A U-Turn on "A U-Turn on Double Descent"

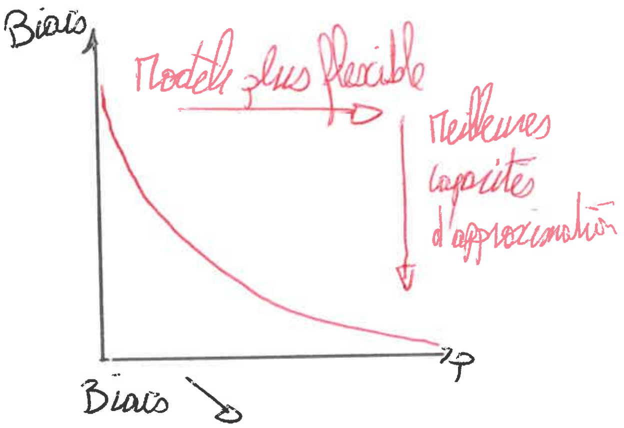
"A U-Turn on Double Descent: Rethinking Parameter Counting in Statistical learning"
[Curth, Jeffares, v.d. Schaar - 2023]

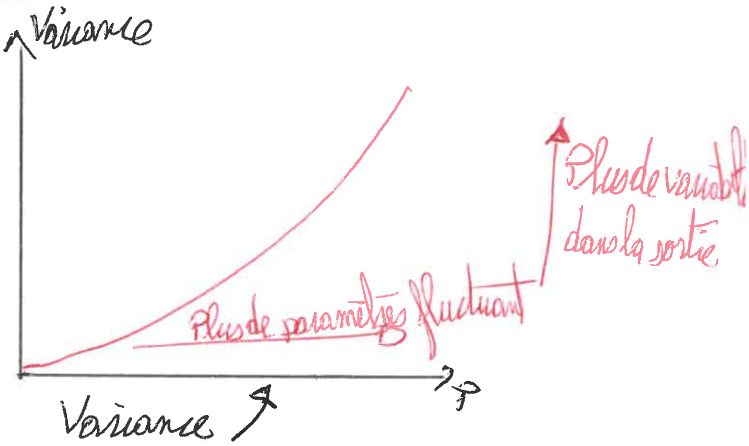
I Biais, Variance, Double descente

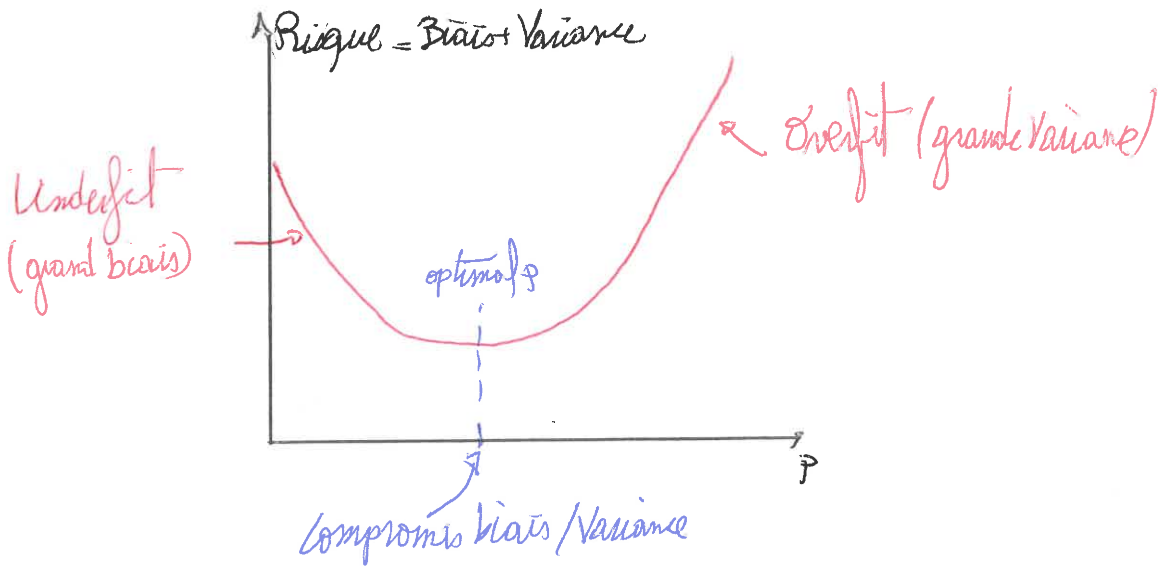
1) Phénomène de double descente

En théorie classique d'estimation, Risque = Biais + Variance

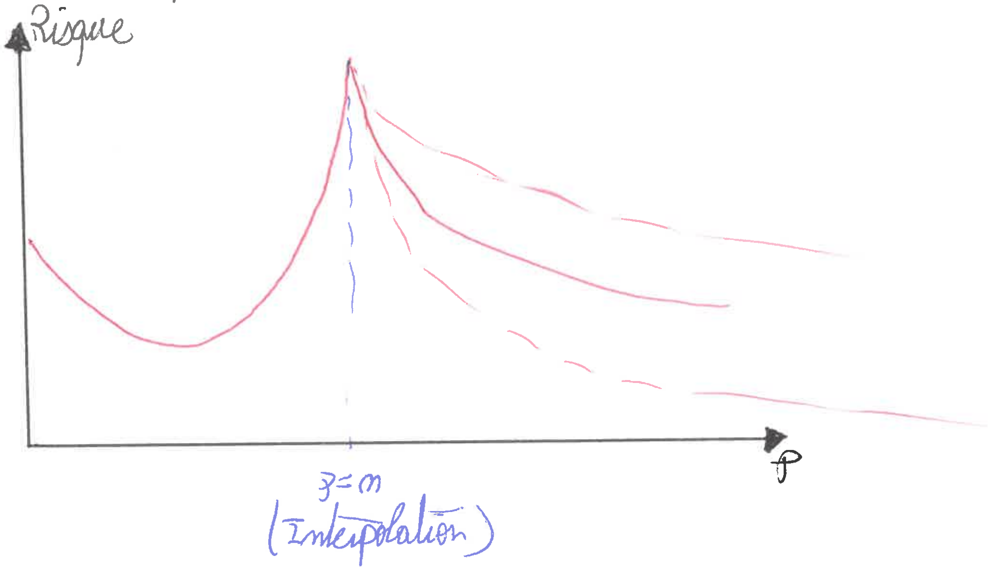
On note $\boxed{p := \#\{\text{paramètres dans le modèle}\}}$







Pourtant, des observations empiriques (notamment sur des reseaux de neurones) donnent



Phénomène de "Double Descente" (Belkin et. al 2019)

(A Brief Prehistory of Double Descent)

Programme :

- Exposer un résultat d'obligation au compromis biais / variance
- Donner des pistes sur une notion de dimension effective p_{eff} en regression linéaire featurisée (Regression RKHS)
- Tenter de faire l'exégèse du papier

2) Le compromis biais/Variance est incontournable

"On lower bounds for the bias-variance trade-off"

[Derumigny, Schmidt-Hieber, 2023]

Dans des régimes surparamétrés, biais et variance ne s'expriment pas de façon simple en fonction de p ($\Longrightarrow$ Pas la courbe en forme de U classique), mais ça ne veut pas dire que ce compromis est caduc.

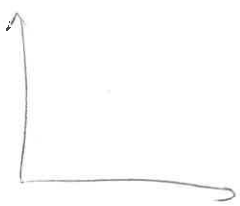
Rappel : Si $\begin{cases} \theta \in \Theta \subset \mathbb{R} \\ \hat{\theta} \text{ est un estimateur} \\ \text{modèle } p_\theta(x)\,dx = P_\theta(dx) \text{ régulier} \end{cases}$ on note $\begin{cases} B(\theta) = \mathbb{E}_\theta \hat{\theta} - \theta \\ V(\theta) = \mathrm{Var}_\theta(\hat{\theta}) \end{cases}$, alors

Cramer-Rao (≃ Cauchy-Schwarz)

$$V(\theta) \geq \frac{\left(1 + B'(\theta)\right)^2}{F(\theta)}$$

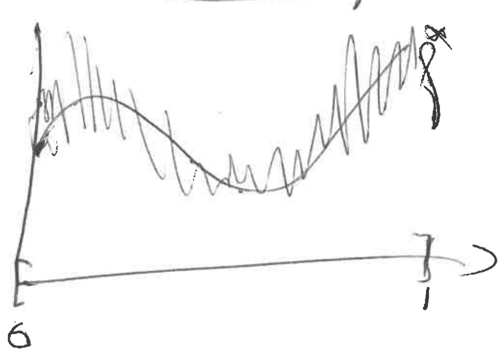
Information de Fisher $\mathbb{E}_\theta\left[\partial^2 \log p_\theta(X)\right]$

- Variance minorée si $B(\theta) = 0$
- ——— si $B'(\theta) \leq \frac{1}{2}$



Pour des problèmes non-paramétrique (regression), l'existence d'estimateurs non-biaisés n'est pas toujours garantie.

On considère le modèle $\boxed{dY_x = f^*(x)\,dx + \frac{1}{\sqrt{n}}\,dW_x}$, $x \in [0,1]$



Pour $\begin{cases} x_0 \in [0,1] \\ h > 0 \end{cases}$ tels que $0 \leq x_0 - h \leq x_0 + h \leq 1$, on considère

$$\hat{f}_h(x_0) := \frac{1}{2h} \int_{x_0-h}^{x_0+h} dY_t$$

$$\text{Biais}_f(\hat{f}_h) = \frac{1}{2h} \int_{x_0-h}^{x_0+h} \left(f^*(u) - f^*(x_0) \right) du$$ → Dépend de la regularité de f.

$$\text{Var}_f(\hat{f}_h) = \left(\frac{1}{2h}\right)^2 \text{Var}\Big(\underbrace{\int_{x_0-h}^{x_0+h} \frac{dW_s}{\sqrt{n}}}_{\frac{1}{\sqrt{n}}(B_{x_0+h} - B_{x_0-h}) \sim N\left(0, \frac{2h}{n}\right)} \Big) = \frac{1}{2nh}$$ → Ne dépend pas de f

Si $f^* \in \mathcal{C}^\beta$, $\text{Biais}(\hat{f}_h)^2 \lesssim h^{2\beta}$, de sorte que

$$\left(\sup_{f \in \mathcal{C}^\beta} |\text{Biais}_f(\hat{f})|^{1/\beta} \right) \left(\sup_{f \in \mathcal{C}^\beta} \text{Var}_f(\hat{f}) \right) \gtrsim \frac{1}{n}$$

→ Derumigny & Schmidt-Hieber demontrent que c'est vrai pour tout estimateur $\hat{f}$ de biais borné (Sinon $\hat{f} = 0$ a variance nulle $\forall f$)

→ Obligation d'avoir les ordres de grandeur classiques pour être minimax
→ Fonction en regression ponctuelle et $L^2([0,1])$.
→ la demonstration utilise :
- Des inégalités d'information
- Une réduction au modèle de la séquence gaussienne
  aux estimateurs invariants par rotation ($\Rightarrow$ Pas de Stein très bon en $\theta = 0$)
- Un argument de symétrie type Borsuk (épaisseur de Bertin ?)

II Double descente en régression featurisée (Kernel regression)

1) Features et moindres carrés

On observe $\boxed{Y_i = f^*(X_i) + \varepsilon_i}$ avec $\begin{cases} X_i \in \mathcal{X} = \mathbb{R}^d \ \boxed{\text{fixes}} \\ \varepsilon_i \text{ iid } \mathbb{E}_\varepsilon[\varepsilon] = 0 \\ \in \mathbb{R} \end{cases}$ Échantillon d'entraînement (train) $\boxed{i \leq n}$

Ici $\boxed{f^* : \mathcal{X} \longrightarrow \mathbb{R}}$ est la fonction de régression qu'on cherche à estimer

⚠ Comme on va se placer dans des régimes sur-paramétrés où $\hat{f}(X_i) = Y_i \ \forall i$, on va évaluer la qualité des estimateurs $\hat{f} : \mathcal{X} \to \mathbb{R}$ sur

$y_j = f^*(x_j) + \varepsilon'_j$ avec $\begin{cases} x_j \in \mathcal{X} = \mathbb{R}^d \ \boxed{\text{fixes}} \\ \varepsilon'_j \text{ iid } \mathbb{E}_\varepsilon[\varepsilon] = 0 \end{cases}$ Échantillon de test (test)

$j \leq m$

Risque de généralisation : $\forall g : \mathcal{X} \to \mathbb{R}$,

$$\boxed{\begin{aligned} R_{test}(g) &:= \sum_{j=1}^{m} \mathbb{E}_{\varepsilon'_j} \left(y_j - g(x_j) \right)^2 \\ &= \mathbb{E}_{\varepsilon'} \| \underbrace{Y_{test}}_{\in \mathbb{R}^m} - g(\underbrace{X_{test}}_{\in (\mathbb{R}^d)^m}) \|^2_{\ell^2(\mathbb{R}^m)} \end{aligned}}$$

Pour construire un estimateur, on se base sur l'erreur empirique d'entraînement

$$\boxed{\hat{L}_{train}(g) := \| \underbrace{Y_{train}}_{\in \mathbb{R}^n} - g(\underbrace{X_{train}}_{\in (\mathbb{R}^d)^n}) \|^2_{\ell^2(\mathbb{R}^{n_{train}})}}$$

Pour tout vectoriser et prendre en compte les non-linéarités, on se donne une feature map $\boxed{\phi : \mathcal{X} = \mathbb{R}^d \longrightarrow \mathbb{R}^p}$

⚠ $\phi = \phi_p$ avec p qui variera

Ex : $\mathcal{X} = \mathbb{R}$, $\phi(x) = (1, x, x^2, \dots, x^{p-1})$ monomes

$\mathcal{X} = [-1,1]$, $\phi(x) = (T_0(x), \dots, T_{p-1}(x))$ Tchebychev

$\mathcal{X} = \mathbb{R}^d$, $\phi(x) = (\cos(\langle v_1, x\rangle), \dots, \cos(\langle v_p, x\rangle))$ Random Fourier Features

$v_k \overset{iid}{\sim} N(0,I)$

$\mathcal{X} = \mathbb{R}^d$, $\phi(x) = (\langle v_1, x - a_1\rangle_+, \dots, \langle v_p, x - a_p\rangle_+)$ Réseaux de neurones ReLU à 1 couche (Perceptron)

Pour tout $\beta \in \mathbb{R}^p$, $x \in \mathcal{X}$

$$\boxed{\begin{aligned} g_\beta(x) &:= \langle \phi(x), \beta \rangle_{\ell^2(\mathbb{R}^p)} \\ &= \phi(x)^T \beta \end{aligned}}$$

est le regresseur associé.

En notant $\Phi_{train} := \begin{pmatrix} \phi(X_1)^T \\ \vdots \\ \phi(X_n)^T \end{pmatrix} \in \mathbb{R}^{n \times p}$, on obtient l'expression matricielle

$$\hat{L}_{train}(g_\beta) = \| Y_{train} - \Phi_{train} \beta \|^2_{\ell^2(\mathbb{R}^n)}$$

2) Moindres carrés explicites et biais implicites

$\beta \longmapsto \hat{L}_{train}(g_\beta) = \|Y_{train} - \Phi_{train}\beta\|^2_{\ell^2(\mathbb{R}^n)}$ est une forme quadratique minorée par 0, elle admet donc un minimum global (point critique). De plus,

$$\hat{L} = \langle \Phi_{train}\beta - Y_{train}, \Phi_{train}\beta - Y_{train} \rangle$$

$$= \langle \beta, \Phi_{train}^T \Phi_{train}\beta \rangle - \langle 2\Phi_{train}^T Y_{train}, \beta \rangle + \langle Y_{train}, Y_{train} \rangle$$

D'où $\nabla_\beta \hat{L} = 2\Phi_{train}^T(\Phi_{train}\beta - Y_{train}) = 0$

$$\iff \Phi_{train}^T \Phi_{train}\beta = \Phi_{train}^T Y_{train}$$

Aparté algèbre linéaire

Si $A \in \mathbb{R}^{u\times v}$, $A^+ \in \mathbb{R}^{v\times u}$ est appelée inverse généralisée de Moore-Penrose de A lorsque

(i) $AA^+A = A$

(ii) $A^+AA^+ = A^+$

Existe toujours

(iii) AA^+ est symétrique $(= \pi_{Im(A)})$

(iv) A^+A est symétrique $(= \pi_{Im(A^T)} = \pi_{Ker(A)^\perp})$

Rend A^+ unique

Si $Ax = b$ admet au moins une solution, alors $x = A^+b$ est la solution de norme minimale, et toute solution s'écrit

$$x = A^+b + \underbrace{(I_{v\times v} - A^+A)}_{\pi_{Ker(A)}} w, \quad w \in \mathbb{R}^v$$

Construction avec SVD : Si $A = U\Sigma V^T$ avec $\begin{cases} U \in \mathcal{O}_{u\times u} \\ V \in \mathcal{O}_{v\times v} \\ \Sigma \text{ diagonale } \mathbb{R}^{u\times v} \ (\sigma_\ell)_\ell \end{cases}$

alors $A^+ = V\Sigma^+ U$

où $\underset{\in \mathbb{R}^{v\times u}}{\Sigma^+} = (\sigma_\ell^+)_\ell$, $\sigma^+ = \begin{cases} \sigma^{-1} & \text{si } \sigma \neq 0 \\ 0 & \text{sinon} \end{cases}$

Construction analytique $A^+ = \lim_{\lambda \searrow 0} (A^T A + \lambda I_{v\times v})^{-1} A^T$

$$= \lim_{\lambda \searrow 0} A^T (AA^T + \lambda I_{u\times u})^{-1}$$

Cette dernière expression permet de voir que l'inverse généralisé intervient naturellement lorsqu'on obtient $\hat{\beta}$ par descente de gradient :

Appliquons une descente de gradient à pas fixe $\eta > 0$ partant de $\hat{\beta}^0 = 0$ pour $\beta \mapsto \|Y_{train} - \Phi_{train}\beta\|$:

$$\beta^{t+1} \overset{GD}{:=} \beta^t - \eta \times 2\Phi_{train}^T(\Phi_{train}\hat{\beta}^t - Y_{train})$$

$$\overset{\hat{\beta}^0=0}{=} \underbrace{\sum_{k=0}^{t} (I_{p\times p} - 2\eta\Phi_{train}^T\Phi_{train})^k}_{\text{Série de Von Neuman}} \cdot 2\eta\Phi_{train}^T Y_{train}$$

$$\xrightarrow[t\to\infty]{2\eta < \|\Phi_{train}\|_{op}^{-2}} (I - (I - 2\eta\Phi_{train}^T\Phi_{train}))^+ \cdot 2\eta\Phi_{train}^T Y_{train}$$

$$= (\Phi_{train}^T\Phi_{train})^+ \Phi_{train}^T Y_{train} \quad = \Phi_{train}^+ \underbrace{(\Phi_{train}^T)^+ \Phi_{train}^T}_{\pi_{\mathrm{Ker}(\Phi_{train}^T)^\perp} = \mathrm{Ker}(\Phi_{train}^+)^\perp} Y_{train}$$

$$= \Phi_{train}^+ Y_{train}$$

Fin de l'aparté

3) Étude de l'erreur de généralisation

• Le regresseur moindre carrés est $g_{\hat{\beta}_{LS}}(x) = \langle \hat{\beta}_{LS}, \phi(x) \rangle$

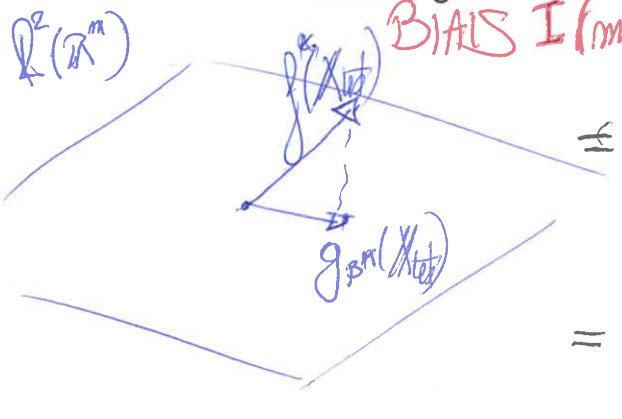
$$= \langle \underbrace{(\Phi_{train}^{+})^T \phi(x)}_{\text{indépendant de } Y_{train}!}, Y_{train} \rangle_{\ell^2(\mathbb{R}^p)}$$

• Ce regresseur ne peut pas faire mieux que le meilleur g_{β^*} pour l'erreur de test

$$\beta^* \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \underbrace{\mathbb{E}_{\varepsilon'_{test}} \| Y_{test} - g_\beta(X_{test}) \|^2_{\ell^2(\mathbb{R}^m)}}_{= \| f^*(X_{test}) - g_\beta(X_{test}) \|^2_{\ell^2(\mathbb{R}^m)} + \mathbb{E}_{\varepsilon'}\left[\| \varepsilon_{test} \|^2_{\ell^2(\mathbb{R}^m)} \right]}$$

On étudie donc l'excès de risque, qui s'écrit par polarisation

$$R(g_\beta) - \underbrace{R(g_{\beta^*})}_{\text{BIAIS I (modèle) (nul si } p \gg m)} = \mathbb{E}_{\varepsilon'}\left[\langle g_{\beta^*}(X_{test}) - g_\beta(X_{test}), 2 f^*(X_{test}) + 2 \overset{\text{centré}}{\varepsilon_{test}} - g_{\beta^*}(X_{test}) - g_\beta(X_{test}) \rangle_{\ell^2(\mathbb{R}^m)} \right]$$



$$= \mathbb{E}\left[\langle g_{\beta^*}(X_{test}) - g_\beta(X_{test}), 2(f^* - g_{\beta^*}) + (g_{\beta^*} - g_\beta) \rangle \right]$$

$$= \| g_{\beta^*}(X_{test}) - g_\beta(X_{test}) \|^2_{\ell^2(\mathbb{R}^m)}$$

D'autre part, on a $\boxed{\beta^* = \Phi_{test}^{+} f^*(X_{test})}$ si l'on considère l'élément de norme minimale

Comme $g_\beta(X_{test}) = \Phi_{test} \beta$ par définition, on obtient

$$R(g_{\hat{\beta}_{LS}}) - R(g_{\beta^*}) = \left\| \Phi_{test} \left(\Phi_{test}^{+} \beta^*(X_{test}) - \Phi_{train}^{+} Y_{train} \right) \right\|^2_{\ell^2(\mathbb{R}^m)}$$

On peut enfin recoller au papier "U-Turn on Double Descent" en effectuant une décomposition biais variance en l'aléa ε_{train} (i.e Y_{train})

$$\boxed{\mathbb{E}_{\varepsilon_{train}} \left(R(g_{\hat{\beta}_{LS}}) - R(g_{\beta^*}) \right) = \underbrace{\left\| \Phi_{test} \left(\Phi_{test}^{+} \beta^*(X_{test}) - \Phi_{train}^{+} \beta^*(X_{test}) \right) \right\|^2_{\ell^2(\mathbb{R}^m)}}_{\text{Biais II (Transfert)}} + \underbrace{\mathbb{E}_{\varepsilon_{train}} \left\| \Phi_{test} \Phi_{train}^{+} \varepsilon_{train} \right\|^2_{\ell^2(\mathbb{R}^m)}}_{\text{Variance}}}$$

Hypothèse : $Cov(\varepsilon) = \sigma^2 I_{n \times n}$.

Alors la variance s'écrit

$$\boxed{\mathbb{E}_{\varepsilon_{train}} \left\| \Phi_{test} \Phi_{train}^{+} \varepsilon_{train} \right\|^2 = \sigma^2 \overbrace{\left\| \Phi_{test} \Phi_{train}^{+} \right\|^2_{Frobenius}}^{p_{\text{Effective}}}}$$

$p_{\text{Effective}}$: dimension effective du modèle.

$$\begin{aligned} \mathbb{E} \|A\varepsilon\|^2 &= \mathbb{E}\, Tr(A \varepsilon \varepsilon^T A^T) \\ &= Tr(A\, \mathbb{E}(\varepsilon\varepsilon^T) A^T) \\ &= \sigma^2 Tr(AA^T) \\ &= \sigma^2 \|A\|_F^2 \end{aligned}$$

Somme sur tous les $(x_j)_j$ (du test) de

$$\begin{aligned} Var_{\varepsilon}\left(g_{\hat{\beta}_{LS}}(x_j) \right) &= Var_{\varepsilon} \left(\langle \Phi_{train}^{+} Y_{train}, \phi(x_j) \rangle \right) \\ &= Var_{\varepsilon} \left(\langle \varepsilon, (\Phi_{train}^{+})^T \phi(x_j) \rangle \right) \\ &= \sigma^2 \left\| (\Phi_{train}^{+})^T \phi(x_j) \right\|^2 \end{aligned}$$

4) Interprétation

a) Géométrie de la variance

Pour le terme de variance uniquement, on a $\forall x_0 \in \mathcal{X}$ fixé

$$\begin{aligned}
\mathrm{Var}_{\varepsilon}\left(g_{\hat{\beta}_{LS}}(x_0)\right) = 0 &\iff \phi(x_0) \in \mathrm{Ker}\left(\left(\Phi_{train}^{+}\right)^{T}\right) \\
&\iff \phi(x_0) \in \mathrm{Im}\left(\Phi_{train}^{+}\right)^{\perp} \\
&\iff \phi(x_0) \in \mathrm{Im}\left(\Phi_{train}^{T}\right)^{\perp} \\
&\iff \phi(x_0) \in \mathrm{Ker}\left(\Phi_{train}\right) \\
&\implies g_{\hat{\beta}_{LS}}(x_0) = 0
\end{aligned}$$

$\Rightarrow$ Pour une donnée dans une région de l'espace jamais vue, on répond 0 et donc
{ variance nulle
{ possible grand biais

$\Rightarrow$ De même, $\phi(x_{test})$ quasi orthogonal à $\mathrm{Vect}(\phi(X_1), \dots, \phi(X_n))$ renvena $\simeq 0$

En sommant sur X_{test}, on mélange alors des régions de l'espace avec
moyennisé

possiblement | petit biais / grande variance
| grand biais / petite variance

b Grande dimension et petite variance

(Si $A : E \to F$, $\text{rang}\, A + \dim \text{Ker}\, A = \dim E$)

Par théorème du rang,

$$\dim \text{Ker}(\Phi_{train}) + \underbrace{\text{rang}\, \Phi_{train}}_{\leq n} = p$$

$$\Rightarrow \boxed{\dim \text{Ker}(\Phi_{train}) \geq p - n}$$

$\Rightarrow$ Quand $p \nearrow$, la taille de l'espace où la variance est petite $\nearrow$

Cela explique pourquoi la variance decroît apres $p = n$, mais ça neglige complètement le biais ☹

Rq : Possibilité d'écrire tout ça avec des noyaux

↳ Decroissance du spectre

↳ Matrice de Gram croisée entre X_{train} et X_{test}

III Ce que semble présenter "A U-Turn"

Le message principal est que pour les estimateurs linéaires (smoothers) du type

$$\hat{g}(x_0) = \langle \underbrace{\hat{S}(x_0)}_{\in \mathbb{R}^n}, Y_{train} \rangle ,$$

le bon paramètre de dimension effective est

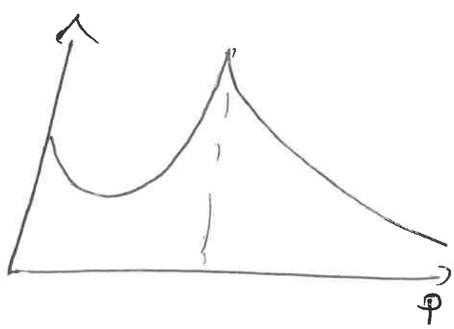
$$p_{eff} := \frac{1}{\sigma^2} \sum_{j \in test} Var_\varepsilon \, \hat{g}(x_j)$$

$$= \frac{1}{\sigma^2} \sum_{j \in test} \|\hat{S}(x_j)\|^2$$

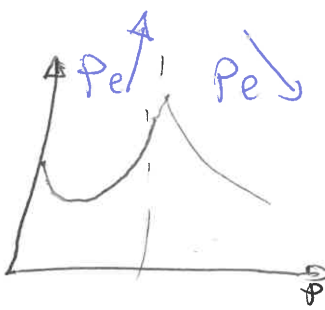
→ Mise en lumière de la notion de transfert (DEUX designs $\rightarrow X_{train}$, $\rightarrow X_{test}$)

→ Inclut
- Regression à noyaux
- Forêts aleatoires
- Gradient Boosting

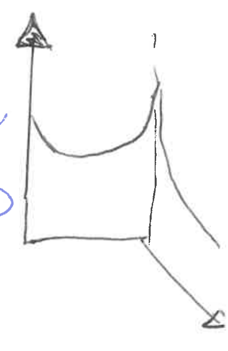
→ Simulations / Données réelles montrant un compromis biais Variance classique si l'on met p_{eff} en abcisse



Réinterprété en

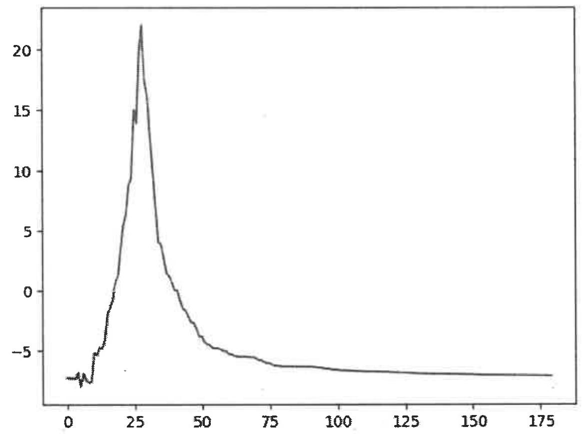
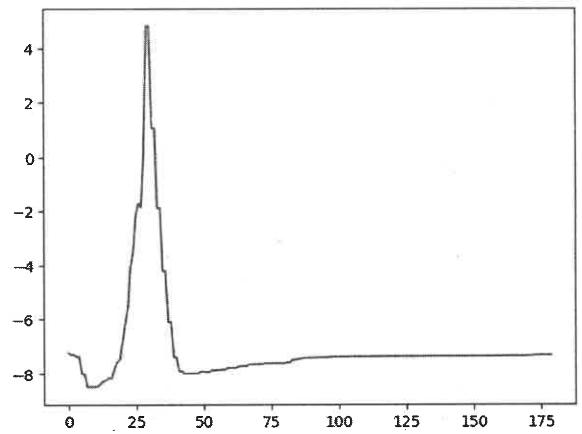


Dessiné en 3D



Chebychev polynomials

$n_{\text{train}} = 30, \; n_{\text{test}} = 1000000$

test \ train	Random	Equidistant
Random		
Equidistant