La théorie de la double descente

Emett Haddad [1]

27/06/2024

[1] Encadrants: Nicolas Vayatis, Samuel Gruffaz

Plan

Contexte

Fonction cible

On pose $\mathcal{X} = \mathbb{R}^D$ l'espace de départ et $\mathcal{Y} = \mathbb{R}$ l'espace d'arrivée.
$\mathbf{X}$ et $\mathbf{Y}$ sont des variables aléatoires tel que $(\mathbf{X}, \mathbf{Y}) \hookrightarrow \mathcal{P}$.

$$y^\star : x \in \mathcal{X} \to \mathbb{E}_{(\mathbf{X},\mathbf{Y})\sim\mathcal{P}}(\mathbf{Y}|\mathbf{X} = x) \in \mathcal{Y}$$

Échantillons

Échantillon d'apprentissage $\mathcal{D} := \{(x_n, y_n) \in \mathcal{X} \times \mathcal{Y}\}_{n=1}^N$ où
$y_n := y^\star(x_n) + \epsilon_n$, et les $x_n \hookrightarrow \mathbf{X}$ et $\epsilon_n \hookrightarrow \epsilon$ iid.
On modélise ici : $\mathbf{Y} = y^\star(\mathbf{X}) + \epsilon$ où ϵ représente le **bruit** tel que $\mathbb{E}(\epsilon|\mathbf{X}) = 0$, et $\mathbb{V}(\epsilon) = \sigma_\epsilon^2$.

Estimateur et risque

Estimateur

- Trouver un **estimateur** $\hat{y} : \mathcal{X} \to \mathcal{Y}$ tel que $\hat{y}(\mathbf{X}) \approx \mathbf{Y}$.
- $\hat{y} \in \mathcal{H}$ un espace de fonctions.

Vrai risque, risque empirique et excès de risque

Vrai risque et risque empirique: $\forall\, \hat{y} \in \mathcal{H}$

$$\mathcal{R}(\hat{y}) = \mathbb{E}_{(\mathbf{X},\mathbf{Y})\sim\mathcal{P}}((\mathbf{Y} - \hat{y}(\mathbf{X}))^2), \quad \hat{\mathcal{R}}_{\mathcal{D}}(\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}(x_i))^2$$

Excès de risque: $\forall\, \hat{y} \in \mathcal{H}$

$$\boxed{\mathcal{E}(\hat{y}) = \mathcal{R}(\hat{y}) - \mathcal{R}(y^\star)}$$

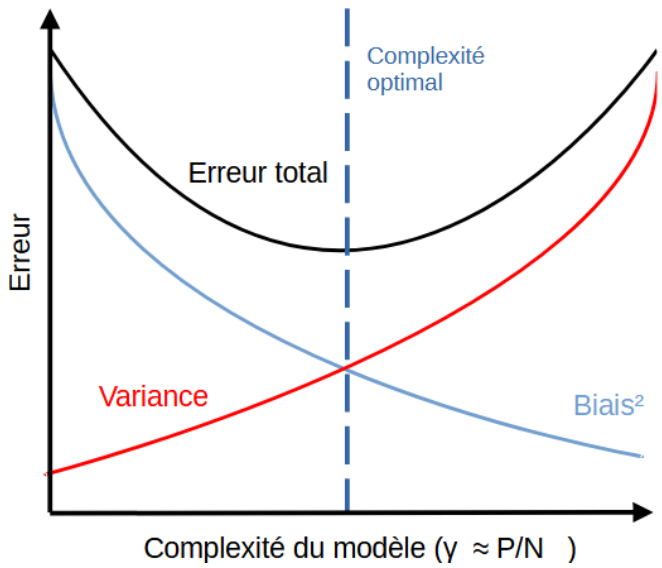
Notion de Double Descente

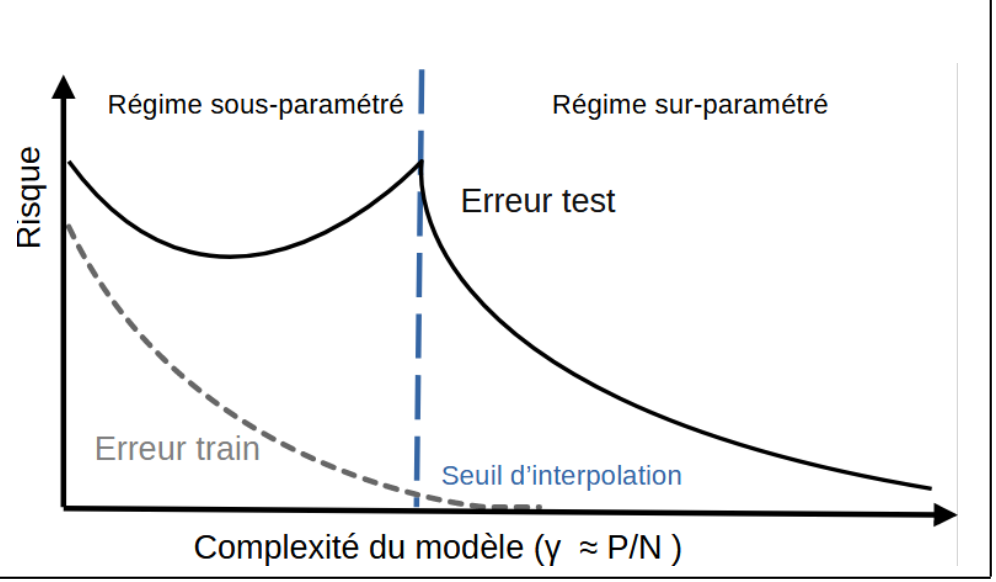
Régime sous-paramétré: $P < N$ et **régime sur-paramétré**: $P > N$
Seuil d'interpolation: $P = N$

Notion de Double Descente:

On dit qu'il y a **double descente** quand l'erreur globale minimal dans le régime sur-paramétré est inférieur à celle dans le régime sous-paramétré et qu'on observe un maximum au seuil d'interpolation.

Compromis biais-variance	Double descente





Modèles linéaire avec features

Modèle linéaire avec features

- $\mathcal{F} = \{f_i : \mathcal{X} \to \mathbb{R}, i \in \mathbb{N}\}$ ensemble de features.
- $X = [x_1, \cdots, x_N]^T \in \mathcal{M}_{N,D}(\mathbb{R})$, $Y = [y_1, \cdots, y_N]^T$
- $\forall x \in \mathcal{X},\ \Phi_P(x) = [f_1(x), \cdots, f_P(x)]^T$ et
  $Z = [\Phi_P(x_1), \cdots, \Phi_P(x_N)]^T = [f_j(x_i)] \in \mathcal{M}_{N,P}(\mathbb{R})$
- Risque empirique $\hat{\mathcal{R}}_{\mathcal{D}}(\hat{y}_\beta) = \frac{1}{N}||Y - Z\beta||_2^2$
- $\forall x \in \mathcal{X}$, $\boxed{\hat{y}_{\hat{\beta}}(x) = \Phi_P^T(x)\hat{\beta}}$, pour $\hat{\beta} = Z^{\dagger}Y$.
- Modèle linéaire simple $f_i(x) = e_i(x)$, $X = Z$, $P = D$ et $\hat{y}_{\hat{\beta}}(x) = x^T\hat{\beta}$.

Théorie modèle linéaire

On considère: $\Phi_P(x) = [f_1(x), \cdots, f_P(x)]^T$ où $f_p : \mathbb{R}^D \to \mathbb{R}$ et $\hat{y}(x) = \Phi_P^T(x)\hat{\beta}$.
Avec $(f,g)_X = \sum_{n=1}^{N} f(x_n)g(x_n)$, $\{f,g\} = \frac{1}{2\pi}\int_0^{2\pi} f(e^{i\theta})\overline{g(e^{i\theta})}$ et $\forall\, z \in \mathbb{C}, G_P^x(z) = \sum_{p=1}^{P} f_p(x)z^p$.

Théorème: Expressions de l'estimateur

Dans le cas linéaire c.à.d tel que $\hat{y}(x) = \Phi_P^T(x)\hat{\beta}$ nous avons:

$$\forall\, x \in \mathcal{X},\ \hat{y}(x) = [f_1(x), \cdots, f_P(x)][(f_i, f_j)_X]^{\dagger} \begin{bmatrix} (f_1, y)_X \\ \vdots \\ (f_P, y)_X \end{bmatrix} \tag{1}$$

$$\forall\, x \in \mathcal{X},\ \hat{y}(x) = ([\{G_P^{x_i}, G_P^{x_j}\}]^{\dagger}[\{G_P^x, G_P^{x_j}\}])^T Y \tag{2}$$

Théorème: Décroissance du paramètre $\hat{\beta}$

Dans le régime sur-paramétré, $||\hat{\beta}_P||_2$ est décroissante à partir du moment où le rang de la matrice Z_P devient maximal i.e $rg(Z_P) = N$.

Théorème modèle linéaire quasi-isotropique

Conjecture: Limite du quotient de matrices aléatoires suivant une distribution orthonormée

$$\lim_{N\to+\infty,\ \frac{P}{N}\to\gamma,\ \frac{E}{N}\to\delta} \frac{1}{min(P,N)}\mathrm{tr}[\mathrm{Z_E}\mathrm{Z_E^T}(\mathrm{Z_P}\mathrm{Z_P^T})^{\dagger}]\Big] = \Big|\frac{1-\delta}{1-\gamma}\Big|$$

Théorème: **Expression de l'excès de risque moyen asymptotique, cas du rang maximal**

En prenant pour hypothèse la conjecture précédente. Et en supposant de plus que les $f_j(x_i)$ sont indépendants deux à deux. On pose $\mathcal{E}_{moyen}(\hat{y}) := \mathbb{E}_{\beta^\star,\epsilon_n}(\mathcal{E}(\hat{y},\beta^\star))$.

En supposant : $\mathbb{E}(\beta^\star\beta^{\star T}) = \frac{||\beta^\star||_2^2}{E} I_E$, $rg(Z_P) = min(P,N)$, et $\mathcal{F}$ features orthonormées par rapport à $\mathbf{X}$.

$$\overline{\mathcal{E}}(\gamma,\delta) := \lim_{N\to+\infty,\ \frac{P}{N}\to\gamma,\ \frac{E}{N}\to\delta} \mathcal{E}_{moyen}(\hat{y}_{\hat{\beta}}) \tag{3}$$

$$\boxed{\overline{\mathcal{E}}(\gamma,\delta) = \sigma_\phi^2\Big[1+\frac{min(\gamma,1)}{\delta}\Big(-2+\Big|\frac{1-\delta}{1-\gamma}\Big|\Big)\Big]\cdot||\beta^\star||_2^2 + \sigma_\epsilon^2\frac{min(\gamma,1)}{|1-\gamma|}} \tag{4}$$

Justification expérimentale de la conjecture

Ici $\gamma = \lim_{N \to +\infty} \frac{P}{N}$ et $\delta = \lim_{N \to +\infty} \frac{E}{N}$.

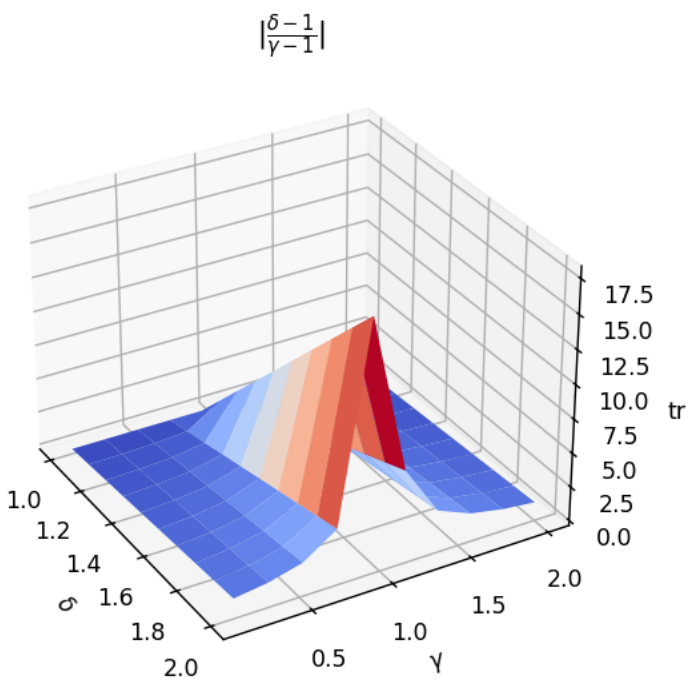


Figure: Conjecture

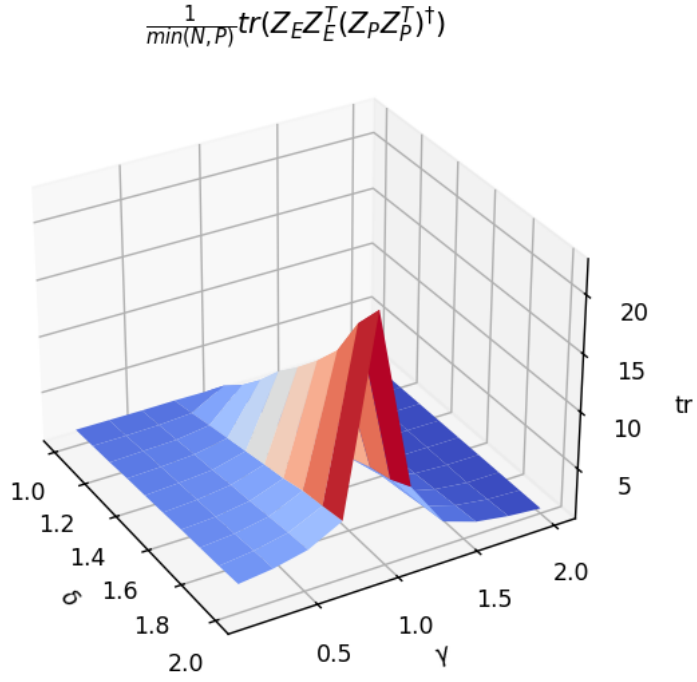


Figure: Expérience numérique pour loi normale standard et N = 500

Modèle linéaire quasi-isotropique

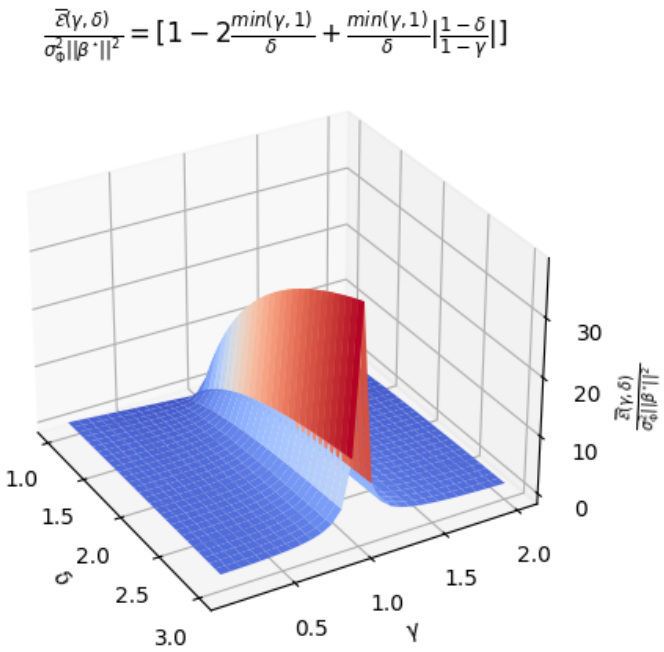


Figure: Modèle linéaire précèdent, avec $\epsilon = 0$

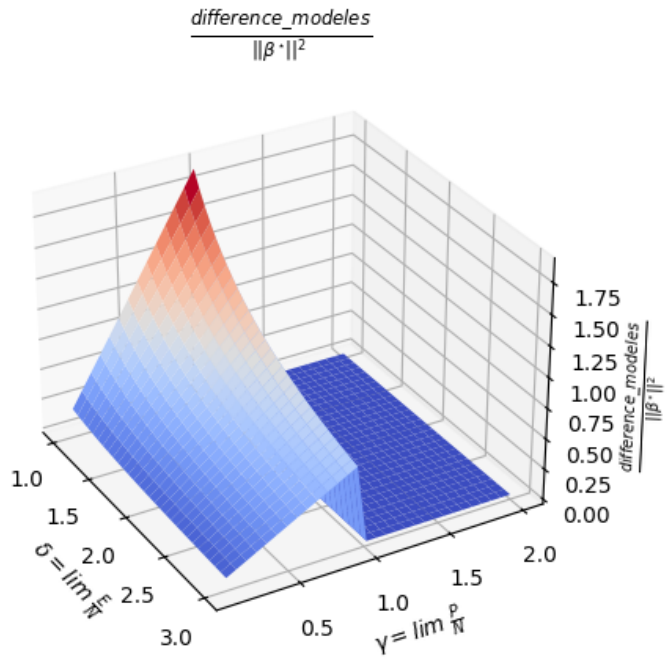


Figure: Différence entre les modèles de Belkin [1] et E. Haddad

Descente de gradient

Remarque:

Décomposition de l'erreur: Si l'on suppose ici que le bruit est nul i.e. $\epsilon_n = 0$.

$$\sigma_\Phi^{-2}\mathcal{E}(\hat{y}_{\hat{\beta}_t}, \beta^\star) = ||\hat{\beta}_t - \beta^\star||_2^2 = ||\hat{\beta}_t - \hat{\beta}||_2^2 + 2\mathrm{tr}[(\hat{\beta}_t - \hat{\beta})^T(\hat{\beta} - \beta^\star)] + ||\hat{\beta} - \beta^\star||_2^2$$

Théorème: Descente de gradient à pas constant par morceaux

On suppose que $\hat{\beta}_0 = 0$ et avec $t_1, \cdots, t_r$ changements de pas de descente de gradient. Alors, avec $t_{r+1} := t > t_r$:

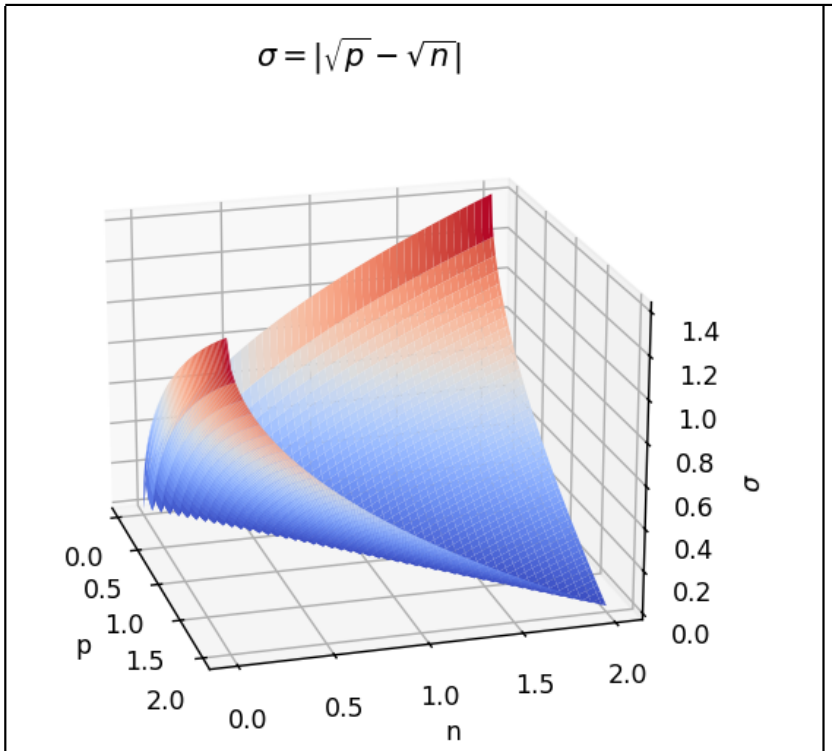
$$\hat{\beta} - \hat{\beta}_t = V \begin{bmatrix} \Sigma_R^{-1} \prod_{j=0}^{r} (I_R - \underline{\alpha_j'}\Sigma_R^2)^{t_{j+1}-t_j} & 0 \\ 0 & 0 \end{bmatrix} U^T Y \tag{5}$$

Corollaire: Approximation par descente de gradient à pas variable

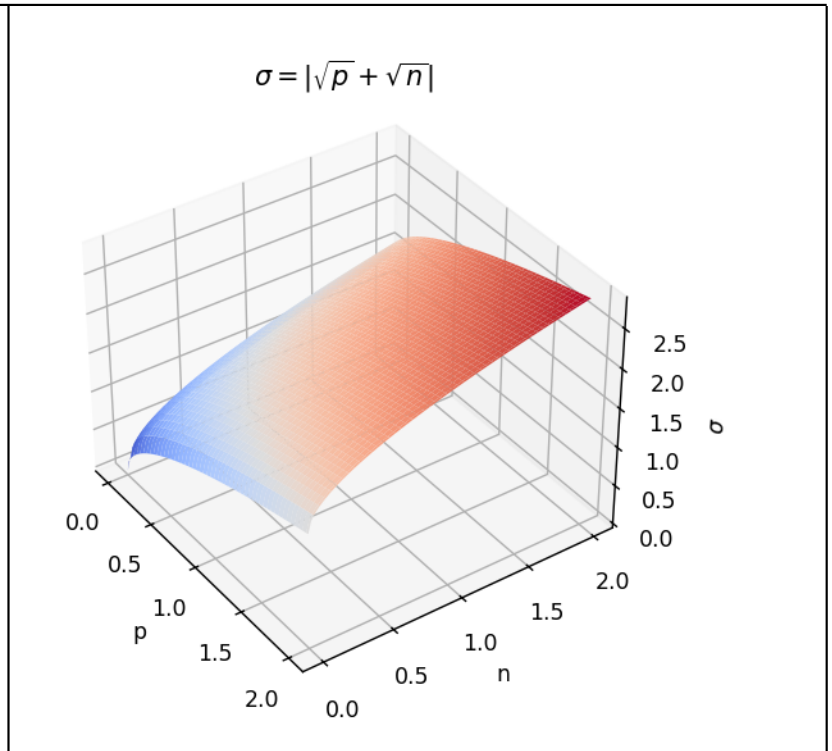
Dans le cas où $\hat{\beta}_0 = 0$ et $\max(\underline{\alpha_j}') < {\sigma_{\max}}^{-2}$, en notant $t_{r+1} := t > t_r$ on a:

$$\sigma_{max}^{-1} \prod_{j=0}^{r} (1 - \underline{\alpha_j}'\sigma_{max}^2)^{t_{j+1}-t_j} ||Y||_2 \leq ||\hat{\beta}_t - \hat{\beta}||_2 \leq \sigma_{min}^{-1} \prod_{j=0}^{r} (1 - \underline{\alpha_j}'\sigma_{min}^2)^{t_{j+1}-t_j} ||Y||_2$$

Étude asymptotique des valeurs singulières



Graphe indicatif de σ_{min}.
En effet : $\sigma_{min} \simeq \sigma_\Phi |\sqrt{P} - \sqrt{N}|$



Graphe indicatif de σ_{max}.
En effet : $\sigma_{max} \simeq \sigma_\Phi |\sqrt{P} + \sqrt{N}|$

"On the limit of the largest eigenvalue" [4] et "Marchenko–Pastur distribution" [2] pour $N \to +\infty$ et $P \backslash N \to \gamma$. Dans le cadre de notre base de features orthonormées,en notant $\mathbb{V}(\Phi_P(x)) = \sigma_\Phi^2 [1, \cdots, 1]^T$, on a alors $\sigma_{min} \simeq \sigma_\Phi |\sqrt{P} - \sqrt{N}|$ qui présente bien une courbe en U, et $\sigma_{max} \simeq \sigma_\Phi |\sqrt{P} + \sqrt{N}|$.

Régression pénalisée optimale

Définition: Régression pénalisée, et régression pénalisée optimale

- Risque empirique pénalisé $\hat{\mathcal{R}}_{\mathcal{D},\lambda}(\hat{y}_\beta) = ||Y - Z\beta||_2^2 + \lambda||\beta||_2^2$ où $\lambda > 0$.
- $\forall x \in \mathcal{X},\ \hat{y}_{\hat{\beta}_\lambda}(x) = \Phi_P^T(x)\hat{\beta}_\lambda$, pour $\hat{\beta}_\lambda = (Z^T Z + \lambda I_P)^{-1} Z^T Y$.
- On pose $\lambda_{D,N}^{opt} := \underset{\lambda \geq 0}{argmin}\ \overline{\mathcal{R}}(\hat{\beta}_\lambda)$ où $\overline{\mathcal{R}}(\hat{\beta}_\lambda) = \mathbb{E}_{\mathcal{D}}(\mathcal{R}(\hat{\beta}_\lambda))$, avec D et N fixés.

  Ce qui fixe $\hat{\beta}_{D,N}^{opt} := \hat{\beta}_{\lambda_{D,N}^{opt}}$.

Théorème: Expression du paramètre optimal et décroissance du risque , Nakkiran et al. [3]

Dans le cadre linéaire simple, avec $\mathbf{X} \sim \mathcal{N}(0, I_D)$ et $P \in \mathcal{O}_{P,D}(\mathbb{R})$:
On note $\tilde{\sigma}_\epsilon^2 = \sigma_\epsilon^2 + \frac{P-D}{P}||\beta^\star||_2^2$

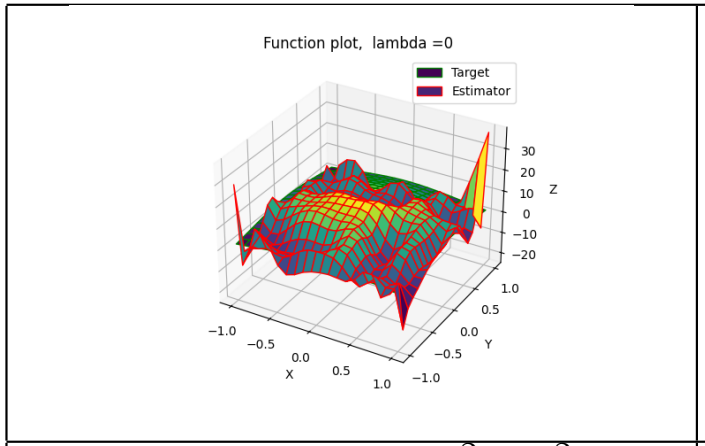
$$\text{Avec : } y^\star(x) = x^T\beta^\star \qquad \text{, on a : } \forall \lambda \geq \lambda_D^{opt} = \frac{D\sigma_\epsilon^2}{||\beta^\star||_2^2},\ \overline{\mathcal{R}}(\hat{\beta}_{N,\lambda}) \geq \overline{\mathcal{R}}(\hat{\beta}_{N+1,\lambda})$$

$$\text{Avec : } y^\star(x) = (Px)^T\beta^\star \text{ , on a : } \forall \lambda \geq \lambda_{D,P}^{opt} = \frac{P^2\tilde{\sigma}_\epsilon^2}{D||\beta^\star||_2^2},\ \overline{\mathcal{R}}(\hat{\beta}_{D,\lambda}) \geq \overline{\mathcal{R}}(\hat{\beta}_{D+1,\lambda})$$

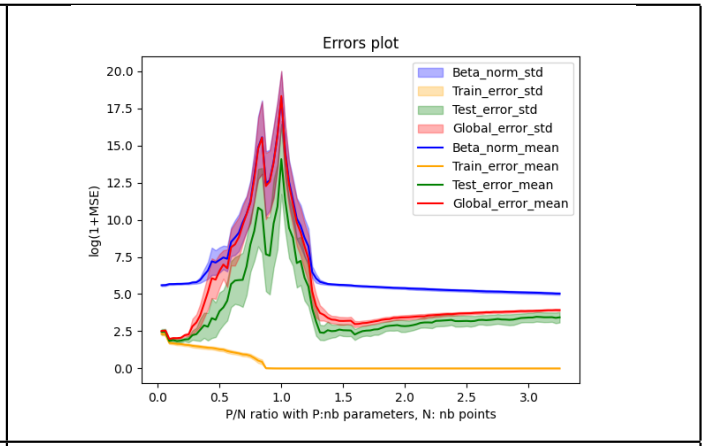
Appendice

Merci de votre écoute

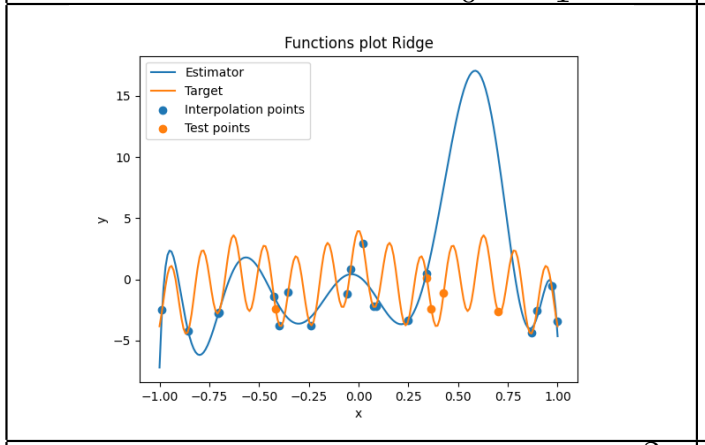
Régression polynomiale



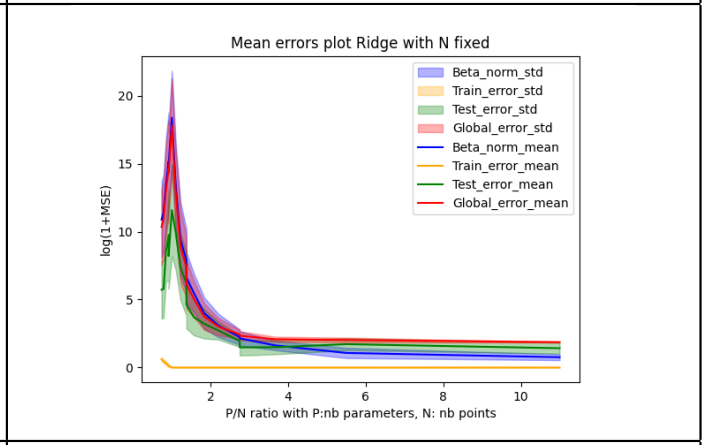
$y^\star(x) = 2exp(2 - (x_0^2 + x_1^2))$



Graphe du risque moyen, N est fixé



$y^\star(x) = 3cos(40x) + cos(10x) - x^2$

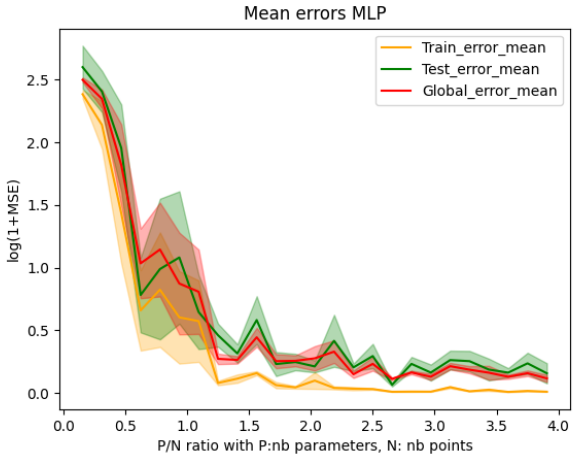


Graphe du risque moyen, P est fixé

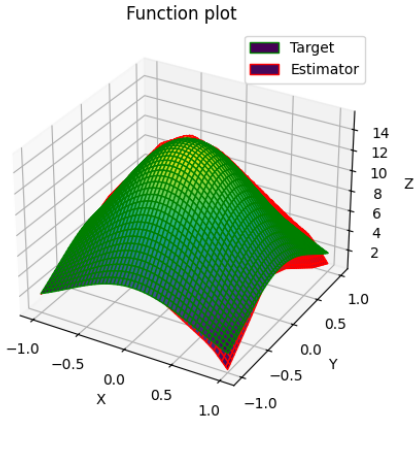
Figure: Résultats

On observe que le phénomène de double descente survient, que l'on commence par fixer N et faire varier P ensuite (premier exemple) ou au contraire, que l'on commence par fixer P et faire varier N ensuite (deuxième exemple).

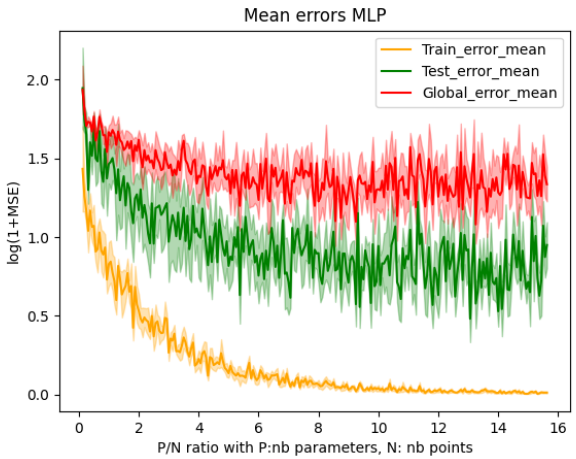
Perceptron Multicouche (MLP)



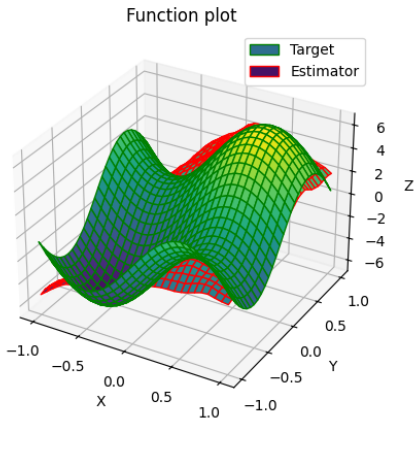
Risque moyen en fonction de $P \backslash N$, MLP



$y^{\star}(x) = 2exp(2 - (x_0^2 + x_1^2))$



Risque moyen en fonction de $P \backslash N$, MLP



$y^{\star}(x) = 20 \cdot [sin(3x_0) + sin(3x_1)] + \cdots$

Figure: Graphe de la MSE d'un algorithme MLP.

Bibliographie I

[1] Mikhail Belkin, Daniel Hsu, and Ji Xu. "Two models of double descent for weak features". In: *SIAM Journal on Mathematics of Data Science* 2.4 (2020), pp. 1167–1180.

[2] Ilja Kuzborskij et al. "On the role of optimization in double descent: A least squares study". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 29567–29577.

[3] Preetum Nakkiran et al. "Optimal regularization can mitigate double descent". In: *arXiv preprint arXiv:2003.01897* (2020).

[4] Yong-Qua Yin, Zhi-Dong Bai, and Pathak R Krishnaiah. "On the limit of the largest eigenvalue of the large dimensional sample covariance matrix". In: *Probability theory and related fields* 78 (1988), pp. 509–521.