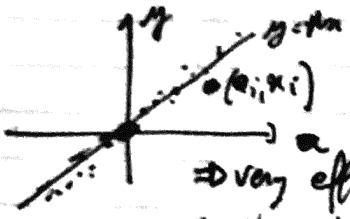
CENTRE
DE RECHERCHES
MATHÉMATIQUES

linear regression in data science.

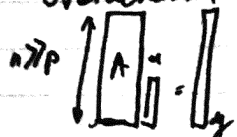
- Inverse pbm in imaging: forward model $\underset{\mathbb{R}^n}{y} = A\underset{\mathbb{R}^p}{x_0} + w$ → inpainting, deconvol°, Radon/tomo, IRM.
  usually $p \ll n$ (ill-posed).
- Regression in ML: supervised learning $(a_i, y_i)_{i=1}^n$ $a_i \equiv$ features, $a_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$

lin. predic° model $y = \langle a, x\rangle$ by fitting $\forall i$ $y_i \approx \langle a_i, x\rangle$

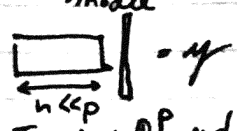


$$\underbrace{\begin{pmatrix} a_1^T \\ a_2 \\ \vdots \\ a_n \end{pmatrix}}_{A} x = y.$$

overdetermined



under determined (high-dim) model



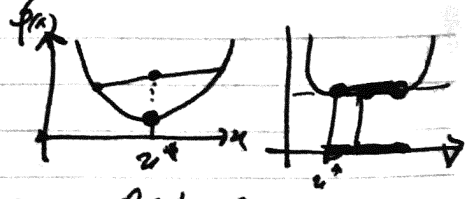
⇒ very efficient in high dim!

- Why we don't want to solve exactly $y = Ax$

PBM: $\left[\begin{array}{l} \mathrm{Ker}(A) \neq \{0\}. \\ Ax = y \text{ ∞ sol°} \end{array}\right] \longleftrightarrow \mathrm{Im}\,A \neq \mathbb{R}^p$, $y \notin \mathbb{R}^p$, no sol°

→ noise
→ over-det $y \notin \mathrm{Im}\,A$ → no sol° → least square fit, Gauss 1795

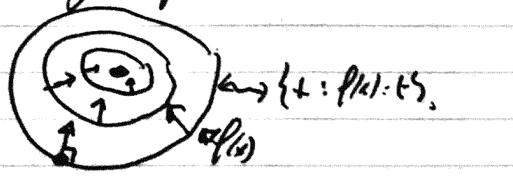
- LS fit $\boxed{\min_{x \in \mathbb{R}^p} f(x) = \frac{1}{2}\|Ax - y\|_2^2}$ f is convex



f is smooth (differentiable)

Thm: $x \in \mathrm{Argmin}\, f \iff \nabla f(x) = 0$

Refresher: $f: \mathbb{R}^p \to \mathbb{R}$, $\nabla f: \mathbb{R}^p \to \mathbb{R}^p$



$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x) \cdots \frac{\partial f}{\partial x_p}(x)\right)^T \in \mathbb{R}^p$$

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

Here: $f(x+\varepsilon) = \frac{1}{2}\|Ax - y + A\varepsilon\|^2 = \frac{1}{2}\|Ax-y\|^2 + \langle A\varepsilon, Ax - y\rangle + \underbrace{\frac{1}{2}\|A\varepsilon\|^2}_{o(\varepsilon)}$

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

$x^* \in \mathrm{Argmin}\, f \iff A^T(Ax^* - y) = 0 \iff (A^TA)x^* = A^Ty$ normal equat°

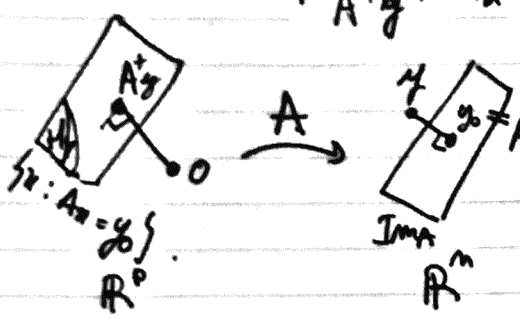
Prop: if $\mathrm{Ker}\,A = \{0\}$ (over-determined) $\boxed{A}\,x = \boxed{}\,y$ A^TA is positive and invertible

$x^* = (A^TA)^{-1}Ay$ is unique

otherwise $\mathrm{Argmin}\{f\}$ is infinite (hyperplane // to $\mathrm{Ker}(A)$)

- Need a select° method, also called Regularization. Use also least square → Ridge regression. → simplest method

$$A^+y = \operatorname{argmin}_x \left\{ \tfrac{1}{2}\|x\|^2 : x \in \underbrace{\operatorname{Argmin} \|Ax - y\|^2}_{= \{x : A^TAx = A^Ty\}} \right\}. \quad \text{(sol° unique)}$$



$= \{x : Ax = y_0\}$.

<u>Prop</u>: If $\mathrm{Im}A = \mathbb{R}^n$ (undedetermined, )

$$A^+y = A^T(AA^T)^{-1}y.$$

<u>Proof</u>: $\mathrm{Im}A = \mathbb{R}^p$ (surjective) $\to y_0 = y$. Lagrange mult: $\nabla \|x\|_2^2 \in (\text{constraint})^\perp$

$\Rightarrow x \in \mathrm{Ker}(A)^\perp = \mathrm{Im}(A^T)$.

$\Rightarrow x = A^Tu \Rightarrow Ax = \underbrace{AA^T}_{\text{invertible}}u = y$

$\Rightarrow u = (AA^T)^{-1}y$

<u>Summary</u>:

- Overdet $n \gg p$ $\leftrightarrow$ A surject $\leftrightarrow$ A^TA inv. $\Leftrightarrow A^+ = \underbrace{(\underbrace{A^TA}_{\in \mathbb{R}^{p\times p}})^{-1}}_{\in \mathbb{R}^{p\times p}} A^T$
- Underdet $\leftrightarrow$ A inj $\leftrightarrow$ AA^T inv $\Leftrightarrow A^+ = A^T(AA^T)^{-1}$
- A invertible $\leftrightarrow$ $A^+ = A^{-1}$ (never the case)
- General case $\to$ SVD (see bellow).

- Solving $A^TAx = A^Ty$ (over-det case): $\begin{cases} \text{conjugate gradient. if } p \text{ is large} \\ \text{Cholesky : if } p \text{ small} \end{cases}$
- SVD: todo.

• Regularis°: noise → ML (model error + small n) $\quad \min_{x\in\mathbb{R}^p} \frac{1}{2}|y-Ax|^2 + \overbrace{\frac{\lambda}{2}|x|^2}^{\lambda R(x)}$.

<u>Prop°</u>: $x_\lambda \overset{①}{=} (A^*A + \lambda Id_p)^{-1}A^*y$. → for small p (under determined)

$\overset{②}{=} A^*(AA^* + \lambda Id_n)^{-1}y$ → ok for small $n \ll p$ (under determined)

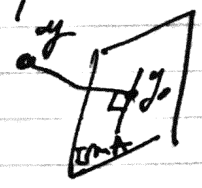
② allows to deal with ∞ dimension feature space (kernel method)

$u = (\underbrace{AA^*}_{\hat{K}} + \lambda Id_n)^{-1}y \qquad \hat{K} = (k(a_i, a_j))_{ij}$.

$\langle x, a\rangle$ becomes $\langle a, A^*u\rangle \qquad \sum_{i=1}^n u_i k(a_i, a)$.

<u>Thm</u>: $x_\lambda \xrightarrow{\lambda\to 0} A^+y$, more generally: $x_\lambda \to \operatorname{argmin}\{R(x) : Ax = y_0\}$. all sub-seq.

<u>Proof</u>: $\|y-Ax\|^2 = \underbrace{\|y_0 - Ax\|^2}_{\in \operatorname{Im} A} + \underbrace{\|y-y_0\|^2}_{\in (\operatorname{Im} A)^\perp}$



• Stat <u>vs</u> Imaging: Rather normalize $\min \frac{1}{n}\|y-Ax\|^2 + \lambda\|x\|^2$

$x_\lambda = (\underbrace{\tfrac{1}{n}A^*A}_{\hat{C}} + \lambda Id_p)^{-1}\underbrace{\tfrac{A^*y}{n}}_{\hat{u}} \qquad \hat{C} = \frac{1}{n}\sum_{i=1}^n a_i a_i^T \in \mathbb{R}^{n\times n}$.

under assump° on $(a_i, y_i) \overset{iid}{\sim} (a, y)$: $\begin{cases}\hat{C} \xrightarrow[n\to\infty]{\mathbb{P}} C = \mathbb{E}(aa^T) \in \mathbb{R}^{p\times p} \\ \hat{u} \qquad u = \mathbb{E}(ya) \in \mathbb{R}^p.\end{cases}$

heuristic: $\begin{array}{l}\text{ML}\\ \\ \end{array}\left[\begin{array}{l}\|\hat{C} - C\| \\ \|\hat{u} - u\|\end{array} \sim \underbrace{\frac{1}{\sqrt{n}}}_{\varepsilon}\right]$ stab $\quad \boxed{VS} \quad$ Imaging IP $\left[\begin{array}{l}\|C - \hat{C}\| = 0 \\ \|\underset{A^TAx_0}{u} - \underset{A^Ty}{\hat{u}}\| = \|A^*w\| \triangleq \varepsilon.\end{array}\right.$ $\quad y = Ax_0 + w$, $\varepsilon \leq \|w\|$

ML is harder to study than IP bc $\hat{C} \neq C$.

- Focus on the IP setup $y = Ax_0 + w$ $\longrightarrow$ see course note for ML

  Quest$^{\underline{n}}$: does $x_\lambda \to x_0$ as $\begin{pmatrix}\lambda \\ \|w\|\end{pmatrix} \to 0$? Rate, ! hyp on x_0 ?
- No if $x_0 \notin \mathrm{Im}(A^*) = \mathrm{Ker}\, A^\perp$!! Because $x_\lambda \in \mathrm{Im}\, A^*$!! $\Rightarrow$ Hyp: $x_0 \perp \mathrm{Ker}\, A$ is $x_0 \in \mathrm{Im}\, A^\perp$
- $x_\lambda - x_0 = (A^*A + \lambda \mathrm{Id})^{-1} A^*(Ax_0 + w) - x_0$.

$$= \underbrace{\left[(A^*A + \lambda \mathrm{Id})^{-1} A^*A - \mathrm{Id}_p\right]}_{\|\cdot\| \leq \lambda \|(A^*A)^{-1}\|} x_0 + (A^*A + \lambda \mathrm{Id})^{-1}(A^*w)$$

$\boxed{\text{one can show (ex)}}$ $\longrightarrow$ $\|\cdot\| \leq \lambda \|(A^*A)^{-1}\| \cdot \|x_0\|$ & $\|\cdot\| \leq \underbrace{\|(A^*A)^{-1}\|_{A^*}}_{\sqrt{\|(A^*A)^{-1}\|}} \cdot \|w\|$

<u>Pbm</u>: $\sigma_{\min}(A)$ can be arbitrary small (and can be 0 in the ∞ dimens$^{\underline{n}}$)

$\longrightarrow$ Need source condit$^{\underline{n}}$, of course.