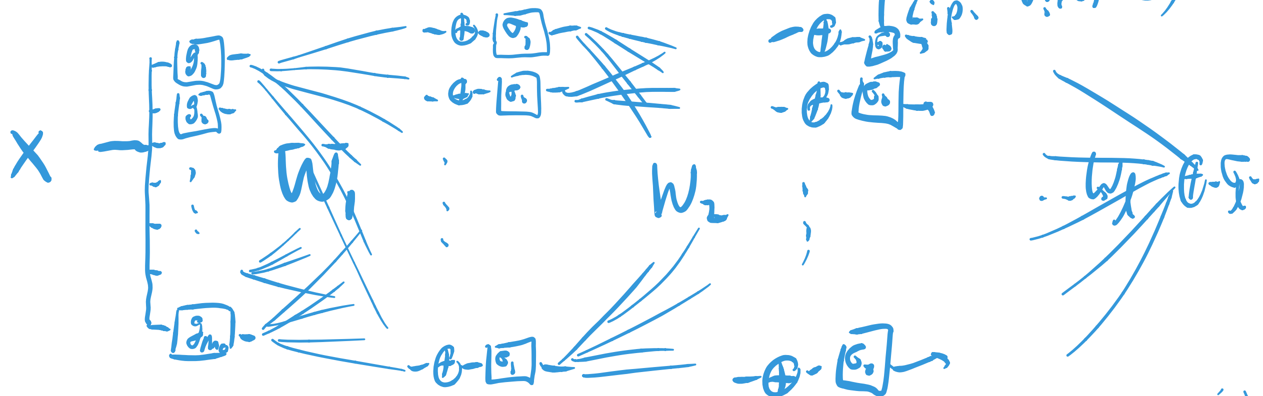
Neural Nets (Second half of section 8.5)

$$N_{\sigma,w}(u_1,\cdots,u_m) = \sigma\left(\sum_{j=1}^{m} w_j u_j\right)$$



$g \qquad g:\mathbb{R}^d \to \mathbb{R}, \quad x \in \mathbb{R}^d$

$\left(\sigma_i:\mathbb{R}\to\mathbb{R} \quad \text{Lip.}, \quad \sigma_i(0)=0\right)$



$\mathrm{sgn}(\cdot) = \pm 1$

Let $\mathcal{F}_0 = g$ and for $1 \le j \le l$,

$$\mathcal{F}_j = \{N_{\sigma_j,w}\circ(f_1,\cdots,f_m) : m\in\mathbb{N},\ |w_1|+\cdots+|w_m| \le B_j,\ f_1,\cdots,f_m \in \mathcal{F}_{j-1}\}$$

$$\mathcal{F}_j = \sigma_j \circ \left(B_j \cdot \mathrm{absconv}(\mathcal{F}_{j-1})\right)$$

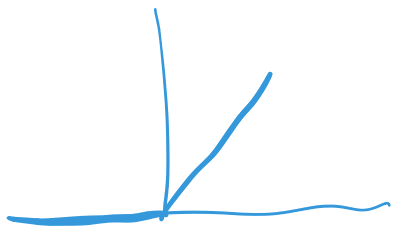
$$R_n(\mathcal{F}_0(x^n)) = R_n(g(x^n))$$

L_j = Lipschitz constant for σ_j

$$R_n(\mathcal{F}_j(X^n)) \le (2L_jB_j)R_n(\mathcal{F}_{j-1}(X^n))$$

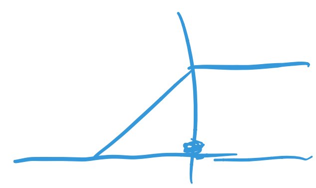
$$\therefore \quad R_n(\mathcal{F}_l(X^n)) \le \underline{\prod^{l}(2L_jB_j)}\, R_n(g(X^n))$$

$$\sqrt{l}$$



$L_j = 1$

± 1



8.5 First part - (neural nets)

Let $\mathcal{F}$ the set of functions : $\mathbb{R}^d \to \mathbb{R}$

$$f_w(x) = \langle w, x \rangle = \sum_{i=1}^{d} w_i x_i$$

such that $\|w\| \leq B$ $\qquad \|w\| = \sqrt{\sum_{i=1}^{d} w_i^2}$

$$R_n(\mathcal{F}(x^n)) = \frac{1}{n} E_{\varepsilon^n}\left[\sup_{\|w\|\leq B} \left|\sum_{i=1}^{n} \langle w, x_i \rangle \varepsilon_i\right|\right]$$

$$= \frac{1}{n} E_{\varepsilon^n}\left[\sup_{\|w\|\leq B} \left|\langle w, \sum_{i=1}^{n} x_i \varepsilon_i \rangle\right|\right]$$

$$= \frac{B}{n} E_{\varepsilon^n}\left[\left\|\sum_{i=1}^{n} x_i \varepsilon_i\right\|\right]$$

$$= \frac{B}{n} E_{\varepsilon^n}\left[\sqrt{\sum_{i=1}^{n} x_i \varepsilon_i \cdot \sum_{j=1}^{n} x_j \varepsilon_j}\right]$$

$$\leq \frac{B}{n} \sqrt{E_{\varepsilon^n}\left[\qquad\right]} \qquad \text{Jensen}$$

$$= \frac{B}{n} \sqrt{\sum_{i=1}^{n} \|x_i\|^2}$$

$$R_n(\mathcal{F}(x^n)) \leq \frac{B}{n} \sqrt{\sum_{i=1}^{n} \|x_i\|^2} \qquad x_i \in \mathbb{R}^d$$

$$\leq \frac{BR}{\sqrt{n}} \quad \text{if} \quad \|x_i\| \leq R$$

with probability one.

Chapter 4.2 spaces determined by kernels

RKHS - reproducing kernel Hilbert space.

It's all about a measure of complexity of functions of the form $\underbrace{\sum_{i=1}^{m} a_i g_i(x)}_{f(x)}$.

$$\underset{(a)}{\text{Complexity}(f)} = \sum_{i=1}^{m} a_i^2 / w_i \quad \left(\text{for some weights } w_1, \cdots w_m\right)$$

Mercer kernel $K(x,y) \to \mathbb{R}$

$x, y \in X$ $\quad (X = \mathbb{R}^d)$

(1) K is continuous in x, y

(2) symmetric $K(x,y) = K(y,x)$

(3) K is positive semidefinite:

$$\sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \geq 0$$

all $m \geq 1$

for any $x_1, \cdots x_m \in X$, and $\alpha_1, \cdots \alpha_m \in \mathbb{R}$

<u>Examples</u> $K(x,y) = (1 + \langle x, y \rangle)^k$

$$K(x,y) = \exp(-\|x-y\|^2)$$

For x fixed let $K_x(y) = K(x,y)$

$$K_x : \mathbb{R}^d \to \mathbb{R}$$

The RKHS is spanned by the functions K_x.

<u>And</u> $\langle K_x, K_{x'} \rangle_K \hat{=} K(x,x')$