**Supervised learning**: Obs: $(x_i, y_i)$, $x_i \in \mathbb{R}^p$

$y_i \in \mathbb{R}$: Regression (loss uses ordering)

$y_i \in \{1 \ldots K\}$: Classification (loss invariant permutation)

**Prediction function**: $y_i \text{ "}\approx\text{" } f(x_i, \beta)$ ← parameter

**Linear models**: $f(x,\beta) = \langle x, \beta \rangle$

bias: $\langle x,\beta\rangle + c = \langle [x,1], [\beta, c] \rangle$

non lin. via lifting: $f(x,\beta) = \langle \varphi(x), \beta \rangle$, $\beta \in \mathbb{R}^{p'}$, $\varphi(x) \in \mathbb{R}^{p'}$, $p' \gg p$

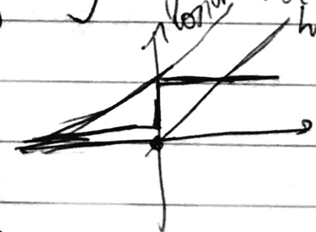
($\rightsquigarrow$ RKHS, $p' = +\infty$).

(P) **ERM**: $\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(\langle x_i, \beta \rangle, y_i) + \lambda R(\beta)$

$\left.\begin{matrix}\ell(\cdot, y) \\ R\end{matrix}\right\}$ cvx $\longrightarrow$ algo optim.

**Regression**: $\ell(y, y') = |y - y'|^2 / 2$

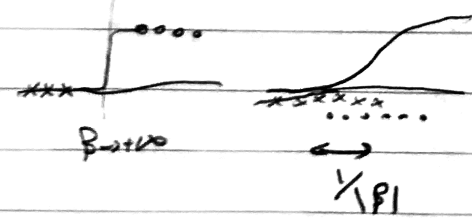
**Classif** (2 classes): $\ell(\bar{y}, y) = h(-\bar{y} y)$, $y \in \{-1, +1\}$

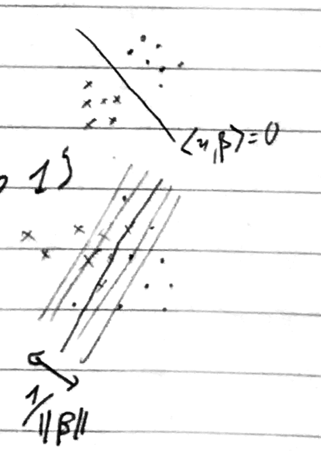



Hard Decision: $\mathrm{sign}(\langle x, \beta \rangle)$

Soft Decision: $\theta(\langle x, \beta \rangle) = \mathbb{P}(y \in \text{class } 1)$

$\theta(y) = \dfrac{e^y}{1+e^y}$



# RIDGE REGRESSION

$$\hat{\beta}_\lambda \triangleq \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\sum_i (y_i - \langle \beta, x_i\rangle)^2 + \lambda\|\beta\|^2 = \frac{1}{n}\|y - X\beta\|^2 + \lambda\|\beta\|^2$$

$$= \Big(\underbrace{\frac{X^*X}{n}}_{\hat{C}} + \lambda \mathrm{Id}_p\Big)^{-1}\underbrace{(X^*y)}_{\hat{\mu}} \qquad X = \begin{bmatrix} x_i \\ \vdots \end{bmatrix} \in \mathbb{R}^{n\times p}$$

Hyp: $x_i \overset{iid}{\sim} X$, $y_i \overset{iid}{\sim} Y$, $E(\|x\|^4) < +\infty$, $E(y^4) < +\infty$ $\Big\} \overset{\sim TCL}{\Longrightarrow}$ $\hat{C} \xrightarrow[\mathbb{P}, E^2]{n} C \triangleq \mathbb{E}(xx^*) \in \mathbb{R}^{p\times p}$, $\hat{\mu} \longrightarrow \mu \triangleq \mathbb{E}(xy) \in \mathbb{R}^p$

Refresher : cv of random variables

**Consistency** : $\hat{\beta}_\lambda \xrightarrow[\mathbb{P}]{n\to+\infty} \beta_\lambda \triangleq (C + \lambda \mathrm{Id}_p)^{-1}\mu$ } random

$\downarrow \lambda \to 0$ } deterministic

$\beta_0 = C^+\mu$

$\lambda = \lambda_n$ ?? consistency ??

Refresher, pseudo-inverse : $C^+\mu = \underset{\beta}{\operatorname{argmin}} \|\beta\|$ st $C\beta = \mu$

- if $\mathrm{Ker}\, C = \{0\}$, $C^+ = C^{-1}$. 
- SVD / eigen: $C = U\mathrm{diag}(\sigma_i)U^* \overset{\sigma_i > 0}{\leadsto} C^+ = U\mathrm{diag}\left(\frac{1}{\sigma_i}\right)U^*$

Example : well specified model : $y_i = \langle \beta^*, x_i\rangle + \varepsilon_i$ $\quad \mathbb{E}(\varepsilon_i) = 0$
iid, indep

$$\mu = \mathbb{E}_{xy}(yx) = \mathbb{E}_x(\langle \beta^*, x\rangle x) + \underbrace{\mathbb{E}(y\varepsilon)}_{=E(y)E(x)=0}$$

$$= \mathbb{E}(xx^T)\beta^* = C\beta^*$$

$$\leadsto \beta_0 = C^+C\beta^* = \mathrm{Proj}_{\mathrm{Im}(C)}(\beta^*)$$

$\leadsto$ Pbm if $\beta^* \notin \mathrm{Im}(C)$ the part of $\beta^*$ in $\mathrm{Ker}(C)$ is lost.
$\mathrm{Im}(C) = \mathrm{Ker}(C)^\perp$

$\leadsto$ to recover the part of $\beta^*$ inside $\mathrm{Ker}(C)$ needs non-linear methods (eg. $\ell^1$, aka LASSO) defined on same proba. space !!

<u>Convergence in proba</u> : $\beta_n \xrightarrow{\mathbb{P}} \beta_0 \iff \forall \varepsilon \ \mathbb{P}(\|\beta_n-\beta_0\| \geq \varepsilon) \xrightarrow{n\to+\infty} 0$

<u>Quantified</u> : $\|\beta_n-\beta_0\| = O_{\mathbb{P}}\left(\frac{1}{n^k}\right)$ : "with proba $1-\eta$, $\forall \eta>0$

$\|\beta_n-\beta_0\| \leq C(\eta) \times \frac{1}{n^k}$ "

very often $\sim \log(1/\eta)$

<u>$L^p$-Convergence in expect°</u> : $\mathbb{E}(\|\beta_n-\beta_0\|^p) \xrightarrow{n\to+\infty} 0$

<u>Rmq</u> : Markov ineq : $\mathbb{P}(\|\beta_n-\beta_0\| \geq \varepsilon) \leq \frac{1}{\varepsilon^p}\mathbb{E}(\|\beta_n-\beta_0\|^p)$

$L^p$ convergence stronger ($\Rightarrow$) convergence in proba

<u>Other convergence</u> : Almost sure, in law (aka optimal transport).

↳ need not be def. on same space!

$$\begin{tikzcd}
\boxed{L^p \text{ conv.}} \arrow[dr, Rightarrow] & & \boxed{\text{A.S. conv}} \arrow[dl, Rightarrow] \\
& \boxed{\text{in proba}} \arrow[d, Rightarrow] & \\
& \boxed{\text{in law}} &
\end{tikzcd}$$

same space

diff space

# ESTIM° & PREDIC° RATES



Def°: $\hat{\beta}_\lambda \triangleq (\hat{C} + \lambda \mathrm{Id})^{-1} \hat{u}$ ; $\tilde{\beta}_\lambda \triangleq (C + \lambda \mathrm{Id})^{-1} \hat{u}$ ; $\beta_\lambda \triangleq (C + \lambda \mathrm{Id})^{-1} u$

VARIANCE

$$\hat{\beta}_\lambda - \beta_0 = \tilde{\beta}_\lambda - \beta_\lambda \xleftarrow{2} \underbrace{(C+\lambda \mathrm{Id})^{-1}}_{\|\cdot\| \leq \frac{1}{\lambda}} \underbrace{(\hat{u} - u)}_{\|\cdot\| \leq \delta} \qquad \boxed{\delta \triangleq \frac{1}{\sqrt{n}}}$$

in proba (1)

$$+ \hat{\beta}_\lambda - \tilde{\beta}_\lambda \xleftarrow{1} \left[ (\hat{C} + \lambda \mathrm{Id})^{-1} - (C + \lambda \mathrm{Id})^{-1} \right] \times \hat{u}$$

in proba $\sim \frac{\delta}{\lambda}$ (why not $\frac{\delta}{\lambda^2}$ ??) (2)

BIAS

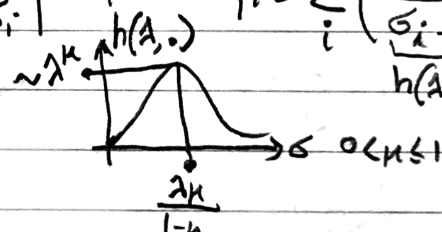
$$+ \beta_\lambda - \beta_0 \xleftarrow{3} \Big[ \underbrace{(C + \lambda \mathrm{Id})^{-1} - C^{+}}_{} \Big] u$$

need hyp. on $\beta_0$

**Lemma** (approx°): if $\beta_0 \in \mathrm{Im}(C^{\mu})$ ($0 \leq \mu \leq 1$, def of $C$) ie $\exists q, \beta_0 = C^{\mu} q$, then $\|\beta_\lambda - \beta_0\| \leq c(\mu) \cdot \|q\| \lambda^{\mu}$ (3')

**Proof**: $\|\beta_\lambda - \beta_0\|^2 = \sum_i \left| \frac{1}{\sigma_i + \lambda} - \frac{1}{\sigma_i} \right|^2 \times \sigma_i^{2(\mu+1)} \times q_i^2 = \sum_i \Big( \underbrace{\frac{\lambda \sigma_i^{\mu}}{\sigma_i + \lambda}}_{h(\lambda, \sigma_i)} \Big)^2 \times u_i^2$

$\boxed{u = C\beta_0 = C^{\mu+1} q}$




**Thm [EST°]**: (with high proba) $\|\hat{\beta}_{\lambda_n} - \beta_0\| = O_P\left( \left(\frac{1}{\sqrt{n}}\right)^{\frac{\mu}{1+\mu}} \right)$ $\lambda_n \triangleq n^{-\frac{1/2}{1+\mu}}$

**Proof**: (1)+(2)+(3') $\Rightarrow \|\hat{\beta}_\lambda - \beta_0\| = O\left(\frac{\delta}{\lambda} + \lambda^{\mu}\right) \rightsquigarrow$ choose $\lambda = \delta^{\frac{1}{1+\mu}}$

balance

**Rates**: $\beta = 1 \to 1/2$ (optimal); $\beta = 1/2 \to 1/3$ (usual one).

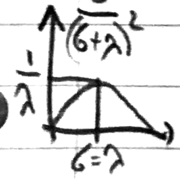
**Rem**: to have faster (eg linear) rates, needs NL meth (ε1)

---

**Prediction**: $\mathbb{E}_x(\langle x, z \rangle^2) = \langle Cz, z \rangle$

(3) $\rightsquigarrow \langle C(\beta_\lambda - \beta_0), \beta_\lambda - \beta_0 \rangle = \sum_i \left( \frac{\lambda \sigma_i^{\mu+1/2}}{\sigma_i + \lambda} \right)^2 u_i^2 \overset{\mu \leq 1/2}{=} O(\|u\|^2 \cdot \lambda^{2\mu+1})$ (A)

(2) $\rightsquigarrow \langle C(\tilde{\beta}_\lambda - \beta_\lambda), \tilde{\beta}_\lambda - \beta_\lambda \rangle = \sum_i \sigma_i \frac{1}{(\sigma_i + \lambda)^2} \times (\hat{u}_i - u_i)^2 \overset{\triangle}{=} O\left( \frac{\|\hat{u} - u\|^2}{\lambda} \right) = O\left( \frac{1}{n\lambda} \right)$ (B)




(1) $\rightsquigarrow \langle C(\hat{\beta}_\lambda - \tilde{\beta}_\lambda), (\hat{\beta}_\lambda - \tilde{\beta}_\lambda) \rangle = ?? = O(1/n\lambda)$ (C)

**Thm** $\mathbb{E}(\langle x, \hat{\beta}_\lambda - \hat{\beta}_0 \rangle^2)^{1/2} = O\left( (1/\sqrt{n})^{\frac{\mu+1/2}{\mu+1}} \right)$ $\lambda_n \triangleq n^{-\frac{1/2}{\mu+1}}$

**Rate**: $\mu = 1 \to 3/4$; $\mu = 1/2 \to 2/3$

**Proof**: (A)+(B)+(C) $\to \langle C(\hat{\beta}_\lambda - \beta_0), \hat{\beta}_\lambda - \beta_0 \rangle^{1/2} = O\left( \delta/\sqrt{\lambda} + \lambda^{\mu + 1/2} \right) \to \lambda = \delta^{\frac{1}{\mu+1}}$

Discussion on the source cond$^{\underline{c}}$ : $\beta_0 = C^{\mu} q = U \operatorname{diag}(\sigma_i^{\mu}) U^T q$.

- In finite dimension, $\beta_0 \in \operatorname{Im}(C) \Rightarrow \forall \mu,\ \beta_0 \in \operatorname{Im}(C^{\mu})$ !

  indeed, $\beta_0 = C^{\mu} q$ for $q = (C^{+})^{\mu} \beta_0$ !

  But can be very bad bound $\|q\| \sim \frac{1}{\lambda_{min}^{\mu}}$ ...

  $\lambda_{min}$ can goes to 0 very fast with dim$^{\underline{c}}$ p.

  so the goal is to have $\|q\|$ small

- In $\infty$ dim$^{\underline{c}}$ (eg RKHS) $\beta_0 \in \operatorname{Im}(C^{\mu})$ not always true, typicaly $\lambda_{min} = 0$, $\sigma_i \xrightarrow[i \to \infty]{} 0$ !

- Rule of thumb : if C trans$^{\underline{c}}$ invariant (convol$^{\underline{c}}$), U = fft, and $\sigma_i$ small for high freq, so that lang $\mu$ corresponds to smoother $\beta_0$ ($\sim$ SOBOLEV space).

# COMPARISON with IP

$Y = X\beta_0 + \varepsilon \qquad \delta = \|\varepsilon\|$

<u>ESI°</u>: $\|\hat{\beta}_\lambda - \beta_0\| = O\left(\frac{\delta}{\sqrt{\lambda}} + \lambda^{\mu}\right) = O\left(\delta^{\frac{\mu}{\mu+1/2}}\right)$.

<u>PRED°</u>: $\frac{1}{\sqrt{n}}\|X(\hat{\beta}_\lambda - \beta_0)\| = O\left(\delta + \lambda^{\mu}\right) = o(\delta)$ (trivial bound).

indep $\lambda$ !!