REGULARIZATION

- The problem $\min_x \|Ax - y\|_2^2$
  - is ill-conditioned (sensitive to noise) if $\kappa(A) = \frac{\sigma_1}{\sigma_n}$ is large.
  - has ∞ solutions if $\ker A \neq \{0\}$.

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

PENALIZED PROBLEM

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

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

- **Regularized problem**

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

with : regularizer J usually convex

regularization parameter $\lambda > 0$

λ small $\longrightarrow$ solution will fit measurements well

λ large $\longrightarrow$ solution will be structured

⚠ tuning the parameter λ is a difficult task

CONSTRAINED FORMULATION

- **Equivalent formulation**

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

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

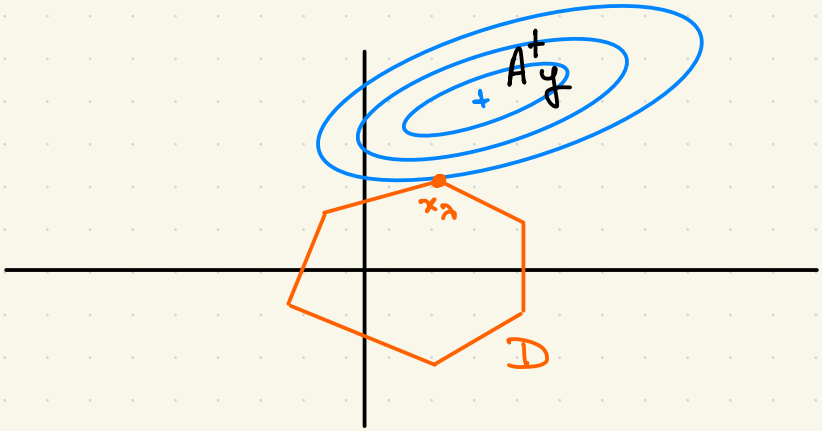
$\forall \tau, \exists \lambda(\tau)$: a solution of (C_τ) is also a solution of (P_λ).

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

- **Also equivalent to**

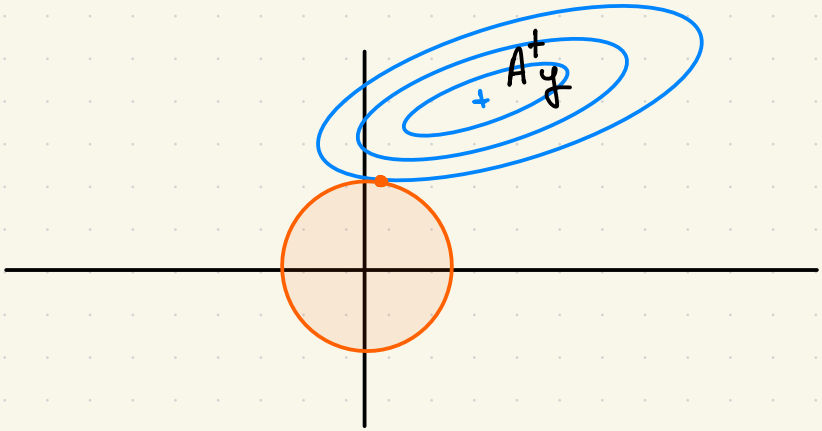
$$\text{minimize } J(x) \quad \text{s.t.} \quad \|Ax-y\| \leq \varepsilon$$

GEOMETRIC PICTURE



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

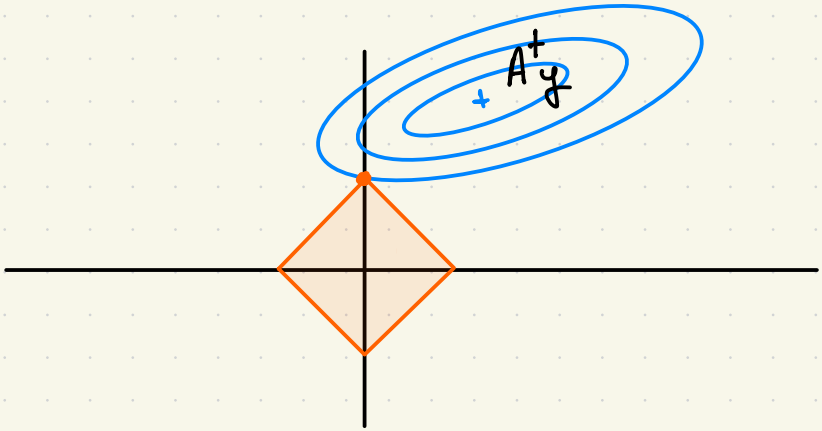
GEOMETRIC PICTURE



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

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

GEOMETRIC PICTURE



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

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

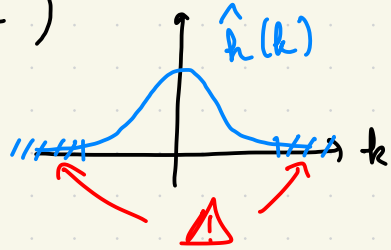
TIKHONOV REGULARIZATION

- **Example** : image deblurring / denoising

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

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

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



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

(aka ridge regression)

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

- 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 explicitely

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 $f : U \to Y$ is differentiable at $a \in U$ 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** : $f : \mathbb{R}^n \to \mathbb{R}$ (scalar-valued function of a vector variable). There exists a unique vector $\nabla f(a) \in \mathbb{R}^n$ such that $f'(a) \cdot h = \langle \nabla f(a), h \rangle$. In an o.b. $\nabla f(a) = (\partial_1 f(a), \dots, \partial_n f(a))$ where $\partial_i f(a)$ partial derivative.

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

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

DIFFERENTIAL CALCULUS 2

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

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

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

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

- **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

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

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

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

- Prop : Problem (T_λ) is equivalent to

$$(A^TA + \lambda I_m)x = A^Ty \quad (*)$$

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

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

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 $A^TA + \lambda I_n$ is invertible.

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

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

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

  $A^+ y$

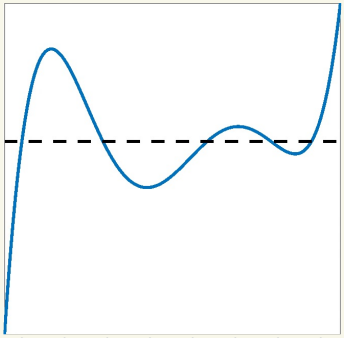
- **Rmk** : We always have that x_λ depends continuously on y

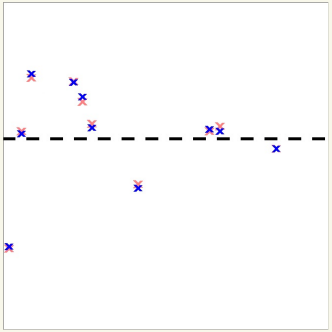
  since $\langle (A^*A + \lambda I_n)x, x \rangle = \|Ax\|^2 + \lambda \|x\|^2$

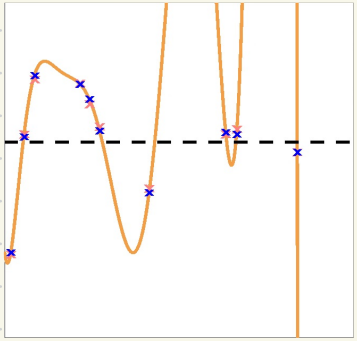
  $= \langle A^*y, x \rangle \leq \|A^*y\| \, \|x\|$ (Cauchy-Schwarz)

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

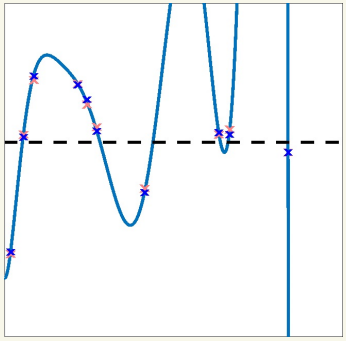
EXAMPLE







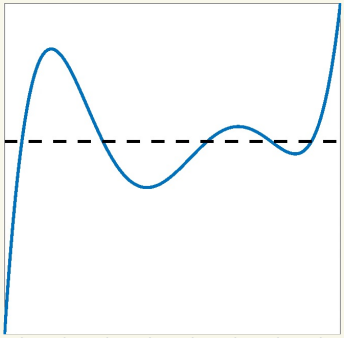
$A^{+}y$

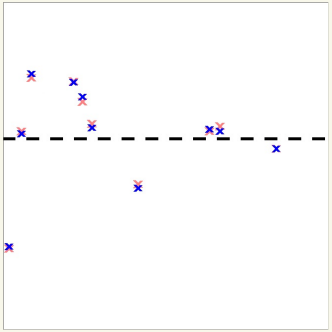


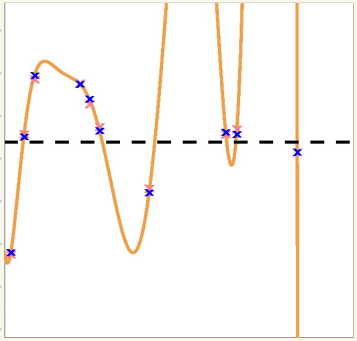
$(A^{T}A + \lambda I)^{-1}A^{T}y$

$\lambda = 10^{-5}$

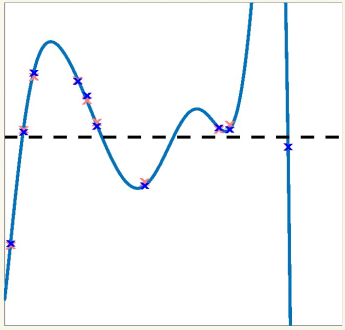
EXAMPLE







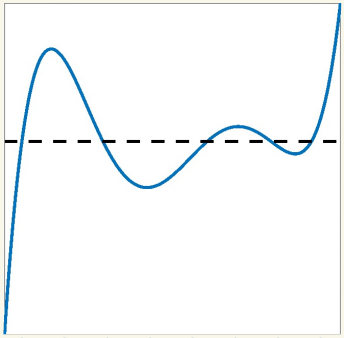
$A^{+}y$

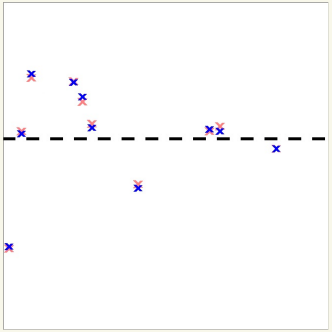


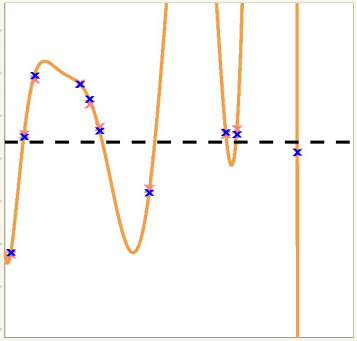
$(A^{T}A + \lambda I)^{-1}A^{T}y$

$\lambda = 10^{-4}$

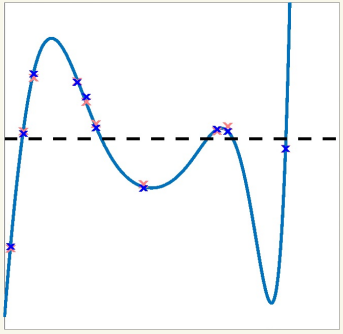
EXAMPLE







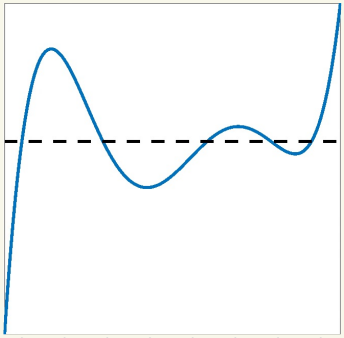
$A^{+}y$

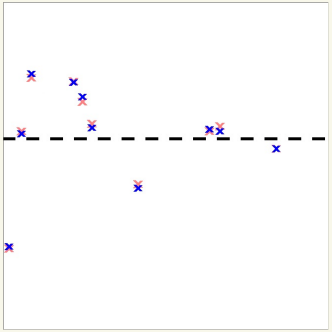


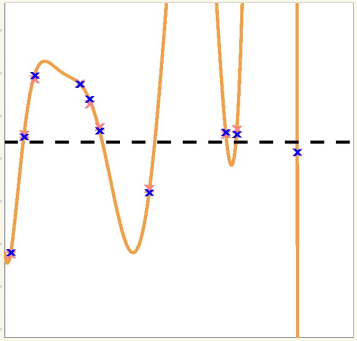
$(A^{T}A + \lambda I)^{-1}A^{T}y$

$\lambda = 10^{-3}$

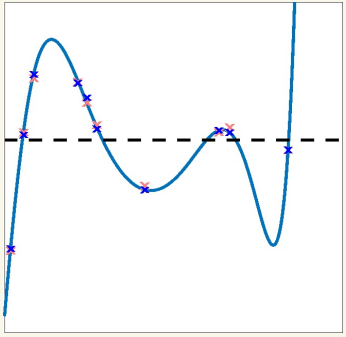
EXAMPLE







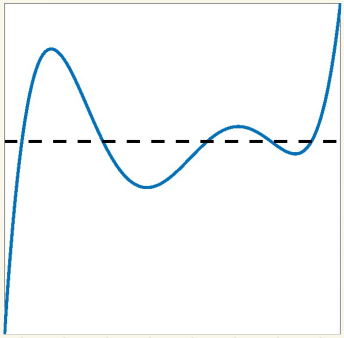
$A^{+}y$

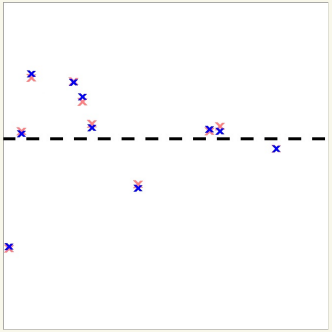


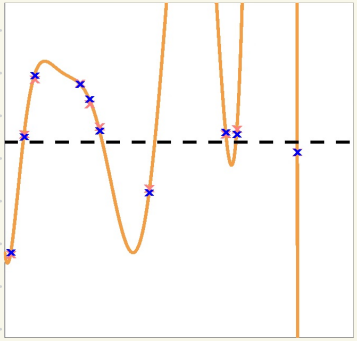
$(A^{T}A + \lambda I)^{-1}A^{T}y$

$\lambda = 10^{-2}$

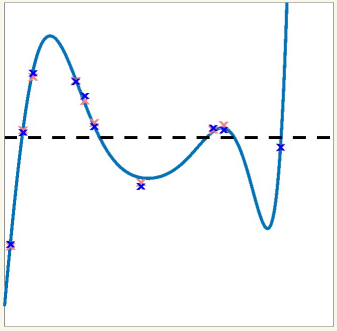
EXAMPLE







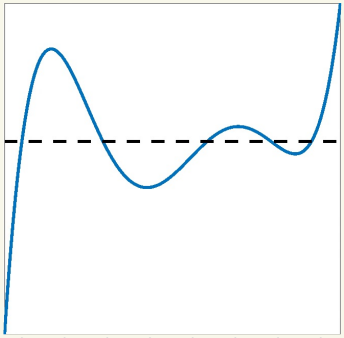
$A^{+}y$

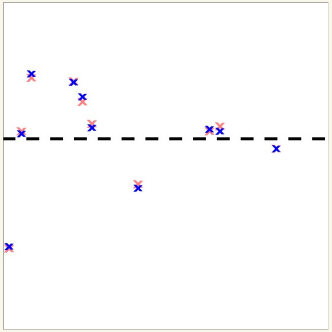


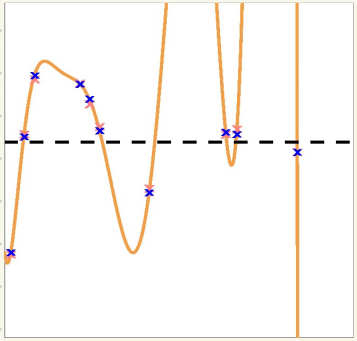
$(A^{T}A + \lambda I)^{-1}A^{T}y$

$\lambda = 10^{-1}$

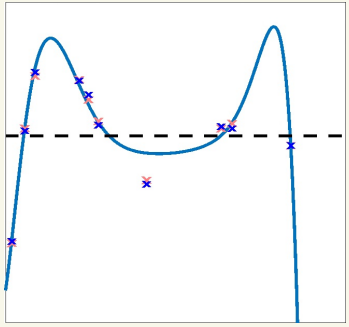
EXAMPLE







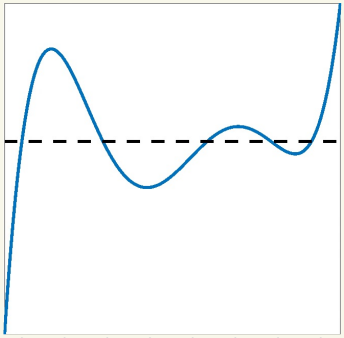
$A^{+}y$

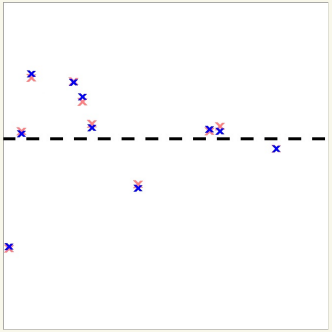


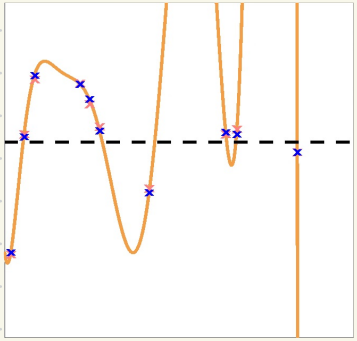
$(A^{T}A + \lambda I)^{-1}A^{T}y$

$\lambda = 1$

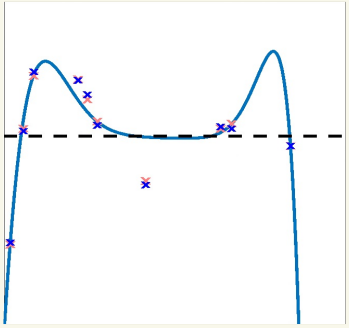
EXAMPLE







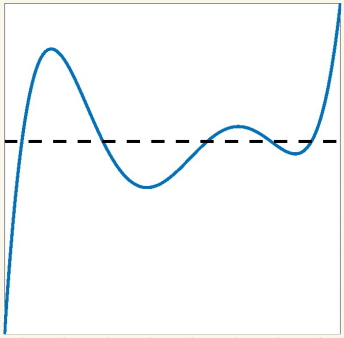
$A^{+}y$

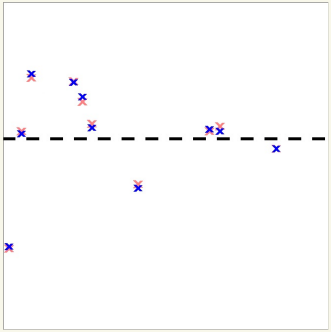


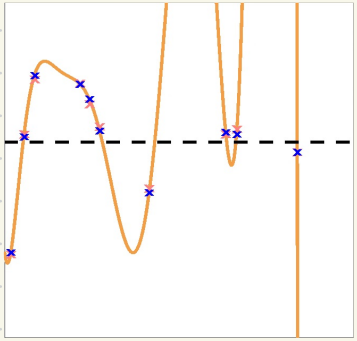
$(A^{T}A + \lambda I)^{-1}A^{T}y$

$\lambda = 10$

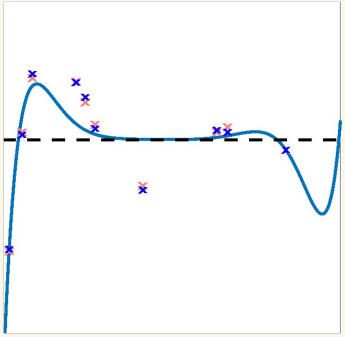
EXAMPLE







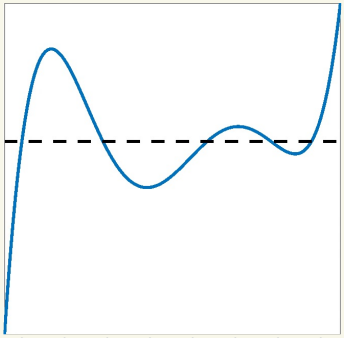
$A^{+}y$

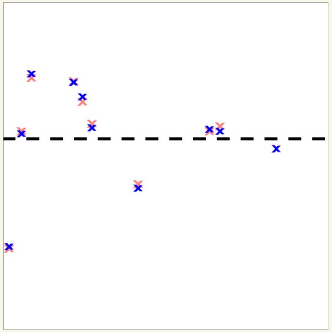


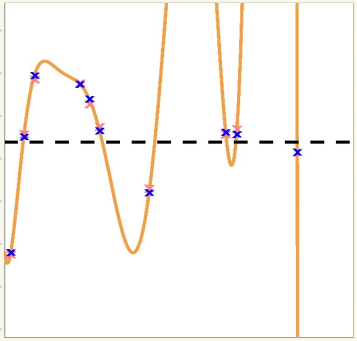
$(A^{T}A + \lambda I)^{-1}A^{T}y$

$\lambda = 10^{2}$

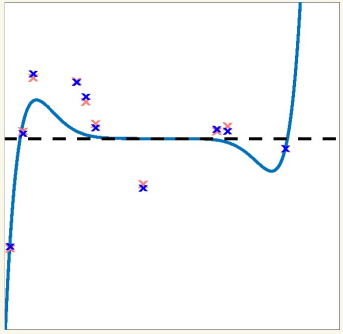
EXAMPLE







$A^{+}y$



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

$\lambda = 10^{3}$

LINK WITH PSEUDO-INVERSE

- Pseudo-inverse :

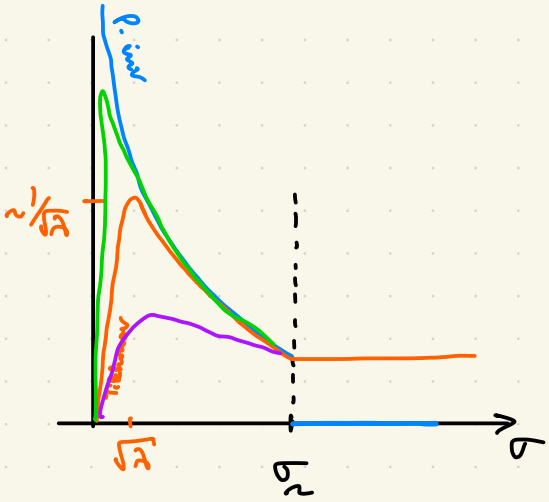
$$x_+ = A^+ y = \sum_{i=1}^{n} \frac{\mu_i^T y}{\sigma_i} v_i$$

- Tikhonov :

$$x_\lambda = (A^T A + \lambda I_n)^{-1} A^T y = \sum_{i=1}^{m} \frac{\sigma_i (\mu_i^T y)}{\sigma_i^2 + \lambda} v_i$$

- Possible strategy :

$$\tilde{x} = \sum_{i=1}^{m} g(\sigma_i)(\mu_i^T y) v_i$$



SENSITIVITY ANALYSIS

$\lambda \to 0$, $\delta \to 0$

$P_\lambda(y+\delta) \longrightarrow P_0(y)$

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

Sensitivity = behavior of the solutions of $P_\lambda(y+\delta y)$ with respect to $P_\lambda(y)$ when $\delta y \to 0$ and $\lambda \to 0$

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 $\frac{\delta}{\lambda} \to 0$, then

$(A^TA + \lambda I)^{-1} A^T y^\delta \longrightarrow x$ for $\delta \to 0$

CONVERGENCE OF TIKHONOV

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

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

let $y_\rho \notin \text{Ran}\, A$ be noisy measurements, with $\delta_\rho = \|y - y_\rho\| \xrightarrow[\rho \to \infty]{} 0$

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

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

$$\forall n, \quad \|x_\lambda^{(\rho)} - x\| \leq \|A^T\| \frac{\delta_\rho}{\lambda} + \sqrt{\frac{\lambda}{2}} \|w\|$$

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

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

- normal equations:

$$A^T A (x_\lambda^{(\rho)} - x_\lambda) + \lambda (x_\lambda^{(\rho)} - x_\lambda) = A^T (y_\rho - y)$$

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

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

$\langle A^T A (x_\lambda^\rho - x_\lambda), x_\lambda^\rho - x_\lambda \rangle + \lambda \|x_\lambda^\rho - x_\lambda\|^2 = \langle A^T (y_\rho - y), x_\lambda^\rho - x_\lambda \rangle$

PROOF CONTINUED

$$\|x_\lambda^{(p)} - x\| \leq \|x_\lambda^{(p)} - x_\lambda\| + \|x_\lambda - x\|$$

- $\lambda \|x_\lambda^{(p)} - x_\lambda\| \leq \|A^T\| \delta_m$
- We have $\|Ax_\lambda - Ax\|^2 = \langle A^TAx_\lambda - A^TAx, x_\lambda - x \rangle$

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

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

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

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

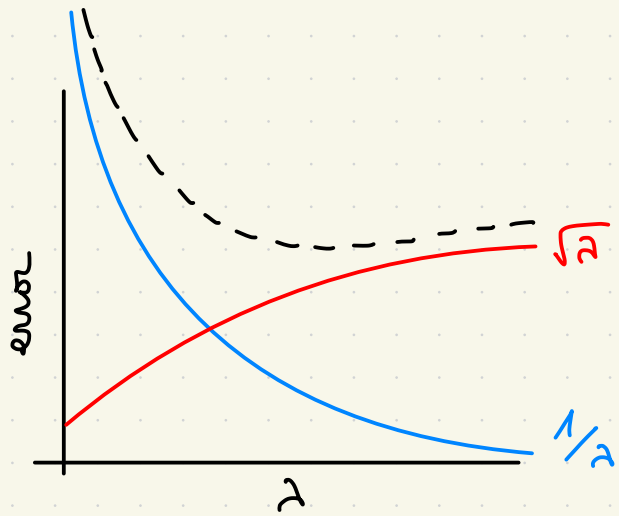
$ab \leq \frac{a^2+b^2}{2}$ $\longrightarrow$ $\leq -\lambda \|x_\lambda - x\|^2 + \frac{\lambda^2 \|w\|^2}{2} + \frac{\|A(x_\lambda - x)\|^2}{2}$

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

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

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

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

noiseless data

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

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

1) $\|A x_p - y\| \xrightarrow[p \to \infty]{} 0$

2) If $\delta_p / \sqrt{\lambda_p} \xrightarrow[p \to \infty]{} 0$ then $\|A x_p - y\| = O(\sqrt{\lambda_p})$ and $x_p \xrightarrow[p \to \infty]{} x$.

3) If $x \in (\ker A)^{\perp}$ and $\delta_p / \lambda_p \xrightarrow[p \to \infty]{} 0$ then $\|A x_p - y\| = O(\lambda_p)$ and $\|x_p - x\| = O(\sqrt{\lambda_p})$.

PROOF

. . .

Made with Goodnotes

REMARKS

- Tradeoff stability vs. accuracy :
  λ_ρ must go to 0 slower than the noise to ensure convergence of the regularized solution.

- Error $\sim \lambda_\rho$, hence larger than noise.

- In practice δ_ρ 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 $\| A x_\lambda^\delta - y^\delta \| \simeq \delta$
- **Cross validation** leave one datum out

$$x_{\lambda,k} \leftarrow \min_x \| A_{\sim k} x - y_{\sim k} \|^2 + \lambda \|x\|^2 \qquad A_{\sim k} = \left[\begin{array}{c}\\ \\ \\ \end{array}\right] \quad y_k = \left[\begin{array}{c}\\ \\ \end{array}\right]$$

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

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

TIKHONOV REGULARISATION 2

- Tikhonov also proposed (1960s)

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

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

- 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 :

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

EXAMPLE

- Tikhonov regularization for image deconvolution / denoising

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

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

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

$$\min \|\hat{y} - \mathrm{Diag}(\hat{a})\,\hat{x}\|^2 + \lambda \|\mathrm{Diag}(\hat{b}) \cdot \hat{x}\|^2$$

solution is $\hat{x} = \left[\mathrm{Diag}(|\hat{a}|^2) + \lambda\, \mathrm{Diag}(|\hat{b}|^2)\right]^{-1} \mathrm{Diag}(\overline{\hat{a}}) \cdot \hat{y}$

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

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

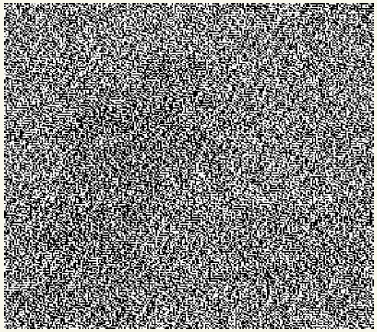
EXAMPLE

A

A^{-1}

+ noise

A^{-1}



EXAMPLE

A

$\tilde{A}$

A

+ noise

Tik.