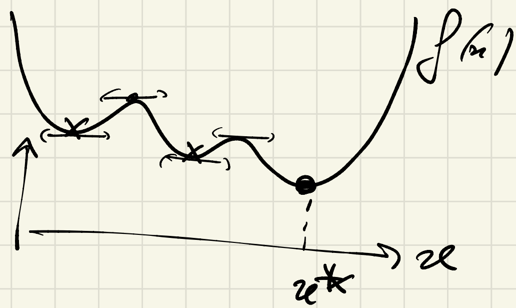
Smooth optimiz°

$$\min_{x\in\mathbb{R}^d} f(x)$$

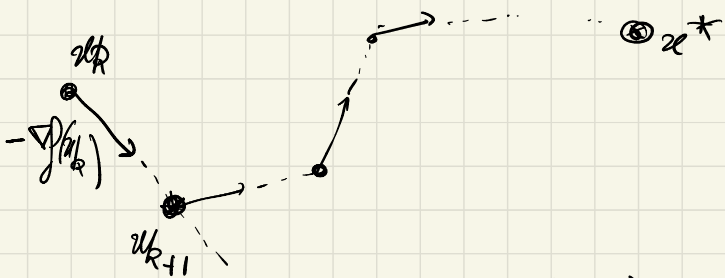


<u>Gradient</u>: $\nabla f(x) \in \mathbb{R}^d$

Gradient Descent : (Batch)

$x_0 \leftarrow$ Init

$$x_{k+1} \triangleq x_k - \tau \cdot \nabla f(x_k)$$



If $x_k \to x^*$

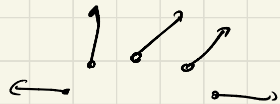
$$x^* = x^* - \tau \nabla f(x^*)$$

$$\Longrightarrow \nabla f(x^*) = 0$$

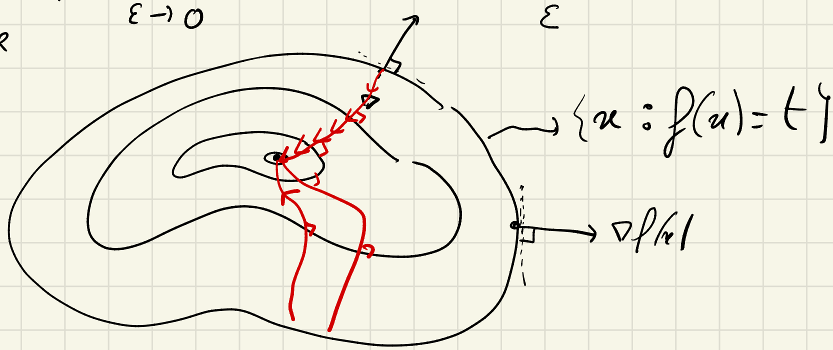
① Gradient : $\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_d}(x) \end{pmatrix} \in \mathbb{R}^d$

$f : x \in \mathbb{R}^d \to f(x) \in \mathbb{R}$

$\nabla f : x \in \mathbb{R}^d \to \nabla f(x) \in \mathbb{R}^d$



$$\frac{\partial f}{\partial x_k}(x) = \lim_{\varepsilon \to 0} \frac{f(\overbrace{x_1, \dots, x_{k-1}, x_k + \varepsilon, x_{k+1}, \dots, x_d}^{x + \varepsilon \delta_k}) - f(x)}{\varepsilon}$$



"Smooth func°" : 1 time differentiable

Def: f is diff. at x if

$$f(x + \varepsilon v) = f(x) + \varepsilon \langle \nabla f(x), v \rangle + \underbrace{o(\varepsilon)}_{\xrightarrow[\varepsilon \to 0]{} 0}$$

$$\Updownarrow$$

$$\frac{f(x + \varepsilon v) - f(x)}{\varepsilon} \xrightarrow{\varepsilon \to 0} \langle \nabla f(x), v \rangle$$

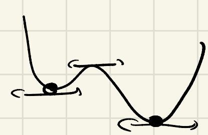
⚠ $\nabla f(x)$ exists $\not\Longrightarrow$ f is diff.

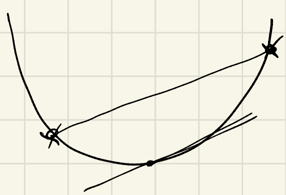
$\Longleftarrow$

$$f(x,y) = \frac{xy(x+y)}{x^2+y^2} \qquad f(0) = 0$$

<u>Thm</u>: if $\nabla f(x)$ for x in a ball around x_0
<u>and</u> $x \mapsto \nabla f(x)$ is continuous
$\Longrightarrow$ f is differentiable

<u>Thm</u>: if x is a local minimizer
$\Longrightarrow$ $\nabla f(x) = 0$





<u>Thm</u>: if f is convex: x global min $\Longleftrightarrow$ $\nabla f(x) = 0$

1D : f cvx $\Leftrightarrow$ $f''(x) \geq 0$ $\forall x$

$\partial^2 f(x) \succcurlyeq 0$

Composi°

$$\underset{\geq 0}{\lambda} \underbrace{f(x)}_{\text{cvx}} + \overset{\geq 0}{\mu} \underbrace{g(x)}_{\text{cvx}} \quad \text{cvx}$$

f cvx, $\quad f(Ax+b)$ cvx

A matrix

$\varphi : \mathbb{R} \to \mathbb{R}$ $\quad \varphi(f(x))$ cvx

$\nearrow$, cvx

$$\varphi(x) = \sum \underline{\underline{\varphi_i}}(\underline{\underline{x_i}})$$

$f : \mathbb{R}^d \to \mathbb{R}^p$ $\quad \underline{\underline{\varphi}} : \mathbb{R}^p \to \mathbb{R}$

$f(x)$ cvx $\quad g(x, \underset{>0}{y}) \overset{\text{def}}{=} \underbrace{y f\left(\frac{x}{y}\right)}$ cvx

Perspective trans.

$$KL(\underbrace{x|y}) = \sum \left(\frac{x_i}{y_i}\right) \log\left(\frac{x_i}{y_i}\right) \times \underline{\underline{y_i}}$$

ML: input: $(a_i, y_i)_{i=1}^{n}$ $\quad a_i \xrightarrow{\psi} y_i$

Supervised $\quad a_i \in \mathbb{R}^d$, $\quad y_i \in \mathbb{R}$, $\{-1,+1\}$ $\quad \psi(a) = y$.

$\psi(a_i) \approx y_i$

linear model $\quad \psi(a) = \langle a, \underbrace{x}_{\text{weight / need train}} \rangle$, $\quad x = \theta$

Regression: ERM

$$\min_{x \in \mathbb{R}} \sum_{i=1}^{n} \ell(\langle a_i, x\rangle, y_i) = f(x), \quad \langle a_i, x\rangle \in \mathbb{R}, \ y_i \in \mathbb{R}$$

least square $\quad \ell(y, y') = (y - y')^2$

Rmq: if ℓ is convex, f is convex

Design matrix: $\quad A = \begin{pmatrix} \text{---} \ a_i \ \text{---} \\ \\ \\ \end{pmatrix} \Big\} n$, $\quad \underbrace{\qquad}_{d}$

$A \in \mathbb{R}^{n \times d}$, $x \in \mathbb{R}^d$

$$Ax = \begin{pmatrix} \langle a_1, x\rangle \\ \langle a_2, x\rangle \\ \vdots \\ \langle a_n, x\rangle \end{pmatrix}$$

$$\min_x f(x) = L(\overset{u}{Ax})$$

$$L(u) = \sum_{i=1}^{m} \ell(u_i, y_i)$$

Least square: $L(u) = \sum_{i=1}^{m} (u_i - y_i)^2 = \|u - y\|^2$

$$\min f(x) = \|\underline{\underline{A}}x - y\|^2$$

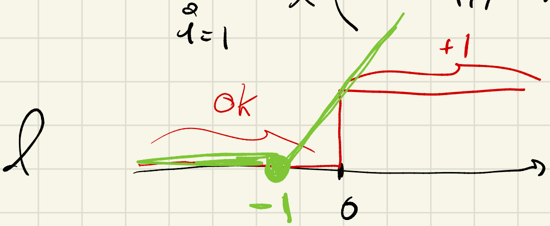
Class: $y_i \in \{-1, +1\}$

predicta: $\text{sign}[\langle a, x\rangle] \in \{-1, 1\}$.

Ultimate goal: 0/1 loss.

$$\min_x \sum_{i=1}^{n} \text{Error}(\text{sign}(\langle a_i, x\rangle), y_i).$$

$$\sum_{i=1}^{m} \ell(-\langle a_i, x\rangle y_i) = \#\text{error}$$

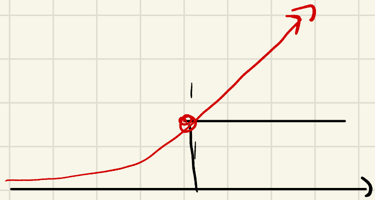


$\ell(s) = \mathbb{1}_{\mathbb{R}^+}(s)$

① SVM: ℓ = HINGE LOSS $\quad \ell(s) \triangleq (s+1)_+$

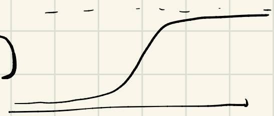
f is non smooth

② Logistic loss / Cross entropy loss



$$\ell(s) \triangleq \frac{\log(1+e^s)}{\log(2)}$$

$$\ell'(s) = \frac{e^s}{1+e^s} \in [0,1]$$



$$f(x) = \sum_i \ell\Big(\underbrace{-y_i \underbrace{\langle a_i, x\rangle}_{(Ax)_i}}_{-\mathrm{diag}(y)\cdot A\cdot x}\Big) = L(\underbrace{-\mathrm{diag}(y)A}_{B}\, x) = L(Bx)$$

$$L(u) = \sum_{i=1} \ell(u_i)$$

Gradient comput°: $\min_x f(x) = L(Bx)$.

$$\left[\; x \in \mathbb{R}^d \xrightarrow[B \in \mathbb{R}^{n \times d}]{} Bx \in \mathbb{R}^n \xrightarrow[\substack{L \\ \text{diff.}}]{} L(Bx) \in \mathbb{R}\right.$$

"Proof":
$$f(x+\varepsilon v) = (\dots) = f(x) + \varepsilon \langle \underset{\substack{\| \\ \nabla f(x)}}{??}, v \rangle + o(\varepsilon)$$

$$L(B(x+\varepsilon v)) = L(Bx + \varepsilon Bv)$$ L is diff at Bx

$$\overset{(*)}{=} \underbrace{L(Bx)}_{f(x)} + \varepsilon \langle \nabla L(Bx), Bv \rangle + o(\varepsilon)$$

Def: $B = (B_{ij})_{ij}$, $B^T = (B_{ji})_{ij}$, $B \in \mathbb{R}^{n \times d}$, $B^T \in \mathbb{R}^{d \times n}$

Prop: $\langle Bu, v \rangle_{\mathbb{R}^n} = \langle u, B^T v \rangle_{\mathbb{R}^d}$

$$f(x+\varepsilon v) \overset{(*)}{=} f(x) + \varepsilon \langle \underbrace{B^T \cdot \nabla L(Bx)}_{= \nabla f(x)}, v \rangle + o(\varepsilon)$$

Prop:
$$\nabla f(x) = \nabla (L \circ B)(x)$$
$$= B^T \times \nabla L(Bx)$$

$$\left[\nabla (L \circ B) = B^T \circ \nabla L \circ B \right]$$

Examples: $L(u) = \frac{1}{2} \| u - y \|^2 = \frac{1}{2} \sum (u_i - y_i)^2$

$$\nabla L(u) = u - y$$

$$\nabla\left(\frac{1}{2}\|\cdot\|^2\right)(u) = u$$

$$f(x) = \frac{1}{2}\|Ax - y\|^2$$

$$\nabla f(x) = A^T(Ax - y)$$

$$\min_x \|Ax-y\|^2 \iff A^T(Ax-y) = 0 = \nabla f(x) = 0$$

$$\iff \underbrace{(A^TA)}_{\text{Covariance } \in \mathbb{R}^{d\times d}} x = A^T y \quad (\text{normal eq}^\circ)$$

⚠ $Ax = y$

$\to \det \neq 0 \iff \text{Ker}(A) = \{0\}$.

if A^TA is invertible ("$n \gg d$"), unique sol$^{\underline{o}}$ "Overdetermined"

$$\boxed{x = \underbrace{(A^TA)^{-1}A^T}_{A^+} y}$$

A^+ Moore-Penrose Pseudo-Inverse

if A^TA not invertible ($d \gg n$) ∞ possibility

- Ridge
- Lasso

Logistic: $L(u) = \sum_i \ell(u_i)$

$$\ell(s) = \log(1 + e^s).$$

$$\ell'(s) = \frac{e^s}{1+e^s}$$

$$\nabla L(u) = \begin{pmatrix} \ell'(u_1) = \frac{e^{u_1}}{1+e^{u_1}} \\ \ell'(u_2) \quad \vdots \\ \vdots \\ \ell'(u_n) = \frac{e^{u_n}}{1+e^{u_n}} \end{pmatrix} = \text{Sigmoid}(u)$$

$$\nabla f(x) = B^T \, \nabla L(Bx) = \underbrace{B^T}_{\text{Back Prop}} \times \text{Sigmoid}(\underline{Bx})$$

Grad descent: $x_{k+1} = x_k - t_k \nabla f(x_k)$

t_k ↳ step size
↳ learning rate

t_k large → fast.

t_k small → avoid explosion

<u>Baby case</u>: $f(x) = \frac{1}{2}\|Ax - y\|^2$

$$\underset{\mathbb{R}^{d\times d}}{C} \triangleq A^T A \qquad C_{kl} = \text{" how much feature } k, l \text{ correlated "}$$

C is a symetric matrix: $C^T = C$

$$C^T = (A^T A)^T = A^T (A^T)^T = A^T A = C$$

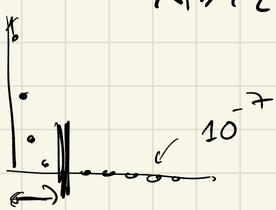
$(AB)^T = B^T A^T$

<u>Thm</u>: since C is symetric, $(\underset{\in \mathbb{R}^d}{u_1}, \dots, \underset{\in \mathbb{R}^d}{u_d})$ eigenvectors

$(\underbrace{\lambda_1}_{\geq 0}, \dots, \underbrace{\lambda_d}_{\geq 0})$ eigenvalues

$$C u_i = \underbrace{\lambda_i}_{\geq 0} u_i$$

$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$.



$$\underset{A^T A}{C} u_i = \lambda_i u_i \Rightarrow \langle A^T A u_i, u_i \rangle = \lambda_i \langle u_i, u_i \rangle$$

$$\langle A u_i, A u_i \rangle = \lambda_i \langle u_i, u_i \rangle$$

$$\|A u_i\|^2 = \lambda_i \|u_i\|^2$$

$$\lambda_i = \frac{\|A u_i\|^2}{\|u_i\|^2} \geq 0$$

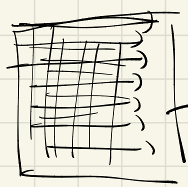
Positive Semi-definite Matrix SDP matrices.

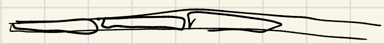
Thm: For gradient descent if

$$0 < t_k < \frac{2}{\lambda_d}$$

then $x_k \to$ sol^o $x^* = (A^TA)^{-1}A^Ty$.

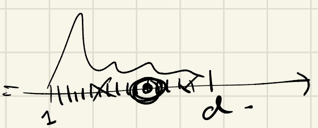
$$A = \begin{pmatrix} \text{---}\ a_1\ \text{---} \\ \text{---}\ a_2\ \text{---} \\ \vdots \\ \text{---}\ a_n\ \text{---} \end{pmatrix} \qquad C_{kl} = \frac{1}{n}\sum_{i=1}^{n} a_i[k] \cdot a_i[l] \approx \mathbb{E}_a\big(a[k]\,a[l]\big)$$

$a_i =$  $\to p_1 \times p_2 \to n = p_1 \times p_2$.

$\to a_1$.flatten() 

$$a_i = (a_i[1], a_i[2] \dots a_i[d])$$

$a_i = (\text{weight}, \text{height}, \text{age}, \dots)$

$a =$ 

$C =$