

REGULARIZATION

- The problem $\min_x \|Ax - y\|_2^2$
 - is **ill-conditioned** (sensitive to noise) if $K(A) = \frac{\sigma_1}{\sigma_n}$ is large.
 - has ∞ solutions if $\text{Ker } A \neq \{0\}$.
- Regularization = replace original problem with a **close**, **well-posed** one. Boils down to enforcing structure on the solution.
- Strategy: balance $\|Ax - y\|^2$ with a **prior**

data fidelity prior

$$\text{low } \|Ax - y\|^2 + \text{low } J(x)$$

PENALIZED PROBLEM

- Standard approach ($\tilde{x} = A^+ \tilde{y}$)

$$\text{minimize } \frac{1}{2} \|Ax - y\|_2^2$$

- Regularized problem

$$\text{minimize } \frac{1}{2} \|Ax - y\|_2^2 + \gamma J(x) \quad (P_\gamma)$$

with : regularizer J usually convex

regularization parameter $\gamma > 0$

γ small \rightarrow solution will fit measurements well

γ large \rightarrow solution will be structured

⚠ tuning the parameter γ is a difficult task

CONSTRAINED FORMULATION

- Equivalent formulation

$$\text{minimize } \frac{1}{2} \|Ax - y\|^2 \text{ subject to } J(x) \leq \tau \quad (C_\tau)$$

Prop : $\forall \lambda, \exists \tau(\lambda)$: a solution of (P_λ) is also a solution of (C_τ) .

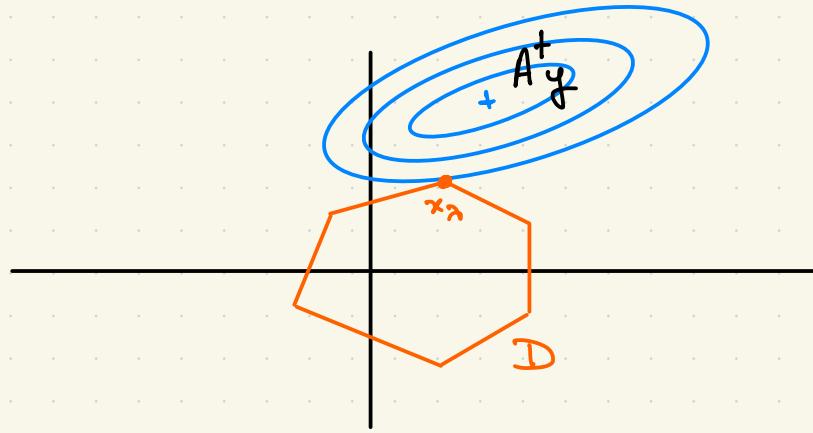
 $\forall \tau, \exists \lambda(\tau)$: a solution of (C_τ) is also a solution of (P_λ) .

Rank : determining $\tau(\lambda)$ or $\lambda(\tau)$ is not obvious.

- Also equivalent to

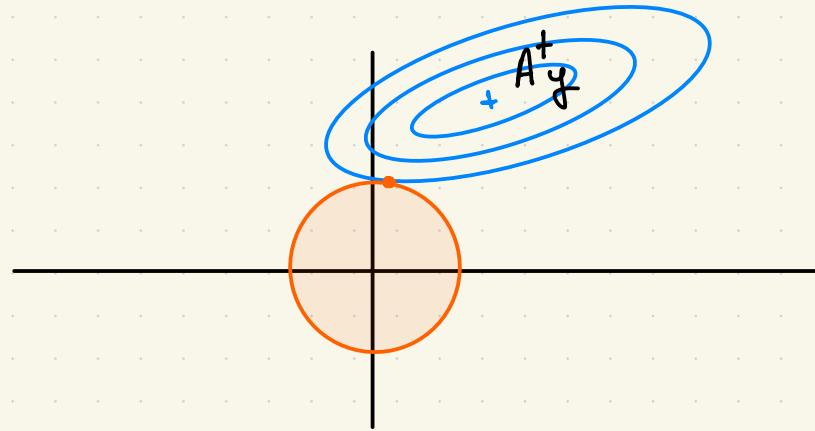
$$\text{minimize } J(x) \text{ s.t. } \|Ax - y\| \leq \varepsilon$$

GEOMETRIC PICTURE



$$D = \{x / J(x) \leq 1\} \text{ convex}$$

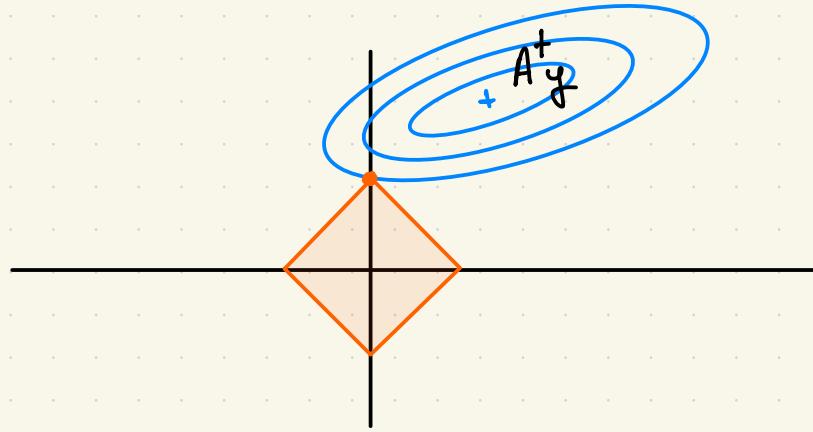
GEOMETRIC PICTURE



$$D = \{x / J(x) \leq 1\} \text{ convex}$$

e.g. $J(x) = \frac{1}{2} \|x\|^2$

GEOMETRIC PICTURE



$$D = \{x / J(x) \leq 1\} \text{ convex}$$

e.g. $J(x) = \frac{1}{2} \|x\|_2^2$, $J(x) = \|x\|_1$

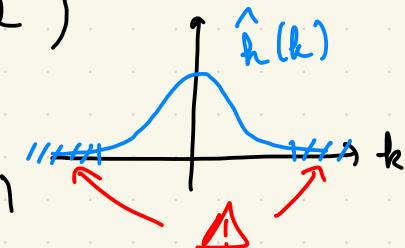
TIKHONOV REGULARIZATION

- Example : image deblurring / denoising

$$y = a * x_0 + e \quad (\text{a blur kernel})$$

$$\hat{y}(k) = \hat{a}(k) \hat{x}_0(k) + \hat{e}(k)$$

$$\hat{x}(k) = \hat{y}(k) / \hat{a}(k) = \hat{x}_0(k) + \hat{e}(k) / \hat{a}(k)$$



- Tikhonov regularization : push $\|x\|_2$ down

(aka ridge regression)

$$\text{minimize} \quad \frac{1}{2} \|Ax - y\|_2^2 + \frac{\lambda}{2} \|x\|_2^2 \quad (T_\lambda(y))$$

$$\bullet \text{Let } E_\lambda(x) := \frac{1}{2} \|Ax - y\|_2^2 + \frac{\lambda}{2} \|x\|_2^2$$

E_λ is differentiable, and strictly convex. This will allow us to give the solution explicitly

First-order optimality : x_* solution of $(T_\lambda) \iff \nabla E_\lambda(x_*) = 0$

DIFFERENTIAL CALCULUS

- Definition : let X, Y be normed spaces, and $\Omega \subset E$ an open set. A function $f: \Omega \rightarrow Y$ is differentiable at $a \in \Omega$ if there exists $L \in \mathcal{L}(X, Y)$ such that

$$f(a+h) - f(a) = L(h) + o(\|h\|)$$

If it exists, L is unique. We denote it $f'(a)$.

- Gradient : $f: \mathbb{R}^m \rightarrow \mathbb{R}$ (scalar-valued function of a vector variable). There exists a unique vector $\nabla f(a) \in \mathbb{R}^m$ such that $f'(a) \cdot h = \langle \nabla f(a), h \rangle$. In an s.b.

$$\nabla f(a) = (\partial_1 f(a), \dots, \partial_m f(a)) \text{ where } \partial_i f(a) \text{ partial derivative.}$$

- Example : $f(x) = \|x\|_2^2$

$$\|x+h\|_2^2 = \|x\|_2^2 + \underbrace{2 \langle x, h \rangle}_{\text{linear}} + \underbrace{\|h\|^2}_{o(\|h\|^2)}, \text{ hence } \nabla f(x) = 2x$$

DIFFERENTIAL CALCULUS 2

- Jacobian : $f: \mathbb{R}^m \rightarrow \mathbb{R}^m$ (vector-valued function of a vector variable). Then $f'(a) \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^m)$ is represented in the canonical bases by the Jacobian matrix

$$J(a) = \begin{pmatrix} \frac{\partial_1 f_1(a)}{\partial_1 x_1} & \dots & \frac{\partial_m f_1(a)}{\partial_m x_1} \\ \vdots & & \vdots \\ \frac{\partial_1 f_m(a)}{\partial_1 x_m} & \dots & \frac{\partial_m f_m(a)}{\partial_m x_m} \end{pmatrix} \quad \text{where } f(x_1, \dots, x_m) = \begin{pmatrix} f_1(x_1, \dots, x_m) \\ \vdots \\ f_m(x_1, \dots, x_m) \end{pmatrix}$$

- Example : $f(x) = Ax - y$ (linear)

$$A(x+h) - y = Ax - y + \underbrace{Ah}_{\text{linear}} \text{, hence } J(x) = A$$

- Chain rule : $(f \circ g)'(a) = f'(g(a)) \circ g'(a)$

Example : $f(x) = \|Ax - y\|_2^2$,

$$f'(x) \cdot h = \langle 2(Ax - y), Ah \rangle, \quad \nabla f(x) = 2A^T(Ax - y)$$

NORMAL EQUATIONS

$$\|\tilde{A}x - \tilde{y}\|^2$$

- $E_\lambda(x) = \frac{1}{2} \|Ax - y\|_2^2 + \frac{\lambda}{2} \|x\|_2^2$

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

- Prop : Problem (T_λ) is equivalent to

$(A^T A + \lambda I_m) x = A^T y$ (*)

- In fact, (T_λ) is equivalent to a modified least-squares :

minimize $\left\| \underbrace{\begin{pmatrix} A \\ \sqrt{\lambda} I_m \end{pmatrix}}_{\tilde{A}} x - \underbrace{\begin{pmatrix} y \\ 0 \end{pmatrix}}_{\tilde{y}} \right\|_2^2$

and (*) corresponds to the normal equations for this system

SOLUTION OF (T_λ)

$$A^{-1} = V^T (\Sigma^2 + \lambda I_n)^{-1} V$$

- Prop : let $A \in \mathbb{R}^{m \times n}$. For any $\lambda > 0$, the matrix

$A^T A + \lambda I_m$ is invertible.

proof : with $A = U \Sigma V^T$, $A^T A + \lambda I_m = V (\Sigma^2 + \lambda I_n) V^T$. Since $\lambda > 0$, $\Sigma^2 \succeq 0$, $(\Sigma^2 + \lambda I_n)$ is invertible and so is $(A^T A + \lambda I_m)$.

- Prop : (T_λ) has a unique solution, given by

$$x_\lambda = (A^T A + \lambda I_m)^{-1} A^T y$$

$$A^+ y$$

- Rmk : We always have that x_λ depends continuously on y

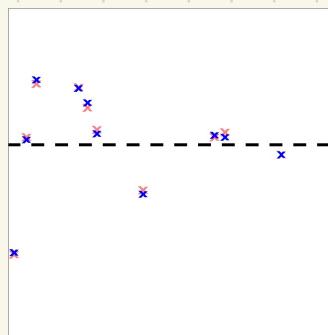
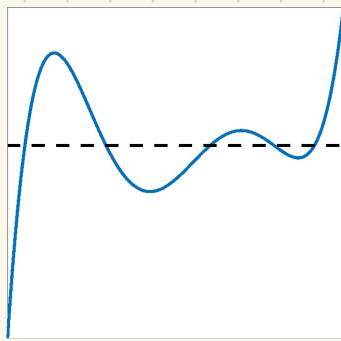
since $\langle (A^* A + \lambda I_n) x, x \rangle = \|A x\|^2 + \lambda \|x\|^2$

$$= \langle A^* y, x \rangle \leq \|A^* y\| \|x\|$$

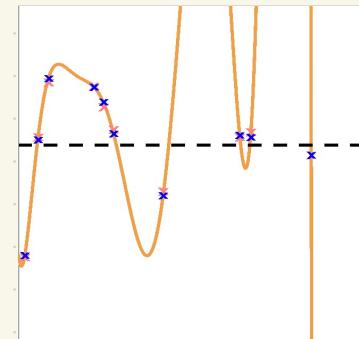
hence $\|x_\lambda\| \leq \frac{1}{\lambda} \|A^*\| \|y\|$

Cauchy-Schwarz

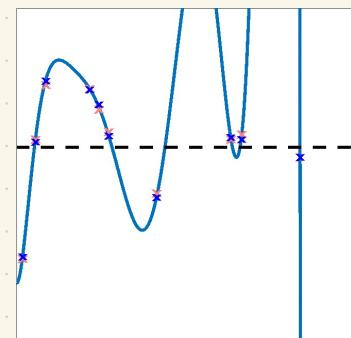
EXAMPLE



{



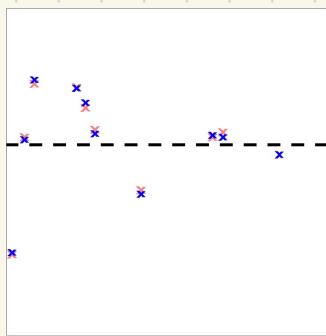
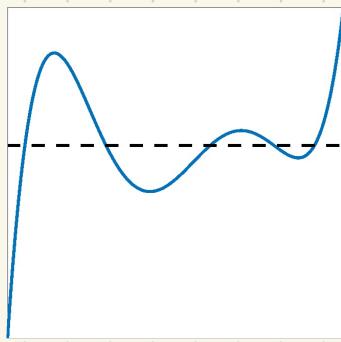
$$A^+ y$$



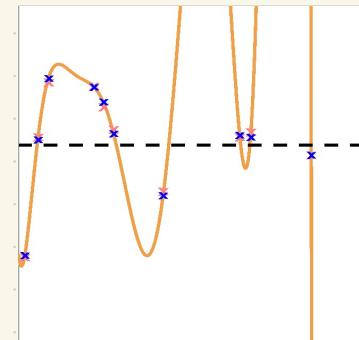
$$(A^T A + \gamma I)^{-1} A^T y$$

$$\gamma = 10^{-5}$$

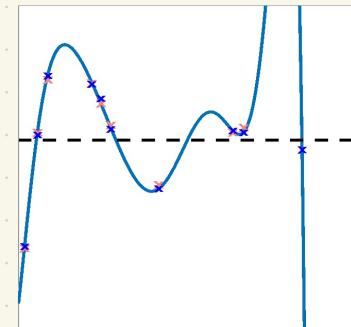
EXAMPLE



{



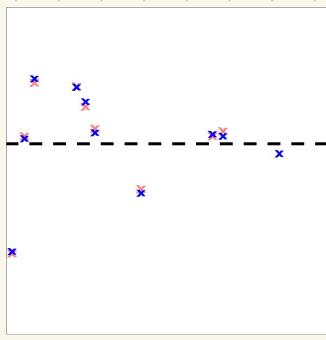
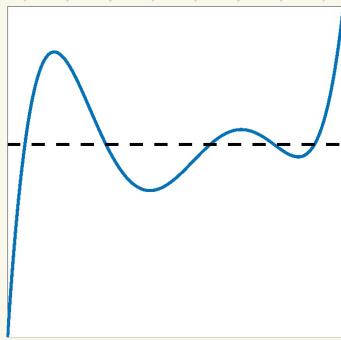
$$A^+ y$$



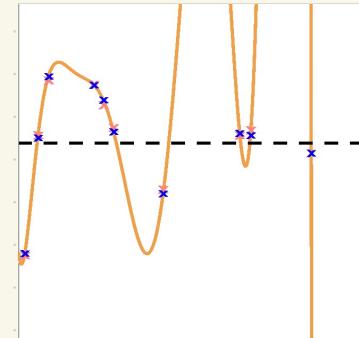
$$(A^T A + \gamma I)^{-1} A^T y$$

$$\gamma = 10^{-4}$$

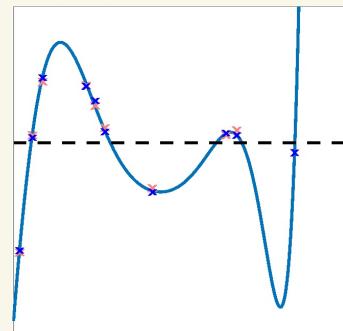
EXAMPLE



{



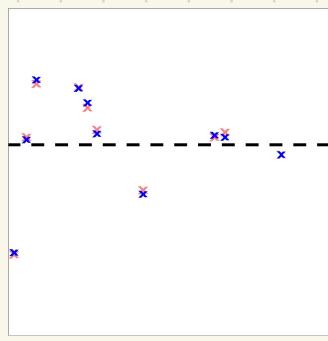
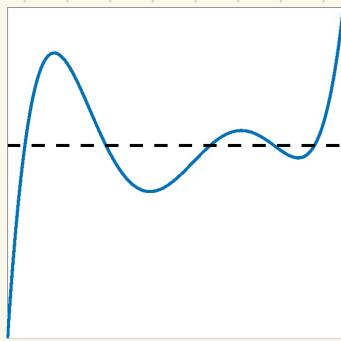
$$A^+ y$$



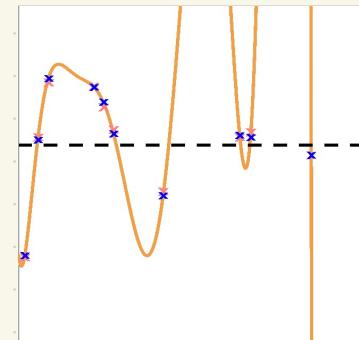
$$(A^T A + \gamma I)^{-1} A^T y$$

$$\gamma = 10^{-3}$$

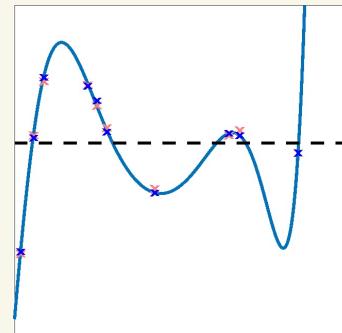
EXAMPLE



{



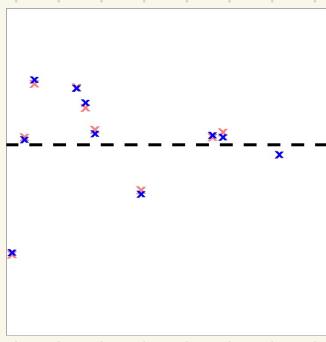
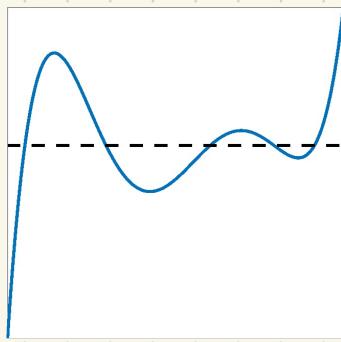
$$A^+ y$$



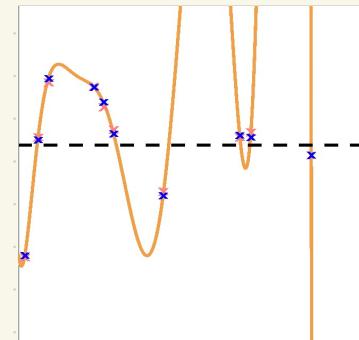
$$(A^T A + \gamma I)^{-1} A^T y$$

$$\gamma = 10^{-2}$$

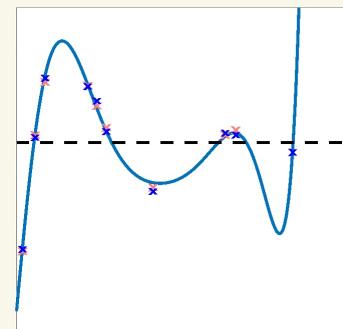
EXAMPLE



{



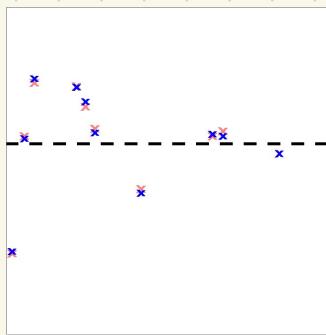
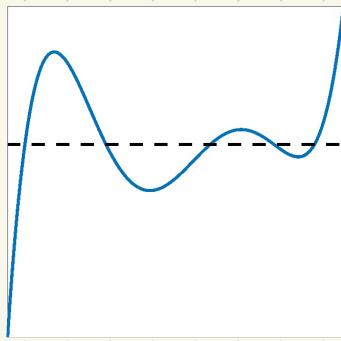
$$A^+ y$$



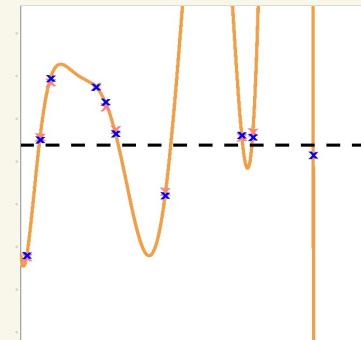
$$(A^T A + \gamma I)^{-1} A^T y$$

$$\gamma = 10^{-1}$$

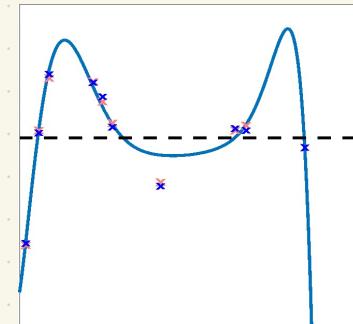
EXAMPLE



{



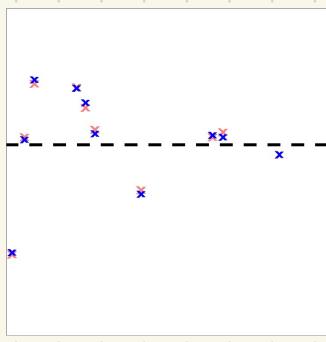
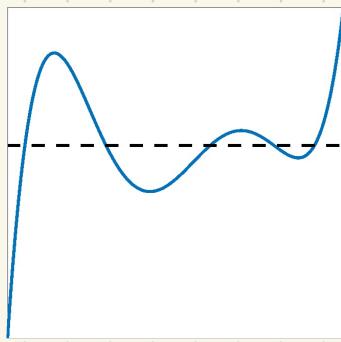
$$A^+ y$$



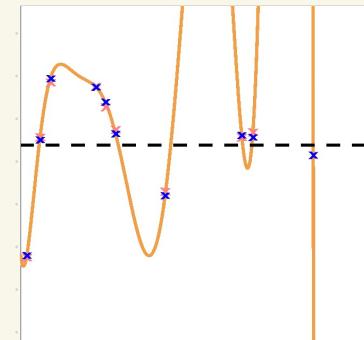
$$(A^T A + \gamma I)^{-1} A^T y$$

$$\gamma = 1$$

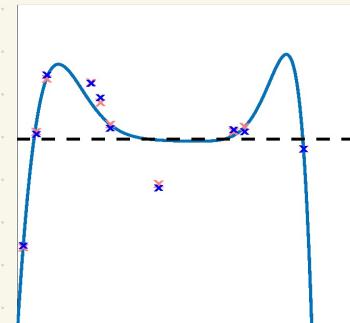
EXAMPLE



{



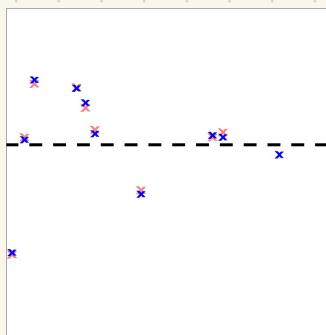
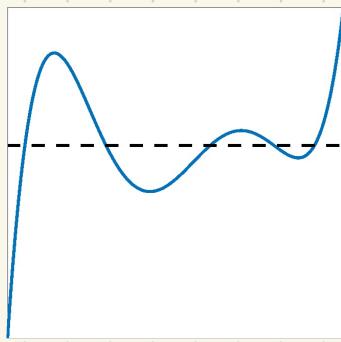
$$A^+ y$$



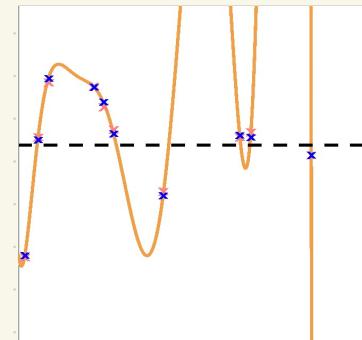
$$(A^T A + \gamma I)^{-1} A^T y$$

$$\gamma = 10$$

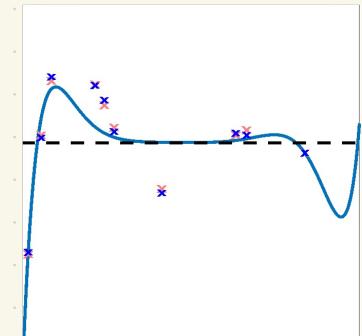
EXAMPLE



{



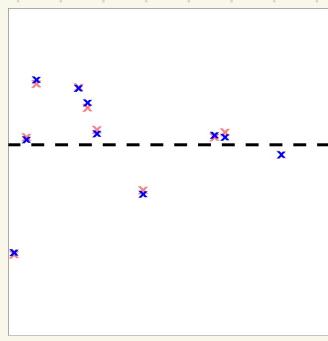
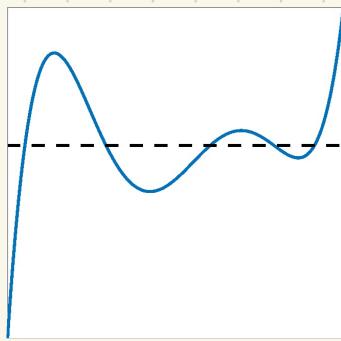
$$A^+ y$$



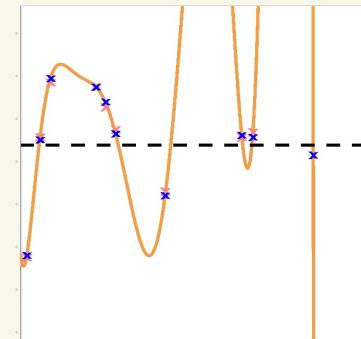
$$(A^T A + \gamma I)^{-1} A^T y$$

$$\gamma = \omega^2$$

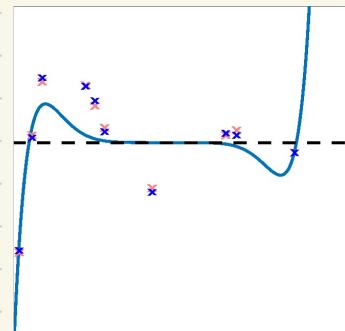
EXAMPLE



{



$$A^+ y$$



$$(A^T A + \gamma I)^{-1} A^T y$$

$$\gamma = 10^3$$

LINK WITH PSEUDO- INVERSE

- Pseudo-inverse :

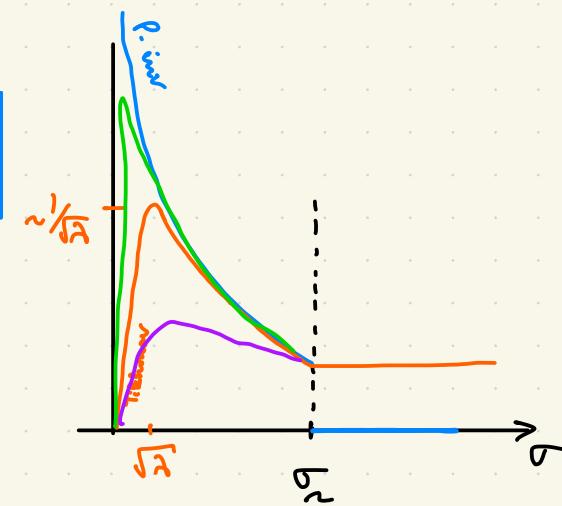
$$x_+ = A^+ y = \boxed{\sum_{i=1}^n \frac{u_i^\top y}{\sigma_i} v_i}$$

- Tikhonov :

$$x_\lambda = (A^\top A + \lambda I_n)^{-1} A^\top y = \boxed{\sum_{i=1}^n \frac{\sigma_i (u_i^\top y)}{\sigma_i^2 + \lambda} v_i}$$

- Possible strategy :

$$\tilde{x} = \boxed{\sum_{i=1}^n g(\tau_i) (u_i^\top y) v_i}$$



SENSITIVITY ANALYSIS

$$\lambda \rightarrow 0 \quad \delta \rightarrow 0 \\ P_\lambda(y + \delta) \rightarrow P_0(y)$$

$$(P_0(y)) \min_x \|Ax - y\|^2 \rightsquigarrow \min_x \|Ax - y\|^2 + \lambda J(x) \quad (P_\lambda(y))$$

Sensitivity = behavior of the solutions of $P_\lambda(y + \delta y)$

with respect to $P_\lambda(y)$ when $\delta y \rightarrow 0$ and $\lambda \rightarrow 0$

i.e. how close are the solutions ?

is there convergence ?

General answer is yes

Thm : assume $y \in \text{Ran } A$, with solution $x \in (\ker A)^\perp$ (just to simplify), and let $y^\delta \in (\mathbb{R}^m)^N$ be noisy measurements with $\|y^\delta - y\| \leq \delta$ ($\delta > 0$).

If $\frac{\delta}{\lambda} \rightarrow 0$, then

$$(A^T A + \lambda I)^{-1} A^T y^\delta \rightarrow x \text{ for } \delta \rightarrow 0$$

CONVERGENCE OF TIKHONOV

$$\min \|Ax - y\| + \lambda \|x\|^2$$

let $y \in \text{Ran } A$ be an ideal measurement vector.

let $y_p \notin \text{Ran } A$ be noisy measurements, with $\delta_p = \|y - y_p\| \xrightarrow{m \rightarrow \infty} 0$

1) let $\lambda > 0$. let $x_\lambda^{(p)}$ be the solution of $(T_\lambda(y_p))$, and x of $Ax = y$

Prop : Assume $x \in (\ker A)^\perp$ (to simplify), and let $w = A^T w$. Then

$$\forall m, \|x_\lambda^{(p)} - x\| \leq \|A^T\| \frac{\delta_p}{\lambda} + \sqrt{\frac{\lambda}{2}} \|w\|$$

proof : let x_λ be the solution of $(T_\lambda(y))$ noisier data

$$\|x_\lambda^{(p)} - x\| \leq \underbrace{\|x_\lambda^{(p)} - x_\lambda\|}_{\text{"err on the data"}} + \underbrace{\|x_\lambda - x\|}_{\text{"approximation"}}$$

- normal equations:

$$A^T A (x_\lambda^{(p)} - x_\lambda) + \lambda (x_\lambda^{(p)} - x_\lambda) = A^T (y_p - y)$$

hence $\|A(x_\lambda^{(p)} - x_\lambda)\|^2 + \lambda \|x_\lambda^{(p)} - x_\lambda\|^2 \leq \|A^T\| \|y_p - y\| \|x_\lambda^{(p)} - x_\lambda\|$ Cauchy-Schwarz

hence $\lambda \|x_\lambda^{(p)} - x_\lambda\|^2 \leq \|A^T\| \delta_p$

$$\langle A^T A (x_\lambda^{(p)} - x_\lambda), x_\lambda^{(p)} - x_\lambda \rangle + \lambda \|x_\lambda^{(p)} - x_\lambda\|^2 = \langle A^T (y_p - y), x_\lambda^{(p)} - x_\lambda \rangle$$

PROOF CONTINUED

$$\|x_\lambda^{(P)} - x\| \leq \underbrace{\|x_\lambda^{(P)} - x_\lambda\|}_{\lambda \|x_\lambda - x\|} + \underbrace{\|x_\lambda - x\|}_{\lambda \|x_\lambda - x\|}$$

- $\lambda \|x_\lambda - x\| \leq \|A^T\| \delta_m$

- We have $\|Ax_\lambda - Ax\|^2 = \langle A^T A x_\lambda - A^T A x, x_\lambda - x \rangle$

normal equations $\rightarrow = \langle -\lambda x_\lambda - \lambda x + \lambda x, x_\lambda - x \rangle$

$$= -\lambda \|x_\lambda - x\|^2 + \lambda \langle x, x - x_\lambda \rangle$$

$x = A^T w$ $\rightarrow = -\lambda \|x_\lambda - x\|^2 + \lambda \langle w, A(x - x_\lambda) \rangle$

Cauchy-Schwarz $\rightarrow \leq -\lambda \|x_\lambda - x\|^2 + \lambda \|w\| \|A(x_\lambda - x)\|$

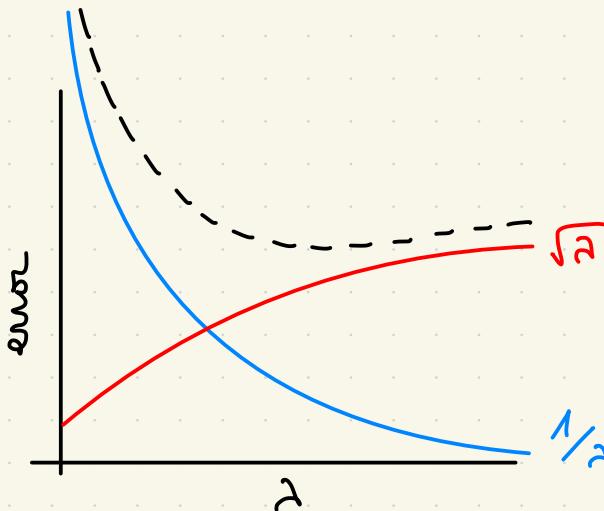
$ab \leq \frac{a^2 + b^2}{2}$ $\rightarrow \leq -\lambda \|x_\lambda - x\|^2 + \frac{\lambda^2 \|w\|^2}{2} + \frac{\|A(x_\lambda - x)\|^2}{2}$

hence $\frac{1}{2} \|A(x_\lambda - x)\|^2 + \lambda \|x_\lambda - x\|^2 \leq \frac{\lambda^2}{2} \|w\|^2$

and $\|x_\lambda - x\| \leq \sqrt{\frac{\lambda^2}{2}} \|w\|$

Finally $\|x_\lambda^{(P)} - x\| \leq \frac{1}{\pi} \|A^T\| \delta_p + \sqrt{\frac{\lambda^2}{2}} \|w\|$

BEHAVIOR OF THE ERROR



$\frac{1}{\lambda} \|A^T\| \delta_p$: error due to the noise (conditioning)

$\sqrt{\frac{\lambda}{2}} \|w\|$: error due to the approximation

regularization strategy needed : adapt λ to the noise level δ_p to control the first term.

⚠ In practice the noise level might not be known ...

CONVERGENCE

noiselss data

Theorem : let $y \in \text{Ran } A$, and $x = A^+y$ (solution of minimal norm). Assume $\delta_p = \|y_p - y\| \xrightarrow[p \rightarrow \infty]{} 0$ and $\lambda_p \xrightarrow[p \rightarrow \infty]{} 0$.

let x_p be the solution of $(T_{\lambda_p}(y_p))$. Then

$$1) \|Ax_p - y\| \xrightarrow[p \rightarrow \infty]{} 0$$

$$2) \text{ If } \frac{\delta_p}{\sqrt{\lambda_p}} \xrightarrow[p \rightarrow \infty]{} 0 \text{ then } \|Ax_p - y\| = O(\sqrt{\lambda_p}) \text{ and } x_p \xrightarrow[p \rightarrow \infty]{} x.$$

$$3) \text{ If } \frac{\delta_p}{\lambda_p} \xrightarrow[p \rightarrow \infty]{} 0 \text{ then } \|Ax_p - y\| = O(\delta_p) \text{ and } \|x_p - x\| = O(\delta_p)$$

PROOF

REMARKS

- Tradeoff stability vs. accuracy :
 γ_p must go to 0 **slower** than the noise to ensure convergence of the regularized solution.
- Error $\sim \gamma_p$, hence larger than noise.
- In practice δ_p is not known, and one has to estimate it from the available data

PARAMETER SELECTION

- δ is unknown
- Discrepancy principle : estimate noise level δ such that $\|A\mathbf{x}_\lambda^\delta - \mathbf{y}^\delta\| \approx \delta$
- Cross validation leave one datum out

$$\mathbf{x}_{\lambda,k} \text{ or } \min_{\mathbf{x}} \|A_{\sim k} \mathbf{x} - \mathbf{y}_{\sim k}\|^2 + \lambda \|\mathbf{x}\|^2 \quad A_{\sim k} = \left[\begin{array}{c c c} & & \\ & & \\ \hline & & \end{array} \right] \quad \mathbf{y}_{\sim k} = \left[\begin{array}{c} \\ \\ \hline \end{array} \right]$$

then minimize over λ : $CV(\lambda) = \sum_k |(A\mathbf{x}_{\lambda,k})_k - y_k|^2$ (prediction of remaining component)

$$CV(\lambda) = \sum_k \frac{|(A\mathbf{x}_\lambda)_k - y_k|^2}{|1 - P_{kk}(\lambda)|^2}, \quad P(\lambda) = A(A^T A + \lambda I)^{-1} A^T$$

TIKHONOV REGULARISATION 2

- Tikhonov also proposed (1960s)

$$\min_x \|Ax - y\|_2^2 + \lambda \|Bx\|_2^2$$

with $B = \begin{bmatrix} -1 & 1 & & & 0 \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ 0 & & \ddots & \ddots & 1 \\ & & & & -1 \end{bmatrix}$

- low $\|Bx\|$ encodes smoothness : if $x_i = f(t_i)$, then

$$Bx = \begin{pmatrix} x_2 - x_1 \\ x_3 - x_2 \\ \vdots \\ -x_n \end{pmatrix} = \begin{pmatrix} \Delta f(x_1) \\ \vdots \end{pmatrix}$$

- Solution :

$$x = (A^T A + \lambda B^T B)^{-1} A^T y$$

EXAMPLE

- Tikhonov regularization for image deconvolution / denoising

$$y = Ax + e \quad \rightsquigarrow \text{direct inversion amplifies noise: } \widehat{A}^{-1}y = \widehat{x} + \frac{\widehat{e}}{\alpha}$$

$\approx \alpha * x$

→ Tikhonov regularization: $\min \|y - Kx\|^2 + \lambda \|Bx\|^2$

assuming periodicity, $Bx = b * x$, and by Parseval:

$$\min \|\widehat{y} - \text{Diag}(\widehat{a})\widehat{x}\|^2 + \lambda \|\text{Diag}(\widehat{b}) \cdot \widehat{x}\|^2$$

Solution is $\widehat{x} = [\text{Diag}(|\widehat{a}|^2) + \lambda \text{Diag}(|\widehat{b}|^2)]^{-1} \text{Diag}(\overline{\widehat{a}}) \cdot \widehat{y}$

i.e. $\widehat{x}(k) = \frac{\overline{\widehat{a}(k)}}{|\widehat{a}(k)|^2 + \lambda |\widehat{b}(k)|^2} \widehat{y}(k)$

(compared with $\widehat{x}(k) = \frac{\widehat{y}(k)}{\widehat{a}(k)}$)

EXAMPLE



A

↓



A^{-1}

↓



A^{-1}

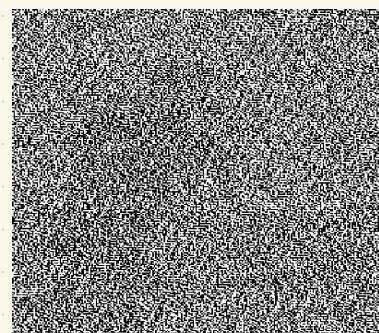
↓



A^{-1}

↓

+ noise



EXAMPLE



$\text{f} \neq T$



$T \neq R$



$\text{f} \neq T$
+ noise