PART 1

- Motiv° opt° convexe en ML $x_i \rightsquigarrow y_i$
  - → Regress° linéaire $\min_\beta \sum_i |y_i - \langle x_i, \beta \rangle| = f(\beta)$
  - → Class. $\mathbb{P}_\beta$ proba $\left(\frac{e^{\langle x,\beta\rangle}}{1+e^{\langle x,\beta\rangle}}, 1-\right)$ $\min_\beta \sum \ell(\langle x_i, \beta\rangle, y_i)$
    - logistic → hinge
  - → Algos de contraintes, pénalités non lisses.

Autres appli: ↗ imagerie
en data sciences.
⚠ $x \to \beta$ de la suite!!

- F° cvxe, Ensemble convexe, ff° étendue $f: \mathbb{R}^n \to \overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$
  - ↳ indicatrice

ex: $\|Ax-b\|^2$ $\|x\|_1$
$\{x \mid f(x) \le c\}$ ↳ cvxe.

$f \circ A$ ← linéaire
$f + g$ ← cvx
$\max(f,g)$ ← cvx

$\min_{x \in C} \|x-y\| \leftrightarrow \|\cdot - y\| + \delta_C$

$\max_y \langle x, y\rangle - g(y)$

- Différentials & Gradients. $n=1$ $\frac{f(u+\varepsilon)-f(u)}{\varepsilon} \to f'(u)$.
- $n \ge 1$ $f(u+\varepsilon h) = f(u) + \varepsilon \langle \nabla f(u), h\rangle + o(\varepsilon)$ $\frac{R(\varepsilon)}{\varepsilon} \to 0$. $\nabla f: \mathbb{R}^n \to \mathbb{R}^n$ (vecta fiel).

ex: $f(x) = \frac{1}{2}\|Ax-b\|^2$ $f(x+\varepsilon h) = \frac{1}{2}\|Ax-b+A\varepsilon\|^2 = \frac{1}{2}f(x) + \langle Ax-b, A\varepsilon\rangle + \varepsilon^2\|A\varepsilon\|^2$

$= \langle \varepsilon, A^T(Ax-b)\rangle$
$= \nabla f(x)$.

ex: multiclass $(\beta_1 - \beta_k)$.
loss logistic
$L(u,y) = \log \sum_i \exp(u_i) - u^T y$
prédiction: $(u_1 - u_k) = (\langle \beta, x_i\rangle)$
Observation $(y_1 - y_k)$ proba vecteur

$\nabla L(u,y) = \left(\frac{e^{u_i}}{\sum_i e^{u_i}} - y_i\right)_i$

- **Chain rule:** $\nabla(f \circ A) = A^T \times \nabla f \circ A$

$$f(A(x+\varepsilon)) = f(Ax) + \langle \nabla f(Ax), A\varepsilon \rangle + o(\varepsilon).$$

$$\nabla f \circ g(x) = \underbrace{[\partial g(x)]^T}_{\in \mathbb{R}^{n \times p}} \underbrace{(\nabla f(g(x)))}_{\in \mathbb{R}^p} \in \mathbb{R}^n$$

$f: \mathbb{R}^p \to \mathbb{R}^q$, $g: \mathbb{R}^n \to \mathbb{R}^p$. $\in \mathbb{R}^{p \times n}$

- **Differential:** $g: \mathbb{R}^n \to \mathbb{R}^p$

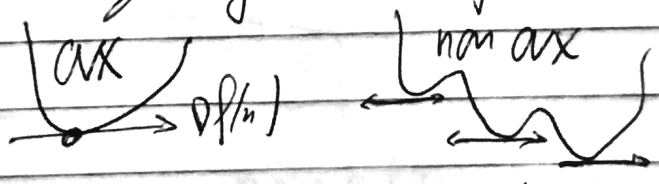
$$g(x+\varepsilon r) = g(x) + \varepsilon \underbrace{\partial g(x)[r]}_{\in \mathbb{R}^{n \times p} \text{ (linear)}} + o(\varepsilon).$$

⚠ $f: \mathbb{R}^n \to \mathbb{R}$ $\quad \partial f(x) \in \mathbb{R}^{1 \times n} \to \boxed{\nabla f(x) = \partial f(x)^T}$

$\nabla f(x) \in \mathbb{R}^{n \times 1}$

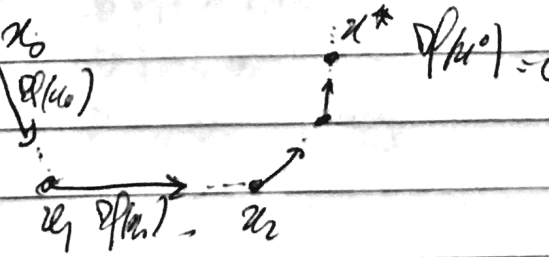
Fundamental thm of calc

$x^* \in \arg\min f \iff \nabla f(x^*) = 0$



- **Steepest descent:**

$x_{k+1} = x_k - \tau_k \nabla f(x_k)$



$$f(x - \tau \nabla f(x)) = f(x) - \tau \langle \nabla f(x), \nabla f(x) \rangle + o(\tau).$$

$\|\nabla f(x)\|^2 > 0$ if x not crit.

$\Rightarrow \exists \tau_k$ small enough such that $f(x_{k+1}) < f(x_k)$

- **Choice of τ** → fixed. → line search

Thm: $\exists [\tau_{min}, \tau_{max}]$, $\tau_k \in [\tau_{min}, \tau_{max}]$, $x_k \to x^*$

TP

Hessian: $f(x+\varepsilon h) = f(x) + \varepsilon f'(x) + \frac{\varepsilon^2}{2} f''(x) h^2 + o(\varepsilon^2)$ (n=1)

$\nabla f(x+\varepsilon h) = \nabla f(x) + \varepsilon \langle \partial^2 f(x) h, h \rangle$

$\partial^2 f(x)^\top = \partial^2 f(x) \in \mathbb{R}^{n \times n}$, symetric

$\partial^2 f(x) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right]$

$f(x+h) = f(x) + \varepsilon \langle \nabla f(x), h \rangle + \frac{\varepsilon^2}{2} \langle \partial^2 f(x) h, h \rangle + o(\varepsilon^2)$

↑ f 2× diff.

thm: f cvx $\iff$ $\partial^2 f(x) \succeq 0$.

Rappel: $A \succeq 0 \iff \forall h, \langle Ah, h \rangle \geq 0 \iff \text{Spectrum}(A) \in \mathbb{R}^+$

- Strong convexity vs Smoothness
- Conditionning
- Convergence speed

- Extension: newton, BFGS, SGD, Nesterov

TP: logistic classification

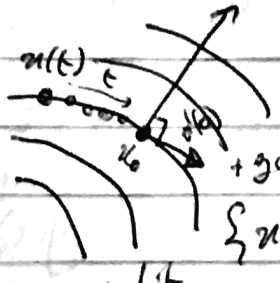
Next
- Non smooth optim (constraints, ℓ^1)
  - Projected gradient
  - Forward Backward
  - Lasso TP
  - Duality

- SGD - TP
- Automatic Differentiation - TP

→ Rappel sur le gradient + descente

→ Correct° exo semaine précédente ☺

→ Steppes? devant chir -



(show pic)

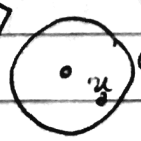
$x(t) \in L_r \iff \underbrace{f(x(t))}_{\triangleq g(t)} = r$ $\quad \{x : f(x) = r\} = L_r$

$g(t+\varepsilon) = f(x(t) + \varepsilon \underbrace{x'(t)}_{+o(\varepsilon)}) = f(x(t)) + \varepsilon \underbrace{\langle \nabla f(x(t)), x'(t) \rangle}_{=g'(t)} + o(\varepsilon)$

donc $g'(t) = \langle \nabla f(x(t)), x'(t) \rangle$

mais ĉ $g = cst \Rightarrow g'(t) = 0 \Rightarrow g'(0) = \langle \nabla f(x_0), \underbrace{x'(0)}_{\text{tangente}} \rangle = 0$

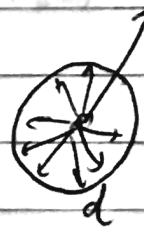
→ $\underset{|x_0 - x_\varepsilon|}{\operatorname{argmin}} f(x_0 + \varepsilon d) = f(x_0) + \varepsilon \langle \nabla f(x_0), d \rangle + o(\varepsilon)$.



$x_\varepsilon = x_0 + \varepsilon d$

$\|d\| \leq 1$? need to solve in the limit

$\underset{|d| \leq 1}{\operatorname{argmin}} \langle \nabla f(x_0), d \rangle \rightsquigarrow d = -\frac{\nabla f(x_0)}{\|\nabla f(x_0)\|}$



→ Rappels hessienne - SDP

→ C^2 smooth function : $\|\nabla^2 f(x)\| \leq L$

$\|M\| = \max_{|x| \leq 1} \frac{|Mx|}{|x|}$ ie $|Mx - My| \leq |x - y| \cdot \|M\|$

↳ lipsch. constant contract° ratio

if M symmetric $\rightarrow$ $= \max$ eigenvalues (M). $\quad Ae_i = \lambda_i e_i$ ↳ BON.

RMQ $\|AB\| \leq \|A\| \|B\|$?

ex : $f(x) = \frac{1}{2}\|Ax - y\|^2$ $\quad \nabla f(x) = A^*(Ax - y)$ $\quad \nabla^2 f(x) = A^*A$

$L = \|A^*A\| \leq \|A\|^2$

ex : $f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2)$ $\quad \nabla f(x) = \begin{pmatrix} x_1 \\ \gamma x_2 \end{pmatrix}$ $\quad \nabla^2 f(x) = \begin{pmatrix} 1 & 0 \\ 0 & \gamma \end{pmatrix}$.

if $\gamma > 1 \rightarrow |\nabla^2 f(x)| = \gamma$

- Thm de convergence : $x_{k+1} = x_k - \tau \nabla f(x_k)$

si $0 < \tau < \frac{2}{L}$ alors $x_k \to x^* \in \operatorname{argmin} f$
et $f(x_k) - f(x^*) \leq C/k$ (difficile).

- Strong convexity : $\partial^2 f(x) \geq \mu \, Id \Leftrightarrow \partial^2 f(x) - \mu Id \geq 0$
$\Rightarrow$ strict cvx ie $\langle \partial^2 f(x) h, h \rangle \geq \|h\|^2 \mu$.
$\Rightarrow$ unicité du min ie $\operatorname{eigen}(\partial^2 f(x)) \geqslant \mu$.

Ex : $f(x) = \frac{1}{2}|Ax - y|^2$ $\partial^2 f(x) = A^*A$
$\mu = \min \operatorname{eigen}(A^*A)$.
$\mu > 0 \Leftrightarrow A$ injective ie $\ker A = \{0\}$
ie $\operatorname{argmin} f$ singleton
$x^* = (A^*A)^{-1} A^* y$. "least square"

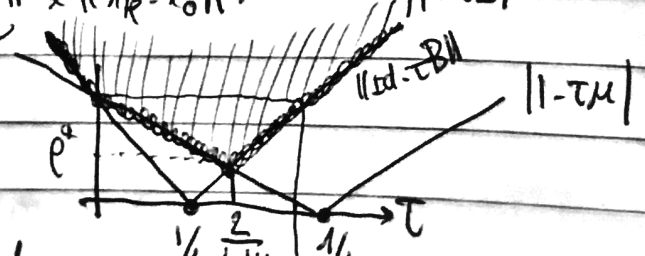
Thm de convergence linéaire : si $0 < \mu \leq L < +\infty$

$$|x_k - x^*| \leq C \cdot \rho^k \|x_0 - x_k\| \qquad 0 \leq \rho < 1$$

(x^* : unique ; $\|x_0 - x_k\|$: cor.)

Dans le cas quadratique :
si $\tau = \frac{2}{L+\mu}$ alors $\rho = \frac{L-\mu}{L+\mu} = 1 - \frac{2\varepsilon}{1+\varepsilon}$ $\varepsilon = \frac{\mu}{L} < 1$
(ε : Conditionnement (ratio valeur propre)).

en effet : $f(x) = \frac{1}{2}\langle Bx, x \rangle - \langle b, x \rangle \rightsquigarrow Bx^* = b$.
$\nabla f(x) = Bx - b$.
$x_{k+1} = x_k - \tau(Bx_k - b)$ $Bx^* - b = 0$.
$(x_{k+1} - x^*) = (x_k - x^*) - \tau B(x_k - x^*) = (Id - \tau B)(x_k - x^*)$.
$\Rightarrow \|x_k - x^*\| \leq \underbrace{\|Id - \tau B\|^k}_{\leq \rho} \times \|x_k - x_0\|$.



Newton : $x_{k+1} = x_k - \partial^2 f(x_k)^{-1} \nabla f(x_k)$ si f Quad : 1 ite ! sinon quad locale $\|x_{k+1} - x^*\| \leq C \|x_k - x^*\|^2$ $\rho < 2/L$ $\rho |x_k - x^*|^2 \leq \rho^{2^k}$

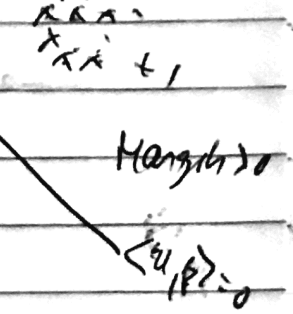
⚠ $\beta \leftrightarrow w$

TP classif logistic : input $(x_i, y_i)_{i=1}^n$

$x_i \in \mathbb{R}^p$ $\quad y_i \in \{-1, +1\}$.

Parametre : $\beta \in \mathbb{R}^p$

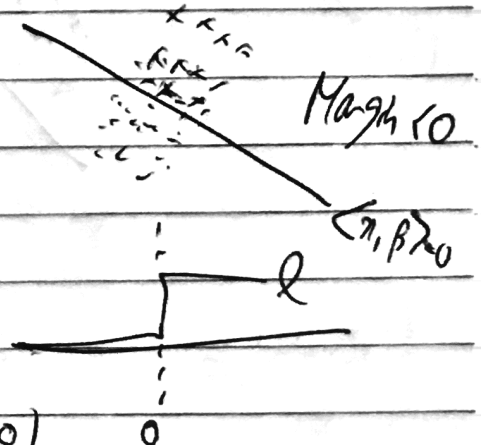
Frontière de décision $\langle x, \beta \rangle = 0$



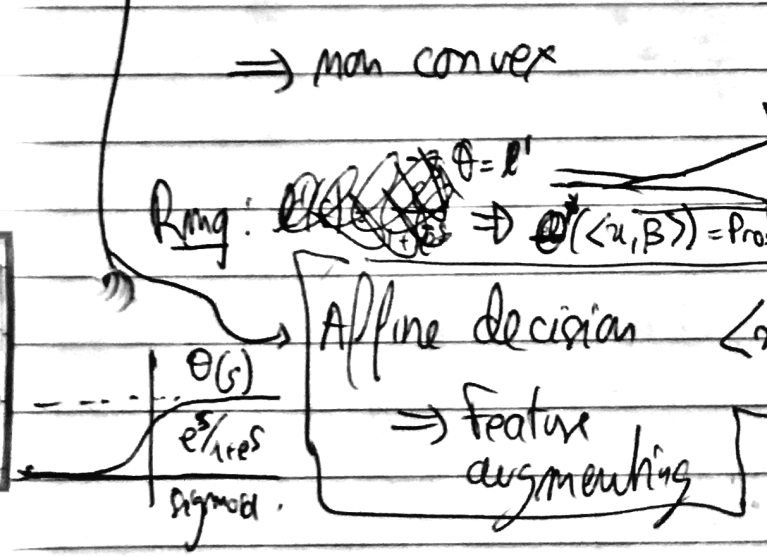
Décision ok if $\text{sign}(\langle x_i, \beta \rangle \cdot y_i) > 0$

0-1 loss : on paye 1 par mauvais choix

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \underbrace{\text{sign}(-\langle x_i, \beta \rangle \cdot y_i)}_{\ell(y_i \langle x_i, \beta \rangle)}$$



$\Rightarrow$ non convex



Rmq : $\Rightarrow \theta(\langle x, \beta \rangle) = \text{Proba}(y_i = +1) \Rightarrow$ smooth decision. cf GLM

$\ell(s) = \log(1 + \exp(s))$

Affine decision $\langle x, \beta \rangle = b$ (b : biais) $\Leftrightarrow \left\langle \begin{bmatrix} x \\ 1 \end{bmatrix}, \begin{bmatrix} \beta \\ b \end{bmatrix} \right\rangle = 0$

$\Rightarrow$ feature augmenting

$\tilde{x} \in \mathbb{R}^{p+1}$

$$\min_{\beta} f(\beta) = \frac{1}{n} \sum_i \ell(-y_i \langle x_i, \beta \rangle).$$

exo $\quad \nabla f(\beta) = \frac{1}{n} \sum_i -y_i \cdot \ell'(-y_i \langle x_i, \beta \rangle) x_i$

Preuve : $f(\beta + \varepsilon d) = \frac{1}{n} \sum \ell(-y_i \langle x_i, \beta \rangle + \varepsilon y_i \langle x_i, \delta \rangle)$

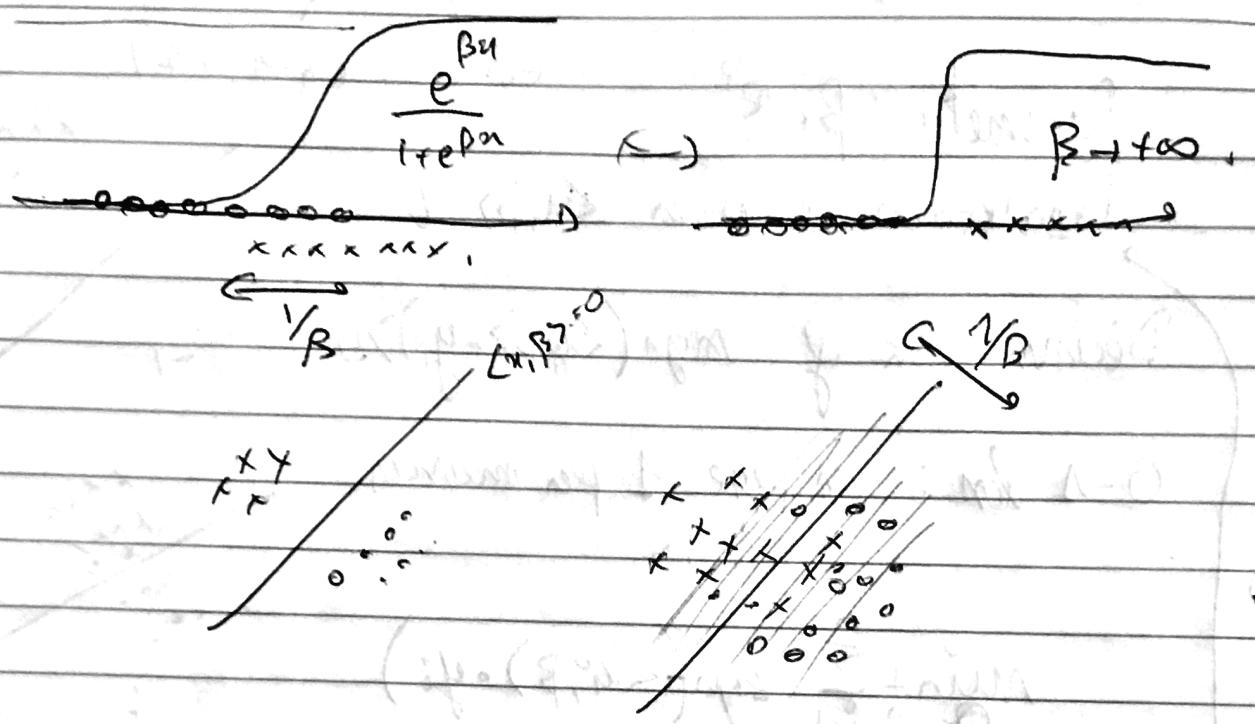
Logistic $\ell(s) = + \frac{e^{+s}}{1+e^{+s}}$

$= \frac{1}{n} \sum \ell(-y_i \langle x_i, \beta \rangle) + \varepsilon y_i \ell'(y_i \langle x_i, \beta \rangle) \langle x_i, \delta \rangle + o(\varepsilon)$

$= f(\beta) + \varepsilon \langle \delta, -\frac{1}{n} \sum_i x_i \times y_i \ell'(y_i \langle x_i, \beta \rangle) \rangle + o(\varepsilon)$.

6'

Logistic Log intuition



Autre notation. $X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^{n \times p}$.

$f(\beta) = \mathcal{L}(X\beta, y)$

Propᵉ : $\nabla f(\beta) = X^T \cdot \nabla \mathcal{L}(X\beta, y)$.

Preuve : $\mathcal{L}(X(\beta + \varepsilon\delta), y) = \mathcal{L}(X\beta + \varepsilon X\delta, y)$.

$= \mathcal{L}(X\beta, y) + \varepsilon \langle \nabla \mathcal{L}(X\beta, y), X\delta \rangle + r(\varepsilon)$

Id : $\mathcal{L}(u, y) = \frac{1}{n} \sum l(u_i, y_i)$.

$\nabla \mathcal{L}(u, y) = \frac{-1}{n} (y_1 l'(u_1, y_1), y_2 l'(u_2, y_2) \dots, y_n l'(u_n, y_n))$

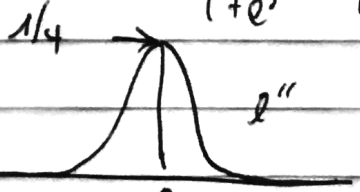
$= \frac{-1}{n} y \odot l'(u \odot y)$

↳ not: matlab/python

Logistic $l' = \Theta$

Hessian : $l(s) = \log(1 + \exp(s))$ $l'(s) = \frac{e^s}{1+e^s} = \frac{1}{1+e^{-s}}$

$l''(s) = \frac{e^{-s}}{(1+e^{-s})^2}$



$f(\beta) = L(y \odot (X\beta))$ $\nabla f(\beta) = X^T \cdot \mathrm{diag}(y) \cdot \nabla L(y \odot X\beta)$.

$\partial^2 f(\beta) = X^T \times \underbrace{\mathrm{diag}(y)}_{\|\cdot\| = 1} \times \partial^2 L(y \odot X\beta) \times \mathrm{diag}(y) \times X$.

$\|\partial^2 f(\beta)\| \leq \|X\|^2 \times \underbrace{\|\partial^2 L(y \odot X\beta)\|}_{\leq 1/4}$.

Multi-class Logistic : Histogram represent.

K classes, K vecteurs $(\beta_1, \dots, \beta_K) = B \in \mathbb{R}^{p \times K}$

Data : (x_i, d_i) $\quad d_i \in \Sigma_k$: histo $\quad d \in \Sigma_k$

loss between histo : $\ell(u, d) = \underbrace{\log\left[\sum_k \exp(u_k)\right]}_{\text{LSE}} - \sum_k u_k d_k$

$$\nabla \ell(u, d) = \left[\underbrace{\frac{e^{u_k}}{\sum e^{u_k}}}_{\text{soft max}} - d_k \right]_k$$

Rmq : $\text{LSE}((u_k) - c) = \text{LSE}((u_k)_k) - c$ $\quad c = \max(u_k)$

Regression linéaire : $y_i = \langle x_i, \beta \rangle$.

$$\min f(\beta) = \frac{1}{n} \sum_i |y_i - \langle x_i, \beta \rangle|^2 = \frac{1}{n} \|X\beta - y\|^2.$$

β sol° $\iff X^*(X\beta - y) = 0$.

si $\ker(X) = \{0\}$, sol° $\beta = (X^*X)^{-1} X^* y$.

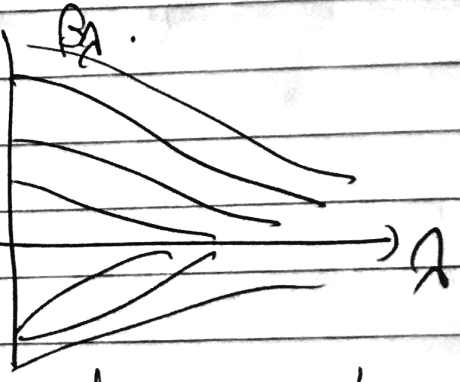
Regul° - Quad : $\min_\beta \frac{1}{n} \|X\beta - y\|^2 + \lambda \|\beta\|^2$ (Ridge)

$\|\beta\|^2$ strictement convexe

$$\beta_\lambda = (X^*X + \lambda \mathrm{Id})^{-1} X^* y.$$

→ LU / Cholesky

→ conjugate gradient



Cross val : $\min_\lambda \|X\beta_\lambda - y\|$.