1) Grad desc.

2) Regulariz° → Ridge
→ Lasso ... Nucl. norm
TV
...

3) Non smooth convex optim → Large scale
↳ proximal

~~Interior point~~

GD : $\min_{x \in \mathbb{R}^d} f(x)$ $\qquad x_{k+1} \overset{(*)}{=} x_k - \tau \nabla f(x_k)$.

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

① $x_{k+1} = x_k - \tau A^T(Ax_k - y)$

if x^* sol° $\nabla f(x^*) = 0$ $\quad A^T(Ax^* - y) = 0$

② $x^* = x^* - \tau A^T(Ax^* - y)$

①-② $\underbrace{x_{k+1} - x^*}_{\varepsilon_{k+1}} = \underbrace{x_k - x^*}_{\varepsilon_k} - \tau A^T A(\underbrace{x_k - x^*}_{\varepsilon_k})$

$$\varepsilon_{k+1} = \underbrace{\left(Id - \tau A^T A\right)}_{U_\tau} \varepsilon_k$$

<u>Cond°</u> : $\varepsilon_k = \left(U_\tau\right)^k \varepsilon_0$

<u>Q°</u>: $\|U_\tau\|_{op} < 1$??

$\|U_\tau\|_{op}$ = operator norm   np.linalg.norm

<u>Def</u>: If B is a matrix

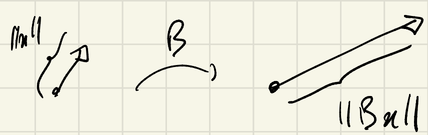
$$\|B\|_{op} = \sqrt{\lambda_{max}(B^T B)} = \sigma_{max}(B)$$

If B is sym. $B^T = B$

$$\|B\|_{op} = \max_i |\lambda_i(B)|$$

Why "operata"? $\|Bx\|_2 \le \|B\|_{op} \cdot \|x\|_2$

Lipschitz constant



<u>Q°</u>: find τ s.t. $\max_i \left|\lambda_i\left(Id - \tau A^T A\right)\right| \overset{(*)}{<} 1$

Theorem: • if $\tau < \frac{2}{\|A\|_{op}^2}$ then it's true $(*)$

• Overdetermined A^TA is invertible.

$$0 < \underbrace{\mu = \lambda_{min}(A^TA)}_{\mu \le \lambda_i(A^TA)} \le \underbrace{\lambda_{max}(A^TA) = L}_{\le L}$$

$\underline{A^TA}$ correl°

if $\mu > 0$, then fast ("linear") convergence

the optimal $\tau = \frac{2}{\mu + L}$

"LINEAR" ⟷ Geometrical

$$\|x_k - x^*\| \le \left(\frac{L-\mu}{\mu+L}\right)^k \|x_k - x_0\|$$

$\mu > 0$ $\quad < 1$

$$\frac{L-\mu}{L+\mu} = \frac{(L/\mu)-1}{(L/\mu)+1}$$

$\frac{L}{\mu} = \kappa \geq 1$ conditionning

If $\mu = 0$ $\quad \tau \leq \frac{2}{L} \longrightarrow$ converge.

<u>Thm</u>: $\tau \leq \frac{2}{L}$ $\qquad f(x) = \frac{1}{2}\|Ax - y\|^2$

$$f(x_k) - f(x^*) \leq \frac{f(x_0) - f(x^*)}{k}$$

"SUB-LINEAR"

<u>General Case</u> : $f(x)$ convex $F^{\underline{o}}$.

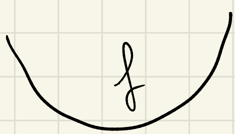
$A^T A \longrightarrow$ Hessian of f

$$\partial^2 f(x) = \left(\frac{\partial f(x)}{\partial x_i \, \partial x_j} \right)_{i,j=1}^{d} \in \mathbb{R}^{d \times d}$$

<u>Prop</u>$^{\underline{o}}$: $\partial^2 f(x)$ symmetric

f is convex $\iff \partial^2 f(x) \succeq 0$

eigenvalues



$f'' \geq 0$

Thm: (f twice diff)

$$\mu = \inf_x \inf_i \lambda_i(\partial^2 f(x))$$

$$L = \sup_x \sup_i \lambda_i(\partial^2 f(x))$$

$\frac{L}{\mu} = \kappa$ cond.

① if $\tau < \frac{2}{L} \longrightarrow$ convergence

② if $\kappa < +\infty$ $(\mu > 0)$ $\longrightarrow$ Fast convergence (Linear).

Stochastic opt^{c}:

$$\min_x f(x) = \frac{1}{n} \sum_i f_i(x).$$

$i \to$ data

$$\nabla f(x) = \underbrace{\frac{1}{n} \sum_i \nabla f_i(x)}_{\text{too slow}} \rightsquigarrow \hat{\nabla} f(x) = \frac{1}{|I|} \sum_{i \in I} \nabla f_i(x)$$

Random

$$\mathbb{E}(\hat{\nabla} f(x)) \overset{\text{CRUCIAL}}{=} \nabla f(x) \quad \text{"Unbiased"}$$

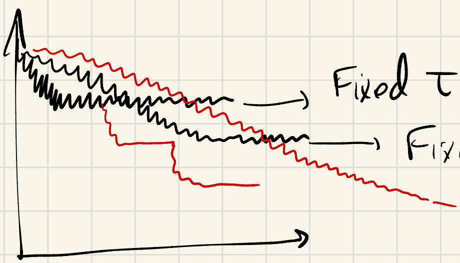
SGD: $\underline{\underline{x_{k+1}}} = x_k - \tau_k \hat{\nabla} f(x)$. ⚠ RANDOM.

D $\to$ "Descent" FALSE

Stochastic Approx:
ROBIN & MONROE

<u>Thm</u>: $\tau_k \searrow 0$ for instance $\tau_k = \frac{1}{k}$

$x_k \longrightarrow x^*$ almost everywhere.



→ Accelerat° / Momentum → Extrapolat°
└→ Nesterov / Heavy Ball

GD : $O(1/k)$ ⟶ Nesterov $\underbrace{O(1/k^2)}_{\text{OPTIMAL}}$.

Regulariz°: $\min_{x} \frac{1}{2}\|Ax-y\|^2$

Pbm: if $n < d$. (underdetermined).

- → non unique sol° $Ker(A) = \{0\}$.
- → Overfitting

Ridge regul° / Tichonov / Weight decay

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}\|Ax-y\|^2_{\mathbb{R}^n} + \underbrace{\frac{\lambda}{2}\|x\|^2_{\mathbb{R}^d}}_{\text{Ridge penalty}} = f(x).$$

$C = 1/\lambda$

λ = Lagrange mult.

strong overfit, large noise $\}$ → $\lambda \nearrow$ → Favor "small" x

Least square + ridge :

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

$$\nabla f(x) = 0 \iff A^TAx + \lambda x = A^Ty.$$

$$\iff (\underbrace{A^TA}_{\geq 0} + \underbrace{\lambda \, Id_{\mathbb{R}^d}}_{>0})x = A^Ty.$$

Invertible ALWAYS!

Ridge → unique sol$^{\underline{o}}$.

Cond$^{\underline{o}}$: $x_\lambda = (A^TA + \lambda \, Id_d)^{-1} A^T y$

Thm : (WOODBURY formula).

$$(\underbrace{A^TA}_{=C} + \lambda \, Id_d)^{-1} A^T = A^T(\underbrace{AA^T}_{=K} + \lambda \, Id_n)^{-1}$$

covell$^{\underline{o}}$ Kernel

Feature space $(\mathbb{R}^d)$ Kernel space $(\mathbb{R}^n)$.

If $d \gg n$ → GO KERNEL!

Works $d = +\infty$ Kernel (RKHS)

[Reproducing Kernel Hilbert Space.

LASSO : regalizer ℓ^1 norm

$\|x\|_1 = \sum_i |x_i|$. $\neq$ $\sum_i |x_i|^2$ RIDGE

$\hookrightarrow \min_x \frac{1}{2}\|Ax - y\|^2 + \lambda \|x\|_1$.

"ℓ^1 promotes lots of zeros in x^*"

$\hookrightarrow$ sparsity

→ modeling : image processing.

→ model selection / Explainable.

$x \in \mathbb{R}^d$ d very large.

select only a "few" feature $x_i = 0$ for a lot of i's.

Lasso does a selection

LASSO : $x_\lambda = \arg\min_x \frac{1}{2}\|Ax-y\|^2 + \lambda\|x\|_1$

$$\underbrace{\min_x \frac{1}{2}\|Ax-y\|^2}\xrightarrow{\text{true}} \infty \text{ \# of sol}^{\underline{o}}$$

$d \gg n \qquad \mathrm{Ker}(A) \neq \{0\}$.

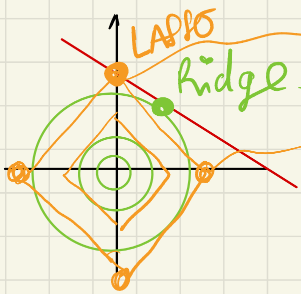
$\mathcal{H} = \{x : Ax = y\}$.

x_λ sol$^{\underline{o}}$ of LASSO

Thm : $x_\lambda \xrightarrow{\lambda \to 0}$ to a sol$^{\underline{o}}$ of . Basis Pursuit

$$\min_{Ax=y} \|x\|_1$$

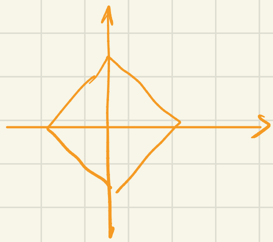
$n = 1 \quad d = 2$. $H = \{Ax = y\}$ line.



$x = (0, x_2)$

Ridge. $x = (x_1, x_2)$

$x = (x_1, 0)$



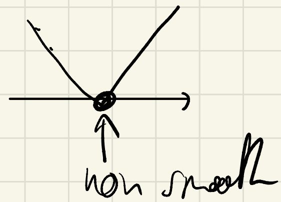
Ridge : $\min_{Ax=y} \|x\|_2^2$ $\qquad$ lasso. $\min_{Ax=y} \|x\|_1$

Good: x sol° is sparse.

$$f(x) = \frac{1}{2}|Ax-y|^2 + \lambda |x|_1 .$$

is convex

Bad: f is non diff. $|x|$



Support vector machine (Hinge loss): same.

<u>Idea</u>: splitting.

<u>1 part</u>: "Proximal operator" ℓ^1.

$A = \mathrm{Id}$.

<u>Def</u>: For some func° $g(x)$ (ex. ℓ^1).

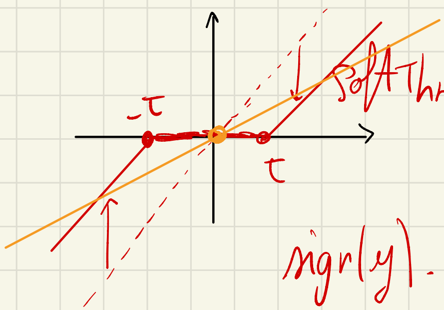
$$\mathrm{Prox}_{\tau g}(y) \overset{def}{=} \underset{(x)}{\mathrm{argmin}} \ \frac{1}{2}\|x-y\|^2 + \tau g(x)$$

"Resolvant" operator

Ridge : $\mathrm{Prox}_{\tau \|.\|_2^2/2}(y) = \arg\min_x \frac{1}{2}\|x-y\|^2 + \frac{\tau}{2}\|x\|^2$

$$0 = x - y + \tau x \Rightarrow y = \frac{x}{1+\tau}$$

Lasso : $\mathrm{Prox}_{\tau\|.\|_1}(y) = \mathrm{Soft\,thresh}_\tau(y) = \left(\mathrm{SoftThresh}_\tau(y_i)\right)_{i=1}^d$



$\mathrm{SoftThresh}_\tau(y)$

$\|$

$\mathrm{sign}(y) \cdot \max(|y| - \tau, 0)$

ISTA : Iterative Soft Thresholding.

$\ell 1$

(~2003 Daubechies / De Mol / De Frise).

[Special case] Forward - Backward algo.

Step ISTA

- Gradient Descent on $\frac{1}{2}\|Ax-y\|^2$ with step τ

  $\tilde{x}_k := x_k - \tau A^T(Ax_k - y) = x_k - \tau[Cx_k - \mu]$

  $\mu = A^T y$
  $C = A^T A$

- Soft Threshold $\lambda\tau$

  $x_{k+1} := \mathrm{SoftThreshold}_{\lambda\tau}[\tilde{x}_k]$

$k \leftarrow k+1$

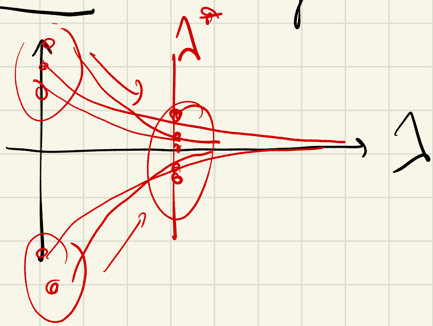
Thm : If $\tau < \dfrac{2}{\|A\|_{op}^2}$, then $x_k \to x^*$ sol° LASSO.

$$f_\lambda(x_k) - f_\lambda(x^*) \sim \left[\frac{1}{k}\right]$$

Accelerate ~~ISTA~~ $O(1/k)$ $\xrightarrow{}$ <u>FISTA</u> $O(1/k^2)$.

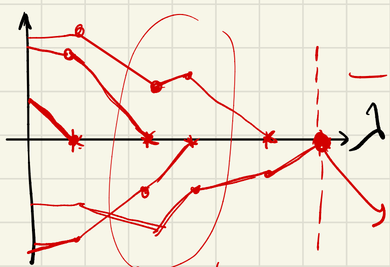
$\uparrow$ Nesterov — optimal

Regul° path : influence $\underline{\lambda}$



$$x_\lambda = (\underbrace{x_\lambda^1, x_\lambda^2 \dots x_\lambda^d}_{d \text{ feature}})$$

$$\lambda \mapsto x_\lambda^i$$



$\to$ sol° = 0

$\lambda_{max} = \|A^T y\|_\infty$

$\lambda_{max}/10$