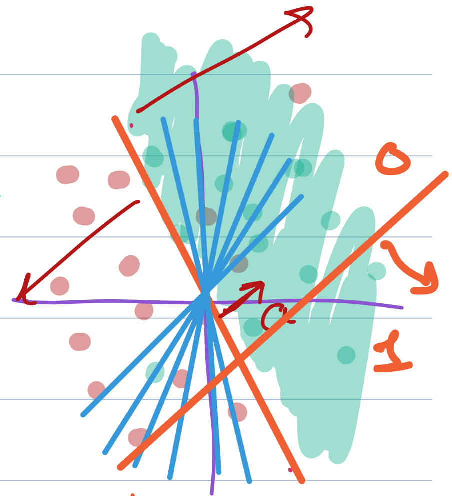
# Regularization (explicit / implicit)




Input: $X_1, X_2, \ldots, X_n \sim D$ on $\mathbb{R}^d$ ← iid

$f(x_1)\ f(x_2) \cdots f(x_n)$

Ground truth: $f: \mathbb{R}^d \to \{0,1\}$

$g: \mathbb{R}^d \to \{0,1\}$
predictor

Thm: If the ground truth

$f \in \{f_1, f_2, \ldots, f_m\}$

and $\boxed{n \geq \frac{1}{\varepsilon}\left(\ln(m) + \ln(1/\delta)\right)}$

Output $g \in \{f_1, \ldots, f_m\}$
that min. Train Err $(g)$

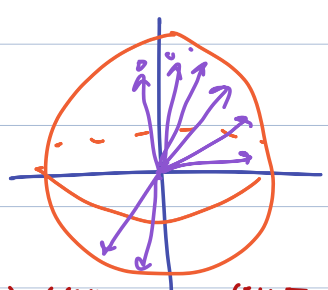
then ERM produces $g: \mathbb{R}^d \to \{0,1\}$

s.t. w/ prob $\geq 1-\delta$: $\underbrace{\text{Gen Error}(g)}_{\mathbb{P}_{x \sim D}[g(x) \neq f(x)]} \leq \varepsilon$.

## Linear classifiers

$a \in \mathbb{R}^d$

$$f_a(x) = \begin{cases} 1 & \langle x, a\rangle \geq 0 \\ 0 & \langle x, a\rangle < 0 \end{cases}$$



$m \simeq c^d \quad (\simeq 5^d)$

$n \geq \frac{1}{\varepsilon}\left(d \log(c) + \log(1/\delta)\right)$

$\simeq \frac{c_0}{\varepsilon}\left(d + \ln(1/\delta)\right)$

← free params

$100 \times 100 \times \cdots \times 100 = 100^d$

$(X_1, X_2, \ldots, X_d)$

$[0,1]$ $\underbrace{\{0.01, 0.02, \ldots, 0.99, 1.00\}}$

$\boxed{d \gg n}$

100

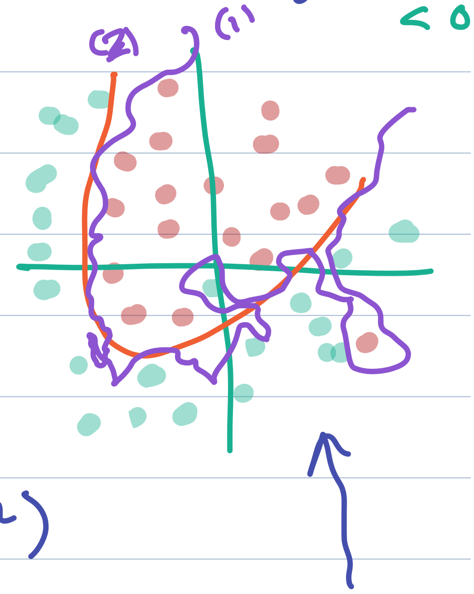
$$\langle a, x\rangle = a_1 x_1 + a_2 x_2 + \cdots + a_d x_d \quad \begin{matrix}\geq 0\\ < 0\end{matrix}$$

$$\underset{x}{(x_1,\dots,x_d)} \longmapsto \overbrace{(1, x_1, x_2, \dots, x_d)}^{F(x)}$$

$$F: \mathbb{R}^d \to \mathbb{R}^{d+1} \qquad \langle a, F(x)\rangle$$

$$= \underline{a_0} + a_1 x_1 + \cdots + a_d x_d \quad \begin{matrix}\geq 0\\ < 0\end{matrix}$$

$$\boxed{\sum_{i=0}^{d}\sum_{j=0}^{d} a_{ij} x_i x_j} \qquad \begin{matrix}\geq 0 \Rightarrow 1\\ < 0 \Rightarrow 0\end{matrix}$$




$$F(x) = (1, x_1, \dots, x_d, x_1 x_2, x_1 x_3, \dots x_2 x_1, x_2^2, \dots, x_{d-1} x_d, x_d^2)$$

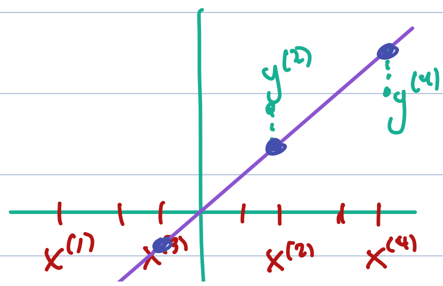
$$F: \mathbb{R}^d \to \mathbb{R}^{(d+1)^2}$$

$$\langle a, F(x)\rangle = \sum_{i=0}^{d}\sum_{j=0}^{d} a_{ij} x_i x_j \qquad \begin{matrix}\geq 0 \Rightarrow 1\\ < 0 \Rightarrow 0\end{matrix}$$

## Linear regression

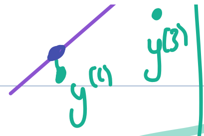
$$x^{(1)}, x^{(2)}, \dots, x^{(n)} \in \mathbb{R}^d$$

$$y^{(1)}, y^{(2)}, \dots, y^{(n)} \in \mathbb{R}$$




$$\boxed{f(a) := \sum_{i=1}^{n} \left(\underline{\langle x^{(i)}, a\rangle} - \underline{y^{(i)}}\right)^2}$$

$a \in \mathbb{R}^d$




$x^{(i)}$: $\underline{\underline{\langle a, x^{(i)}\rangle}}$

ERM:
$\underline{\underline{\text{argmin}}}\{f(a) : a \in \mathbb{R}^d\}$

$$\hat{f}(a) := \sum_{i=1}^{n} (\langle x^{(i)}, a\rangle - y^{(i)})^2 + \lambda \|a\|_2^2$$

encourages smoothness

$|a_1|_2^2 + \dots + |a_d|_2^2$

$\text{argmin}\{\hat{f}(a) : a \in \mathbb{R}^d\}$

$$\left.\begin{cases} x^{(1)} = (z_1, z_1) & y^{(1)} = 10 \cdot z_1 \\ x^{(2)} = (z_2, z_2) & y^{(2)} = 10 \cdot z_2 \\ \vdots & \vdots \\ x^{(n)} = (z_n, z_n) & y^{(n)} = 10 \cdot z_n \end{cases}\right]$$

$(a_1, a_2)$

$(10, 0), (0, 10)$

$\boxed{a_1 + a_2 = 10}$

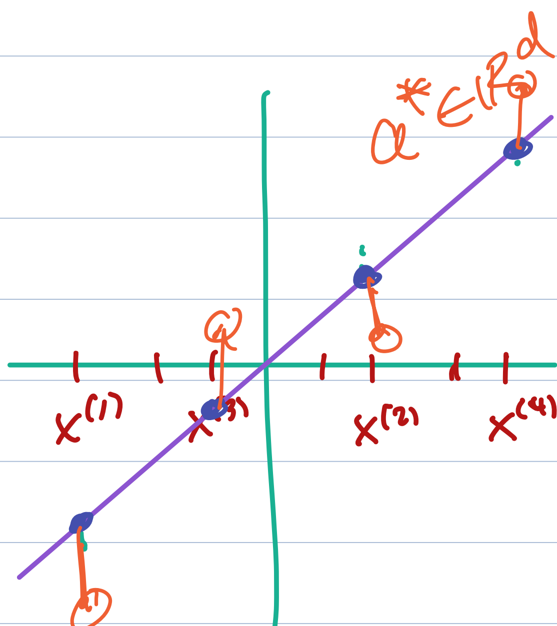
$$\langle a, x^{(i)}\rangle = a_1 z_i + a_2 z_i = (a_1 + a_2) z_i$$

$a = (-1000{,}000, 1{,}000{,}010)$

$a^* = (5, 5)$

$\min\{a^2 + b^2 : a + b = 10\}$

$a = b = 5$




MLE:

$\hat{f}(a) = f(a)$

$$y^{(i)} = \langle a^*, x^{(i)}\rangle + N(0, \sigma^2)$$

$$\boxed{\hat{f}(a) = \sum_{i=1}^{n} (\langle x^{(i)}, a \rangle - y^{(i)})^2 + \frac{1}{\sigma^2} \|a\|_2^2}$$

$$\operatorname{argmin} \{ \hat{f}(a) : a \in \mathbb{R}^d \}$$

---

## Sparsity and $\ell_0$ regularization

$$\hat{f}(a) = \sum_{i=1}^{n} (\langle x^{(i)}, a \rangle - y^{(i)})^2 + \cancel{\lambda \|a\|_0}$$

$$\boxed{\|a\|_0 = |\{ i : a_i \neq 0 \}|}$$

$(0, 0, \dots, 0)$ $\quad 0$

$\downarrow$

$(0, \varepsilon, 0, \dots, 0)$ $\quad 1$

Proxy: In many cases, encourages sparsity:

$$\hat{f}(a) = \sum_{i=1}^{n} (\langle x^{(i)}, a \rangle - y^{(i)})^2 + \lambda \|a\|_1$$

$|a_1| + |a_2| + \dots + |a_n|$

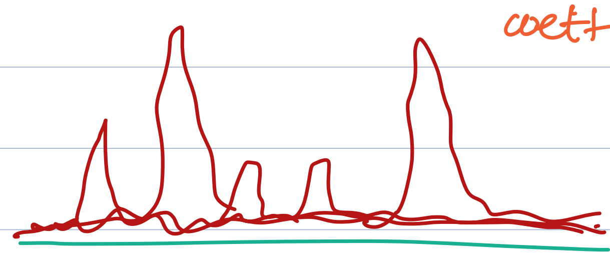
Optimized eff (+ linear prog.)

$(1, 0, 0, \dots, 0)$ $\quad 1$

$\left(\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}, \dots, \frac{1}{\sqrt{d}}\right)$ $\quad \sqrt{d}$

$k$ large coeff

Compressive sensing:



$\boxed{A : \mathbb{R}^N \longrightarrow \mathbb{R}^d}$ $N >> d$

$(Ax)_i = \langle a_i, x \rangle$

$\uparrow$

$(1,0,0,1,1,0,1,0,1,\dots)$

If $x \in \mathbb{R}^N$ is $k$-sparse then if $d \approx k \log k$, can recover $x$ from $Ax$

Given $\boxed{y = Ax}$

$$\boxed{\min_{\hat{x} \in \mathbb{R}^N} \{ \|\hat{x}\|_1 : A\hat{x} = y \}}$$

$y = A\hat{x} \leftarrow (N-d)$-space of solutions