Surprises in High-Dimensional Ridgeless Least Squares Interpolation.

Hastie, Montanari, Rosset, Tibshirani.

(1) Classical approximation / variance trade-off.

Classical learning problem: $(X_i, Y_i) \overset{iid}{\sim} \mathbb{P}$,

learn $f^*(x) = \mathbb{E}[Y \mid X = x]$:

$$\hat{f}_N \in \operatorname*{argmin}_{f \in \mathcal{F}_N} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i))$$

risk: approximation / variance trade-off

$$\mathbb{E}[d(\hat{f}, f^*)] = \underbrace{\min_{f \in \mathcal{F}_N} d(f, f^*)}_{\text{approximation error}} + \underbrace{\mathbb{E}[d(\hat{f}, f^*)] - \min_{f \in \mathcal{F}_N} d(f, f^*)}_{\text{stochastic error.}}$$

when $d(\hat{f}, f^*) = \|\hat{f} - f^*\|_{\mathcal{H}}^2$ ← Hilbert norm.

we have the bias/variance decomposition

$$\mathbb{E}\left[\|\hat{f} - f^*\|_{\mathcal{H}}^2\right] = \underbrace{\mathbb{E}\left[\|\hat{f} - \mathbb{E}[\hat{f}]\|_{\mathcal{H}}^2\right]}_{\text{variance term}} + \underbrace{\|\mathbb{E}[\hat{f}] - f^*\|_{\mathcal{H}}^2}_{\text{bias term.}}$$

typical examples:

1/ $(\varphi_j)_{j \geq 1}$ Fourier basis, $\ell(y, f(x)) = (y - f(x))^2$

$$\mathcal{F}_N = \left\{ f = \sum_{j=1}^{N} \theta_j \varphi_j, \ \theta_j \in \mathbb{R} \right\}, \quad d(\hat{f}, f^*) = \|\hat{f} - f^*\|_{L^2}^2$$

$Y_i = f^*(x_i) + \varepsilon_i$, $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

- Approximation error: $f^* = \sum_{j \geq 1} \theta_j \varphi_j$

$$\min_{f \in \mathcal{F}_N} \|\hat{f} - f^*\|_{L^2}^2 = \sum_{j \geq N+1} \theta_j^2$$

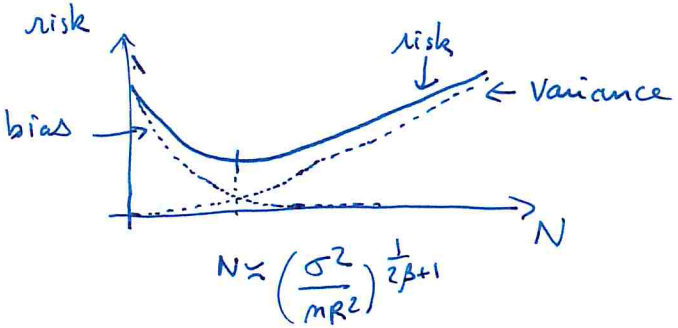
- Stochastic error: $\frac{N}{n} \sigma^2$

$$\Big\} \Rightarrow \mathbb{E}[d(\hat{f}, f^*)] \leq \sum_{j \geq N+1} \theta_j^2 + \frac{N}{n} \sigma^2.$$

case when $f^* \in W_\beta^{\mathrm{Pen}}(R) = \left\{ f : \sum_{j \geq 1} j^{2\beta} \theta_j^2 \leq R^2 \right\}$.

$$\rightsquigarrow \sum_{j \geq N+1} \theta_j^2 \leq \frac{1}{N^{2\beta}} R^2$$

$$\Rightarrow \text{risk} \leq \frac{R^2}{N^{2\beta}} + \frac{N}{n} \sigma^2 \quad \Rightarrow \text{"optimal" } N \quad N \asymp \left(\frac{\sigma^2}{nR^2}\right)^{\frac{1}{2\beta+1}}$$



$$N \asymp \left(\frac{\sigma^2}{nR^2}\right)^{\frac{1}{2\beta+1}}$$

2/ $\mathcal{F}_N$ = ball $B_{\mathcal{H}}(0,N)$ in a RKHS.

remark: in a RKHS $\|f\|_{\mathcal{H}}$ $\leftrightarrow$ regularity as

$$|f(x)-f(y)| = |\langle f, k(x,\cdot)-k(y,\cdot)\rangle_{\mathcal{H}}|$$
$$\leq \|f\|_{\mathcal{H}} \, \|k(x,\cdot)-k(y,\cdot)\|_{\mathcal{H}}.$$

Lagrangian version

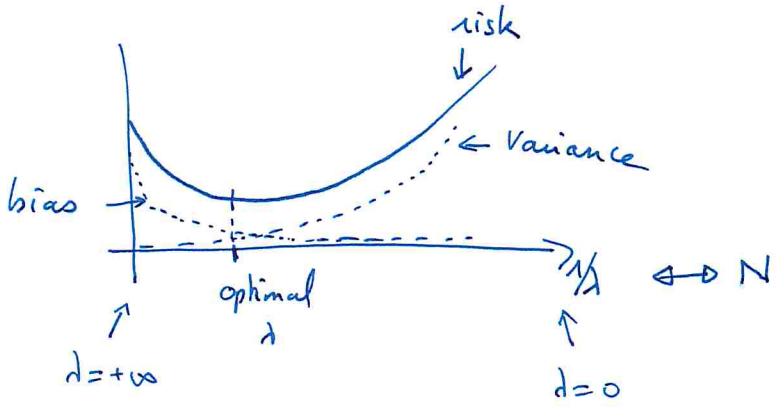
$$\hat{f} \in \operatorname*{argmin}_{\cancel{\|f\|_{\mathcal{H}} \leq N}\ f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2$$

Solution: $\hat{f} = \sum_{j=1}^n \hat{\beta}_j \, k(x_j,\cdot)$ where

$$\hat{\beta} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^n \ell(y_i, (K\beta)_i) + \lambda \beta^T K \beta$$

where $K = (k(x_i,x_j))_{i,j=1,\dots,n}$.

again we have



Yet: in deep-learning or Random Forests

→ number p of parameters $\gg n$ = sample size

→ no regularisation (interpolation of data points)

$\Leftrightarrow$ work empirically very well!

Why? Something to do with optimisation algorithm?

② Simple least square regression

$$\underset{\in\mathbb{R}^n}{Y} = \underset{\in\mathbb{R}^p}{X\beta} + \varepsilon \quad \text{, with } \mathbb{E}[\varepsilon]=0, \ \operatorname{cov}(\varepsilon)=\sigma^2 I_n.$$

$$\ell(y,y') = (y-y')^2$$

So $\mathcal{L}(\beta) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, x_i^T\beta) = \frac{1}{n}\|Y-X\beta\|^2$

$$\Rightarrow \hat{\beta} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \|Y-X\beta\|^2.$$

classical case: $p \leq n$ and $\operatorname{rank}(X)=p$ (full rank)

Then $\hat{\beta} = (X^TX)^{-1}X^TY$

- $\mathbb{E}[\hat{\beta}] = \beta^*$ (no bias)
- $\operatorname{cov}(\hat{\beta}) = \sigma^2 (X^TX)^{-1}X^TX(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$

Hence

$$\mathbb{E}[\|\hat{\beta}-\beta\|^2] = \|\hat{\beta}-\mathbb{E}(\hat{\beta})\|^2 + \operatorname{Tr}\operatorname{cov}(\hat{\beta}) = \sigma^2 \operatorname{Tr}(X^TX)^{-1}$$

typically diverge when $p \to n$.

Interpolation case: $p \geq n$, $\text{rank}(X) = n$.

$\rightarrow \min_{\beta} \|Y - X\beta\|^2 = 0$

$\rightarrow$ no unique solution to $Y = X\beta$

Lemma: $\mathcal{G} = \{\beta : Y = X\beta\}$ is equal to $\hat{\beta}_0 + \ker(X)$ where $\hat{\beta}_0 = X^T(XX^T)^{-1}Y$

\# $X\hat{\beta}_0 = Y$ and $X(\beta - \hat{\beta}_0) = 0 \quad \forall \beta \in \mathcal{G}$ \#

Remark: $X : \mathbb{R}^p = \underbrace{\ker(X)}_{\dim = p-n} \oplus \underbrace{\text{Im}(X^T)}_{\dim = n} \rightarrow \mathbb{R}^n = \text{Im}(X)$

So $\hat{\beta}_0 \perp \ker(X)$.

question: if we minimise $\mathcal{L}(\beta) = \frac{1}{n}\|Y - X\beta\|^2$ by gradient descent started at 0, where do we converge?

GD: $\beta^{(k+1)} = \beta^{(k)} - \frac{2\eta}{n} X^T(X\beta^{(k)} - Y)$ as $\nabla\mathcal{L}(\beta) = \frac{2}{n}X^T(X\beta - Y)$

if $\beta^{(0)} = 0$ then $\beta^{(k)} \in \text{Im}(X^T) \quad \forall k \geq 0$.

So $\beta^{(k)} \xrightarrow[k\to\infty]{} \hat{\beta}_0$!

Message: Gradient Descent select a specific solution.

Lemma: if $X \overset{SVD}{=} \sum_{k=1}^{n} \sigma_k u_k v_k^T$

$$\hat{\beta}_0 = \sum_{k=1}^{n} \frac{1}{\sigma_k} v_k u_k^T Y$$

$$= \underset{\beta \in \mathcal{G}}{\text{argmin}} \|\beta\|^2$$

$$= \lim_{\lambda \searrow 0} \underbrace{(X^TX + \lambda I)^{-1}X^TY}_{\underset{\beta}{\text{argmin}}\{\|Y - X\beta\|^2 + \lambda\|\beta\|^2\}}.$$

Proof:

- for $\hat{\beta} \in \mathcal{G}$, $\hat{\beta} = \hat{\beta}_0 + \underset{\in \ker(X)}{v}$

as $\hat{\beta}_0 \perp \ker(X)$: $\|\hat{\beta}\|^2 = \|\hat{\beta}_0\|^2 + \|v\|^2 \geq \|\hat{\beta}_0\|^2$

so $\hat{\beta}_0 \in \underset{\beta \in \mathcal{G}}{\text{argmin}} \|\beta\|^2$.

- $X^T(XX^T)^{-1} = \sum_k \sigma_k v_k u_k^T \times \sum_k \frac{1}{\sigma_k^2} u_k u_k^T$

$$= \sum_k \frac{1}{\sigma_k} v_k u_k^T$$

- $(X^TX + \lambda I)^{-1}X^T = \left(\sum_{k=1}^{n} \frac{1}{\sigma_k^2 + \lambda} v_k v_k^T + \frac{1}{\lambda} P^{\perp}\right) \sum_{k=1}^{n} \sigma_k v_k u_k^T$

$$= \sum_{k=1}^{n} \frac{\sigma_k}{\sigma_k^2 + \lambda} v_k u_k^T$$

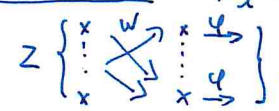
$$\xrightarrow[\lambda \searrow 0]{} \sum_{k=1}^{n} \frac{1}{\sigma_k} v_k u_k^T$$

□

Paper: analyse $\hat{\beta}_0$ in different settings:

- linear: $X_i = \Sigma^{1/2} Z_i$, Z_i iid, $\mathbb{E}[Z_i]=0$, $\mathrm{cov}(Z_i)=I_p$, $X_i \in \mathbb{R}^p$

- non-linear: $X_i = \varphi(W Z_i)$, $Z_i \overset{iid}{\sim} \mathcal{N}(0, I_d)$, W prd with $W_{ij} \overset{iid}{\sim} \mathcal{N}(0, \frac{1}{d})$.



, mis-specified: some explanatory variables are missing.

- asymptotic setting: $p \sim \gamma n$ with $n \to \infty$ and $d \sim \psi_p$

Focus on: sections 2, 3 and 5 (linear setting).

- risk: $R_X(\hat{\beta}, \beta) = \mathbb{E}\left[(x_0^T\hat{\beta} - x_0^T\beta)^2 \mid X\right]$ where $x_0 \perp\!\!\!\perp X, Y$.

$= \mathbb{E}\left[\|\hat{\beta}-\beta\|_\Sigma^2 \mid X\right]$ where $\|\beta\|_\Sigma^2 = \beta^T\Sigma\beta$.

Lemma: $\mathbb{E}\left[\|Z\|_\Sigma^2\right] = \|\mathbb{E}[Z]\|_\Sigma^2 + \mathrm{Tr}(\Sigma\,\mathrm{cov}(Z))$

\# $Z^T\Sigma Z = \mathbb{E}[Z]^T\Sigma\,\mathbb{E}[Z] + \underbrace{(Z-\mathbb{E}[Z])^T\Sigma(Z-\mathbb{E}[Z])}_{=\langle \Sigma, (Z-\mathbb{E}[Z])(Z-\mathbb{E}[Z])^T\rangle,\ (Z-\mathbb{E}[Z])(Z-\mathbb{E}[Z])^T = \mathrm{cov}(Z)}$

$+ \underbrace{2\,\mathbb{E}[Z]^T\Sigma(Z-\mathbb{E}[Z])}_{\mathbb{E}[\cdots]=0}$

\#

So $R_X(\hat{\beta}, \beta) = \underbrace{\|\mathbb{E}[\hat{\beta}\mid X]-\beta\|_\Sigma^2}_{B_X(\hat{\beta},\beta)} + \underbrace{\mathrm{Tr}(\Sigma\,\mathrm{cov}(\hat{\beta}\mid X))}_{V_X(\hat{\beta},\beta)}$.

③ Linear case (sections 2 and 3)

X is full rank: $\mathrm{rank}(X) = \min(n,p)$.

classical regime: $n \geq p$ i.e. $\gamma \leq 1$

$R_X[\hat{\beta},\beta] = 0 + \sigma^2\,\mathrm{Tr}(\Sigma(X^TX)^{-1}) \xrightarrow{\text{Thm 1}} \sigma^2\frac{\gamma}{1-\gamma}$

interpolation regime: $p \geq n$ i.e $\gamma \geq 1$

- $\mathbb{E}[\hat{\beta}_0 \mid X] = \mathbb{E}\left[X^T(XX^T)^{-1}Y \mid X\right]$

$Y = X\beta+\varepsilon \Rightarrow \underbrace{X^T(XX^T)^{-1}X}_{P_{\mathrm{Im}(X^T)}}\beta + 0$

So $\beta - \mathbb{E}[\hat{\beta}_0\mid X] = \underbrace{(I - P_{\mathrm{Im}(X^T)})}_{=P_{\ker(X)}}\beta$

$B_X(\hat{\beta},\beta) = \|P_{\ker(X)}\beta\|_\Sigma^2$

- $\mathrm{cov}(\hat{\beta}_0\mid X) = \sum_{k=1}^n \frac{1}{s_k} v_k u_k^T\,\underbrace{\mathrm{cov}(Y\mid X)}_{\sigma^2 I_n}\left(\sum_{k=1}^n \frac{1}{s_k} v_k u_k^T\right)^T$

$= \sigma^2 \sum_{k=1}^n \frac{1}{s_k^2} v_k v_k^T$

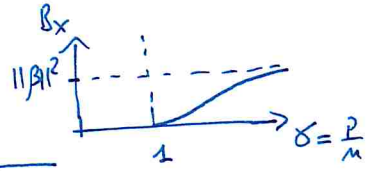
$$V_X(\hat\beta,\beta) = \mathrm{Tr}\left(\Sigma\, \mathrm{cov}(\hat\beta_0 \mid X)\right)$$

$$= \sigma^2 \sum_{k=1}^{m} \frac{1}{\tau_k^2} V_k^T \Sigma V_k .$$

Lemma 2: $\Sigma = I_p$, $\gamma > 1$

$$B_X(\hat\beta,\beta) \longrightarrow \|\beta\|^2 \left(1 - \frac{1}{\gamma}\right)$$

increase with γ !!



\# idea: $u := \frac{\beta}{\|\beta\|}$

$$B_X(\hat\beta,\beta) = \| P_{\ker(X)}\, \beta\|^2 = \|\beta\|^2 \| P_{\ker(X)}\, u\|^2$$

$$\ker(X) = \mathrm{span}\langle w_1, \dots, w_{p-m}\rangle$$

$\uparrow$ $\uparrow$

$\perp$ random vectors (unit norm)

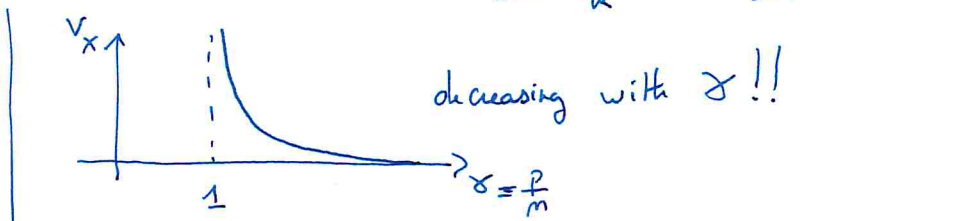
$$\| P_{\ker(X)}\, u\|^2 = \sum_{k=1}^{p-m} \langle w_k, u\rangle^2$$

we have $\mathbb{E}\left[\langle w_k, u\rangle^2\right] = \frac{1}{p}$ as $\sum_{k=1}^{p} \langle w_k, u\rangle^2 = 1$ ($\|u\|^2 = 1$)

so $\| P_{\ker(X)}\, u\|^2 \simeq \frac{p-m}{p} = 1 - \frac{1}{\gamma}$.

\#

Lemma 3: $V_X(\hat\beta,\beta) = \sigma^2 \sum_{k=1}^{\gamma p} \frac{1}{\tau_k^2} \longrightarrow \frac{\sigma^2}{\gamma - 1}$

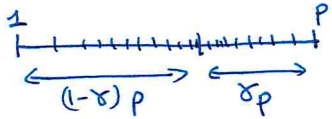


\# idea: Marchenko-Pastur:

$$\frac{1}{p} \sum_{k=1}^{p} \delta_{\sigma_k^2/p}(x) \longrightarrow \mu_\gamma(dx) = \frac{\sqrt{\left((1+\sqrt{\gamma})^2 - x\right)\left(x - (1-\sqrt{\gamma})^2\right)}}{2\pi x}\, dx.$$

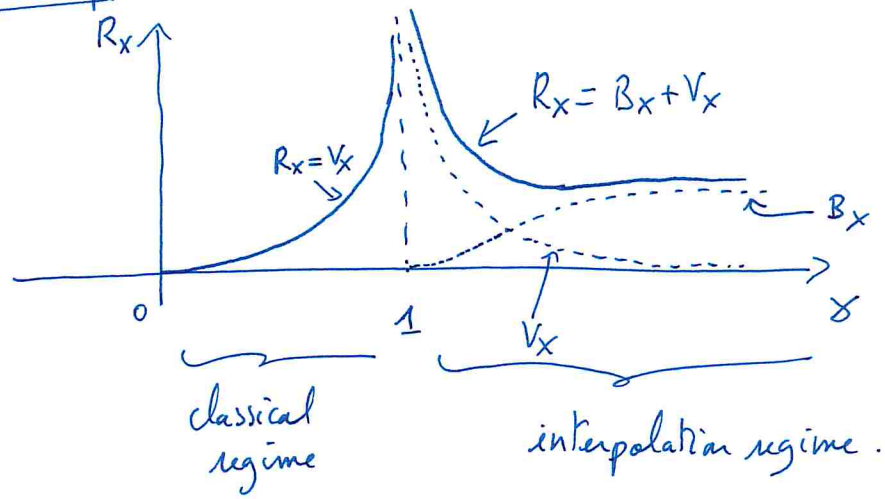
So $\sum_{k=1}^{\gamma p} \frac{1}{\tau_k^2} = \frac{1}{p} \sum_{k=1}^{\gamma p} \frac{1}{\sigma_k^2/p} \longrightarrow \int_{Q_{1-\gamma}}^{+\infty} \frac{1}{x}\, \mu_\gamma(dx) = \frac{1}{\gamma - 1}$

where $F_{\mu_\gamma}(Q_{1-\gamma}) = 1 - \gamma$

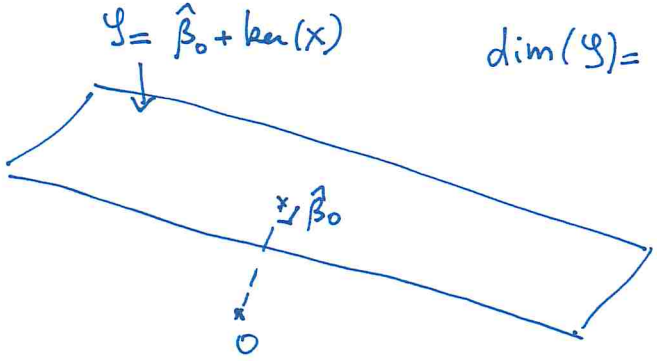


\#

Sum-up



intuition?



when $p \nearrow$, $\mathcal{S}$ becomes larger, so more solutions to $Y = X\beta$

$\Rightarrow \|\hat{\beta}_0\| \searrow 0$

$\Rightarrow$ { - bias $\nearrow$; - variance $\searrow$

Yet: this simple setting miss the approximation / variance trade-off: best at $\gamma \equiv 0$

$\rightsquigarrow$ misspecified setting.

④ Mis-specified case (section 5)

observe (X_i, Y_i) iid with

$$Y_i = X_i^T \beta + \underbrace{W_i^T \theta}_{\text{unobserved, iid.}} + \varepsilon_i$$

risk: compare $X_0^T \hat{\beta}$ to $\mathbb{E}[Y_0 \mid X_0, W_0] = X_0^T\beta + W_0^T\theta$.

$$R_X(\hat{\beta}, \beta) := \mathbb{E}\left[\left(X_0^T\hat{\beta} - \mathbb{E}[Y_0 \mid X_0, W_0]\right)^2 \mid X\right]$$

Pythagore $=$

$$\mathbb{E}\left[\left(X_0^T\hat{\beta} - \mathbb{E}[Y_0 \mid X_0]\right)^2 \mid X\right] + \underbrace{\mathbb{E}\left[\left(\mathbb{E}[Y_0 \mid X_0] - \mathbb{E}[Y_0 \mid X_0, W_0]\right)^2 \mid X\right]}_{\text{approximation bias} \;=\; \Pi_\theta}$$

- Π_θ complex in general.
- when all entries of X_i, W_i are iid:

$$\mathbb{E}[Y_0 \mid X_0] = X_0^T\beta \quad \text{and} \quad \mathbb{E}[Y_0 \mid X_0, W_0] = X_0^T\beta + W_0^T\theta$$

$$\Rightarrow \Pi_\theta = \mathbb{E}\left[(W_0^T\theta)^2 \mid X\right] = \|\theta\|^2$$

$$= r^2(1 - \kappa)$$

$r^2 = \underbrace{\|\theta\|^2 + \|\beta\|^2}_{\text{signal strength}}$; $\kappa = \frac{\|\beta\|^2}{r^2}$) fraction of signal explained by X_0.

more complex behaviours

→ famula theorem 4

→ case $1 - K(\gamma) = (1+\gamma)^{-a}$ (↔ Fourier)

Figure 3 and 4.

Sometimes best R_x for $\gamma > 1$.

Take Home Message:

→ optimisation algorithms select some specific solutions in the interpolation domain

→ can lead to opposite behavior o/s in the classical regime

→ some limits yet

i) $R_x(\hat{\beta}_0) \geq \min_{\lambda \geq 0} R_x(\hat{\beta}_\lambda)$ ← Ridge.

ii) full optimisation is not performed (usually) in the interpolation domain.

↳ early stopping ↔ implicit regularisation

iii) Neural Networking much more complex as features are learnt at the same time as β → the behavior might be very different.