Last time

- SVD + applications
- least squares
- perceptron alg.

Today

- PAC learning
- Gradient descent & SGD
- Linear programming (maybe)

Setting: supervised learning setting

classify email msgs → spam / → not spam

Take a sample of msgs, labelled according to spam Y/N.

Goal: given labelled sample, come up with a good rule for classifying future msgs

$h^* : \{0,1\}^n \longrightarrow \underbrace{\{0,1\}}_{\text{true label}}$

	money	pills	Mr.	bad spelling	known-sender	spam?
1	Y	N	Y	Y	N	Y
2	N	N	N	Y	Y	N
3	N	Y	N	N	N	Y
4	Y	N	N	N	Y	N
5	N	N	Y	N	Y	N
6	Y	N	N	Y	N	Y
7	N	N	Y	N	N	N
8	N	Y	N	Y	N	Y

$h^*(x_1)$ $h^*(x_2)$

return SPAM if ¬known and (money or pills)

$\mathcal{D}$ distn over inputs $x \in \mathcal{X}$

each sample x_i is drawn indep from $\mathcal{D}$

see m samples

future inputs are also drawn from same distn

$$S = \begin{cases} (x_1, y_1) \\ (x_2, y_2) \\ \vdots \\ (x_m, y_m) \end{cases} \longrightarrow \boxed{\text{learng alg}} \longrightarrow h \in \mathcal{H}$$

$y_i = h^*(x_i)$

email spam Y/N

$h: \{0,1\}^n \rightarrow \{0,1\}$

hypothesis class

assume the mapping h^* is the true mapping

$$\mathrm{err}_S(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{h(x_i) \neq h^*(x_i)}$$

training
empirical error
sample

try to find $h \in \mathcal{H}$ that minimizes $\mathrm{err}_S(h)$

empirical risk minimization

$$\mathrm{err}_D(h) = \Pr_{x \sim D}(h(x) \neq h^*(x))$$

generalization error
true error.

$\mathcal{H}$: decision lists $\vec{x} \in \{0,1\}^n = (x_1, x_2, \dots, x_n)$

$x_3 = 1 \rightarrow x_4 = 0 \rightarrow x_5 = 1 \rightarrow 1$

$x_3 = 1 \downarrow 0$, $x_4 = 0 \downarrow 1$, $x_5 = 1 \downarrow 0$

$$\boxed{|\mathcal{H}| = n! \, (2 \cdot 2)^n = n! \, 4^n}$$

x_1	x_2	x_3	x_4	x_5	label
1	0	0	1	1	+
0	1	1	0	0	−
1	1	1	0	0	+
0	0	0	1	0	−
1	1	0	1	1	+
1	0	0	0	1	−

find if-then rule consistent w/ remaining data until either no examples left or can't find anything that works.

$x_1 = 0 \rightarrow x_2 = 1 \rightarrow x_4 = 1 \rightarrow$ remaining examples

$x_1 = 0$: − (rets), $x_2 = 1$: +, $x_4 = 1$: +

when get stuck: no rule consistent w/ remaining ...
⋮

have a nice alg for finding consistent DL if such exists.

Confidence, generalization – Claim: if $|S|$ was $\geq$ ☐ then w.h.p. $\mathrm{err}_D(h)$ small.

h misclassifies $\geq \varepsilon$ ... h classifies correctly

Consider some DL $h \in \mathcal{H}$ that $\boxed{\mathrm{err}_D(h) > \varepsilon}$

Prob(h was consistent w/ our sample)

$\leq (1-\varepsilon)^{|S|}$

$\mathcal{H}$

$\Pr(\exists$ DL h such $\mathrm{err}_D(h) \geq \varepsilon$ but $\mathrm{err}_S(h) = 0)$

$$\leq \sum_{\substack{h \in \mathcal{H} \\ \text{s.t. } \mathrm{err}_D(h) \geq \varepsilon}} \Pr(\mathrm{err}_S(h) = 0) \leq \boxed{|\mathcal{H}|(1-\varepsilon)^{|S|}}$$

$(1-x) \leq e^{-x}$

$|\mathcal{H}| e^{-\varepsilon|S|}$

How big does S need to be so that $> \delta$

ε, δ

$$|\mathcal{H}|(1-\varepsilon)^{|S|} \leq \delta$$

$$n!\,4^n e^{-\varepsilon|S|} \leq \delta$$

$$\frac{n^n 4^n}{\delta} \leq e^{\varepsilon|S|}$$

$n \ln n + n \ln 4 + \ln\left(\frac{1}{\delta}\right) \leq \varepsilon |S|$

$\ln |\mathcal{H}|$

$$\boxed{\frac{2}{\varepsilon}\left(n \ln n + \ln\left(\frac{1}{\delta}\right)\right) \leq |S|}$$

0.01 0.01

if $|S| = \Omega\left(\frac{1}{\varepsilon}\left(n \ln n + \frac{1}{\delta}\right)\right)$ then

$\Pr\left(\exists h \in \mathcal{H} \text{ s.t. } err_D(h) \geq \varepsilon \text{ and } err_S(h) = 0\right) \leq \delta$

If we can find $h \in \mathcal{H}$ that is consistent w/ sample

then rule h we find

is probably approximately correct

$\geq 1-\delta$ error $\leq \varepsilon$

$\boxed{\text{PAC-learning}}$

Turing award.

Leslie Valiant.

$\mathcal{H}$

If $\boxed{|S| \geq \frac{1}{\varepsilon}\left(\ln |\mathcal{H}| + \ln\left(\frac{1}{\delta}\right)\right)}$ sample complexity

then w prob $\geq 1-\delta$ any $h \in \mathcal{H}$

that has $err_D(h) \geq \varepsilon$ will have $err_S(h) > 0$

assumed $\exists h \in \mathcal{H}$ s.t. $err_S(h) = 0$

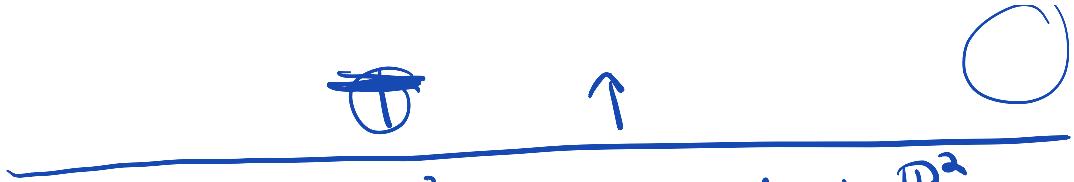
Thm:

If $\boxed{|S| \geq \frac{1}{2\varepsilon^2}\left(\ln |\mathcal{H}| + \ln\left(\frac{1}{\delta}\right)\right)}$ sample complexity

then w prob $\geq 1-\delta$ for every $h \in \mathcal{H}$

$$|err_S(h) - err_D(h)| \leq \varepsilon$$

h^*



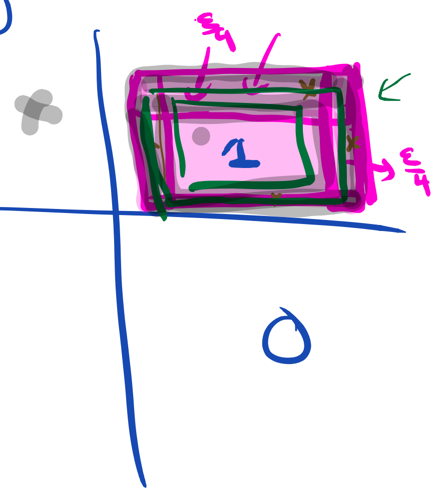
$\mathcal{H} = \{ \text{rectangles} \}$

samples in $\mathbb{R}^2$

$h: \mathbb{R}^2 \to \{0,1\}$

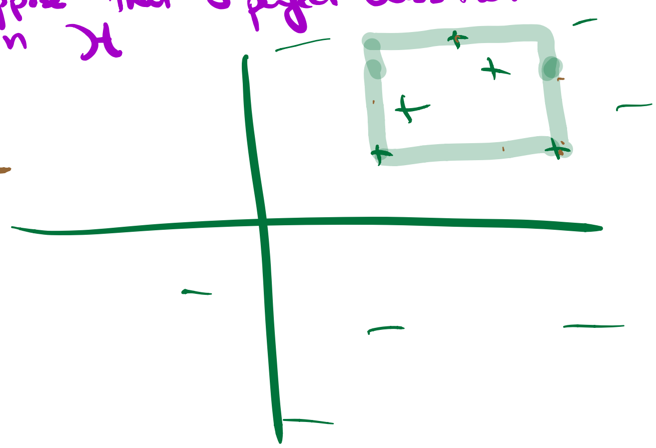
$$h(x,y) = \begin{cases} 1 & x_1 \le x \le x_2,\ y_1 \le y \le y_2 \\ 0 & \text{o.w.} \end{cases}$$

x_1, x_2, y_1, y_2.



Suppose that $\exists$ perfect classifier in $\mathcal{H}$

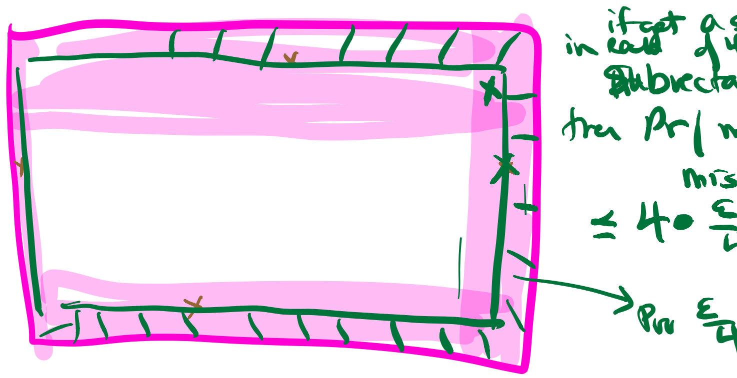
green inside pink



output smallest possible bounding rectangle.

Pr (green rectangle has error $\ge \varepsilon$)

If $\exists$ sample in each 4 little pink rectangles, then Pr (error on random draw $\le \varepsilon$)



if get a sample in each of 4 pink subrectangles then Pr(making a mistake) $\le 4 \cdot \frac{\varepsilon}{4} \le \varepsilon$

$$\Pr[\mathrm{err}_D(\text{smallest bounding rectangle}) > \varepsilon]$$

$$\le \Pr[\exists \text{ pink subrectangle w/ no sample in it}]$$

$$\le 4 \cdot \Pr(\text{no sample in particular } \text{sub rectangle})$$

$$= 4\left(1-\frac{\varepsilon}{4}\right)^{|S|} \le \delta$$

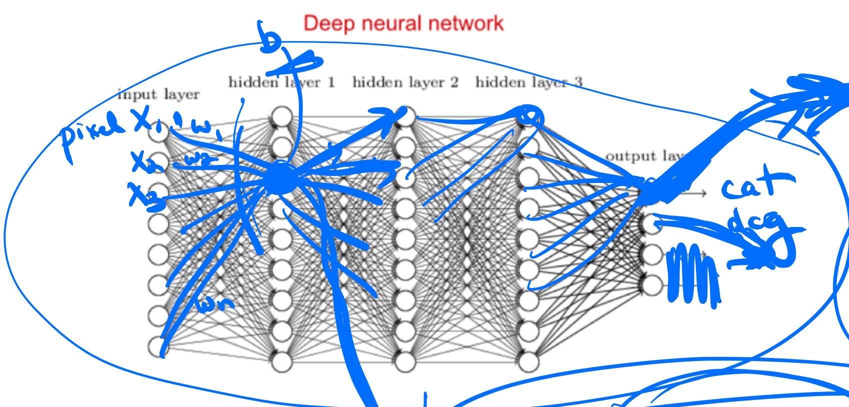
$$\boxed{\frac{4}{\varepsilon}\ln\left(\frac{4}{\delta}\right) \le |S|}$$

---

find $h \in \mathcal{H}$ to minimize $err_S(h)$

optimization problem $\Rightarrow$ $= \frac{1}{m}\sum_{i=1}^{m} loss(h, i^{th} \text{ sample})$

$\ell_i(h)$



given h $\vec{X}$ image.gif

cat(x)
dog(x)

$\ell_i(h) = (cat(x) - 1)^2 + (dog(x) - 0)^2$

$\max\left(\sum w_i x_i + b,\ 0\right)$

Problem:
Find $\boxed{\text{wts \& biases}}$ that minimize $\boxed{err_S(h)}$
h

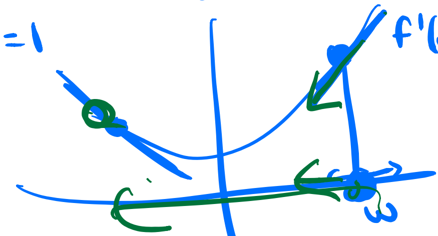
$$\frac{1}{m}\sum_{i=1}^{m} \ell_i(h)$$

Gradient descent

method for "trying" to minimize a fn. $f: \mathbb{R}^n \to \mathbb{R}$

cont, differentiable

$n=1$



$$\boxed{f(w+s) \approx f(w) + s \cdot f'(w)} \qquad \boxed{\frac{f(w+s) - f(w)}{s}} \approx f'(s)$$

$>0 \implies s<0$

$<0 \implies s>0$

$w_0 :=$ arbitrary

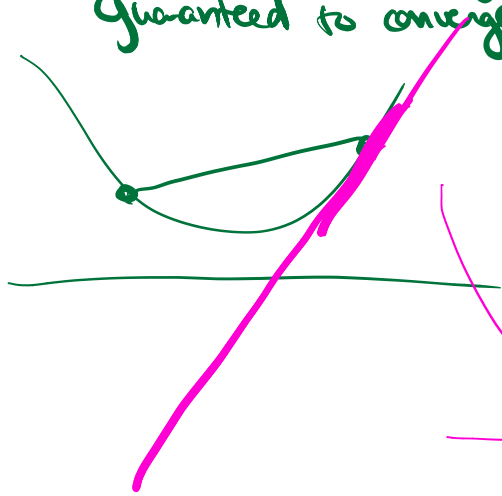
for $t=1, \ldots$ "done"

$w_{t+1} := w_t - \boxed{\eta_t f'(w)}$

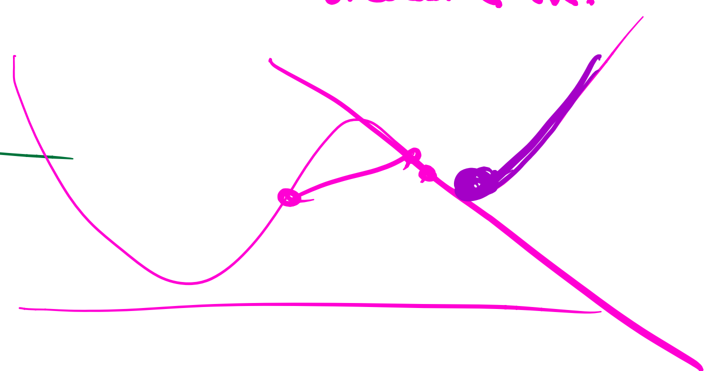
param

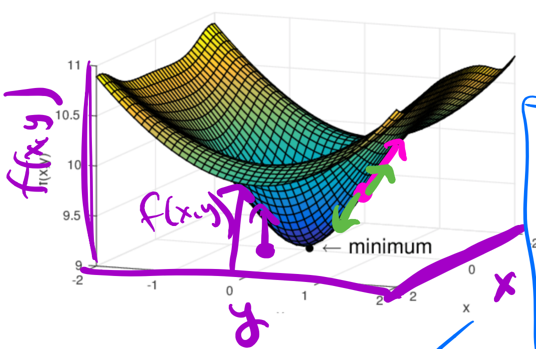
takes opposite direction of derivative

If fn is convex then w/ appropriately selected values of η_t guaranteed to converge to global min



tangent lines globally underestimate fn.





$f: \mathbb{R}^n \to \mathbb{R}$

$f(x,y) = x^2 + 2xy + 4y^2$

$(1,1)$

$\frac{\partial f(x,y)}{\partial x}$ $2x + 2y$

$\frac{\partial f(x,y)}{\partial y}$ $2x + 8y$

$$f(\vec{w} + \vec{s}) \approx f(\vec{w}) + \vec{s} \cdot \nabla f(w)$$

$(w_1, \dots, w_n)$ $(s_1, s_2, \dots, s_n)$

$(w_1 + s_1, w_2 + s_2, \dots)$

gradient of f

$$\begin{pmatrix} \frac{\partial f}{\partial x_1}(\vec{w}) \\ \frac{\partial f}{\partial x_2}(\vec{w}) \\ \vdots \end{pmatrix}$$

$$\nabla f(1,1) = \begin{pmatrix} 4 \\ 10 \end{pmatrix}$$

$$f(1 + s_1, 1 + s_2) \approx f(1,1) + 4s_1 + 10s_2$$

$$(s_1, s_2) \begin{pmatrix} 4 \\ 10 \end{pmatrix}$$

$$\vec{s} \cdot \vec{\nabla f}$$

$$= \|s\| \, \|\nabla f\| \cos\theta$$

-1

∇f

want to move in direction opposite

$$\vec{w}_{t+1} = \vec{w}_t - \eta_t \nabla f(w_t)$$

direction of negative gradient.

gradient descent.

cost of single update

fn of n, m # sample pts

$em_s(w_1, \ldots, w_n)$

fn we're trying to minimize

$f: \mathbb{R}^n \to \mathbb{R}$

$$\frac{1}{m} \sum_{i=1}^{m} \ell_i(\vec{w})$$

error on ith sample pt.

$$\nabla \ldots = \frac{1}{m} \sum_{i=1}^{m} \nabla \ell_i(w)$$

n #wts & biases

m #images

at each t, pick one random image uniformly at random

$I_t \in \{1, \ldots, m\}$

$$E(\nabla \ell_{I_t}(w)) = \sum_{i=1}^{m} \Pr(\text{select } i) \nabla \ell_i(w)$$

$$= \frac{1}{m} \sum_{i=1}^{m} \nabla \ell_i(w)$$

$$= \nabla \text{ loss fn}$$

Define $\underline{\underline{w^*}}$ to be min'g fn $(\mathrm{err}_S(\vec{w}))$

$$\mathbb{E}\left[\mathrm{err}_S(\bar{w}) - \mathrm{err}_S(w^*)\right] \le \sqrt{\frac{RG}{T}} \le \varepsilon$$

consts.

run SGD for T steps

$$\boxed{\underline{\underline{\bar{w}}} = \frac{1}{T}\sum_{t=1}^{T} w_t}$$

$\underline{\underline{\|w_0 - w^*\|^2 \le R}}$ $\qquad$ $\max \underline{\underline{\|\nabla \ell_i(w)\|^2}} \le G$