Rapport - Double Descente

Emett Haddad, Encadrants: Nicolas Vayatis, Samuel Gruffaz

27/06/2024

Table des matières

1 Introduction

Le concept de double descente est un concept introduit en 2019 par Belkin [4] pour décrire un phénomène rencontré en machine learning: lorsque la complexité du modèle augmente l'erreur de généralisation tends à augmenter puis à diminuer de nouveau. Nous formaliserons cela au début du rapport.

Ce phénomène a été observé expérimentalement dans le contexte de diverses méthodes d'optimisations, et divers modèles. Cependant, il reste encore aujourd'hui difficile à prouver analytiquement. Il a été étudié de nombreuses fois par exemple par : [1, 3, 7, 9]. Par exemple, [1] a étudié le cas linéaire et sa résolution par pseudo-inverse, et [7] on étudiés le cas de linéaire par descente de gradient à pas constant. L'enjeu étant de le comprendre, d'en comprendre l'origine et de parvenir à trouver des conditions d'apparitions à ce phénomène.

En effet, il est utile de parvenir à le contrôler afin de pouvoir obtenir une meilleure erreur de généralisation en régime sur-paramétré. L'étude de ce phénomène pourrait contribuer à expliquer le succès et l'efficacité des réseaux de neurones actuels qui se placent pour la plupart dans des régimes largement sur-paramétrés.

2 Résumé

Nous allons donc premièrement présenter le contexte général du problème mathématique, c'est-à-dire l'approximation d'une fonction cible, qui modélise la dépendance entre l'entrée et la sortie d'un système, par un estimateur. Puis nous allons donner une définition possible de la double descente. Ensuite, nous allons présenter un état de l'art à travers quelques théorèmes récents.

Enfin, nous apporterons des éléments de réponse concernant le modèle linéaire par pseudo-inverse (équation 11), et puis par descente de gradient (équation 17). En effet, nous proposerons une expression asymptotique du risque moyen qui généralise en quelque sorte le résultat de Belkin [3], puis nous mettrons en évidence l'apport de la descente de gradient à la double descente à travers l'importance de la plus petite des valeurs singulière de la matrice des features qui généralise le résultat donné par [7]. Nous traiterons ici le cas de la descente de gradient à pas variable.

3 Présentation du problème

Définition: **Fonction cible**

On pose $\mathcal{X} = \mathbb{R}^D$ espace de départ et $\mathcal{Y} = \mathbb{R}$ espace d'arrivée.
On souhaite approximer une fonction $y^\star$ avec $\mathbf{X}$ et $\mathbf{Y}$ des variables aléatoires tel que $(\mathbf{X}, \mathbf{Y}) \hookrightarrow \mathcal{P}$.

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

Ici, $y^\star$ représente la fonction que l'on cherche à estimer, mais nous ne disposons pour cela que d'un échantillon de données, définit ci-dessous.

Définition: **Échantillons**

On se base sur un **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 $\mathbb{V}(\epsilon) = \sigma_\epsilon^2$.

Définition: **Estimateur**

- On veut trouver un **estimateur** $\hat{y} : \mathcal{X} \to \mathcal{Y}$ tel que $\hat{y}(\mathbf{X}) \approx \mathbf{Y}$ (au sens du risque ci-dessous).
- $\hat{y} \in \mathcal{H}$ un espace de fonctions.

La **complexité** de $\mathcal{H}$ est représentée ici par P le nombre de paramètres, bien que la meilleur notion actuelle de complexité est la complexité de Rademacher [10] qui est la capacité d'une classe de fonctions à s'adapter à du bruit.

Définition: **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: $\forall\, \hat{y} \in \mathcal{H}$

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

Ici, le vrai risque correspond à la vraie erreur que fait l'estimateur par rapport à la fonction de perte $\mathcal{L}(y, y') = (y - y')^2$. Cependant, en pratique, nous n'avons pas accès à cette erreur, d'où l'utilité du risque empirique, qui converge par loi des grands nombres vers ce vrai risque.

Pour étudier les différents modèles, il est intéressant d'étudier l'excès de risque, car il s'agit de l'erreur relativement à la fonction cible. Cela revient intuitivement à considérer l'erreur réalisée sans bruit, étant donné que si $\epsilon = 0$ alors $\mathcal{R}(y^\star) = 0$.

4 Régimes et Double Descente

Nous allons ici définir notre problème et donner une définition à la double descente dans le contexte de nos modèle.

Notions de régimes:

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

Définition: **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.

En réalité, la double descente est ce même phénomène mais relativement à la complexité de la classe de fonctions [4].

Illustration du compromis biais-variance

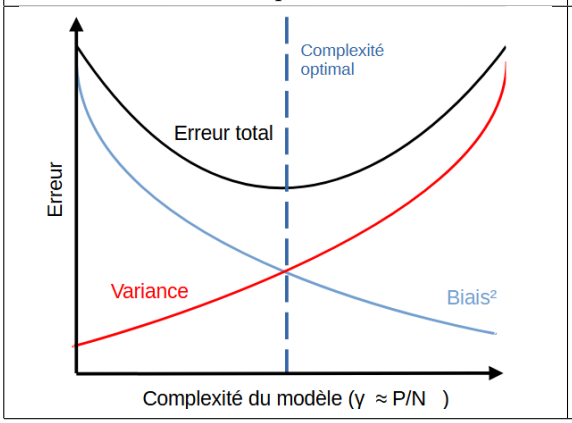
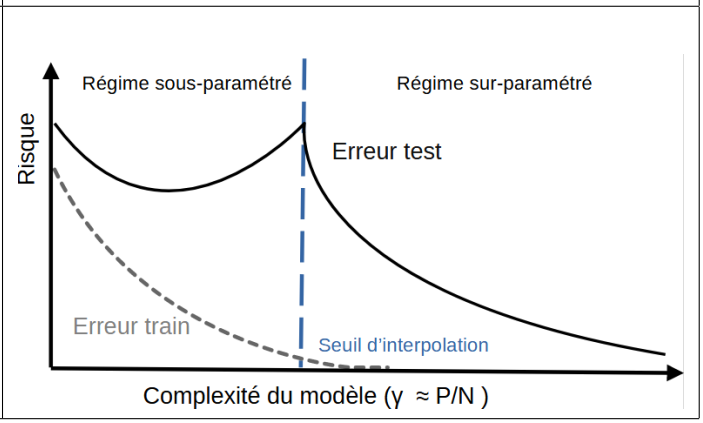


Illustration de la double descente



5 Modèles

Nous allons ici définir les différents modèles utilisés pour traiter notre problème.

5.1 Modèle par MLP: (Perceptron Multicouche)

Définition: **Modèle MLP**

- $\hat{y}_\beta(x) = W_L \circ \sigma \circ \cdots \sigma \circ W_1(x)$, $W_l(x_l) = A_l x_l + b_l$ où $A_l \in \mathcal{M}_{S_l,E_l}(\mathbb{R})$ et $b_l \in \mathbb{R}^{S_l}$ tel que $S_l = E_{l+1}$
- $\sigma(x) = max(0,x)$ fonction ReLU.
- On optimise $\hat{\mathcal{R}}_\mathcal{D}$ par descente de gradient (stochastique) selon le paramètre $\beta = \begin{bmatrix} A_l \\ b_l \end{bmatrix}_{l\in[\![1,L]\!]}$.

En pratique, on utilise ce type de modèle, mais leur analyse étant assez compliqué, on préfère étudier le modèle suivant plus simple.

5.2 Modèles linéaires

Définition: **Pseudo-inverse**

Soit $Z \in \mathcal{M}_{N,P}(\mathbb{R})$, on définit $Z^\dagger$ le pseudo-inverse de Z : en notant R = rg(Z), $U \in \mathcal{O}_N(\mathbb{R})$ et $V \in \mathcal{O}_P(\mathbb{R})$ Avec $Z = U \begin{bmatrix} \Sigma_R & 0_{\text{R,P-R}} \\ 0_{\text{N-R,R}} & 0_{\text{N-R,P-R}} \end{bmatrix} V^T$ la décomposition en valeur singulière de Z, alors $Z^\dagger = V \begin{bmatrix} \Sigma_R^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^T$.

Définition: **Modèle linéaire**

- $\mathcal{F} = \{f_i : \mathcal{X} \to \mathbb{R}, i \in \mathbb{N}\}$ 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 $\hat{\mathcal{R}}_\mathcal{D}(\hat{y}_\beta) = \frac{1}{N}||Y - Z\beta||_2^2$
- $\forall x \in \mathcal{X}$, $\hat{y}_{\hat{\beta}}(x) = \Phi_P^T(x)\hat{\beta}$, pour $\hat{\beta} = Z^\dagger Y$ qui minimise le risque empirique.
- Modèle simple $f_i(x) = e_i(x)$, X = Z et $P = D$.

5.3 Modèle linéaire pénalisé (Ridge Regression)

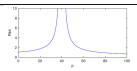
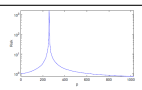
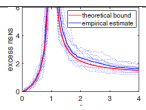
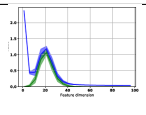
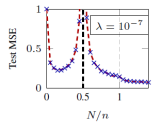
Définition: **Modèle linéaire pénalisé**

- 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 $\hat{\beta}_\lambda = (Z^TZ + \lambda I_P)^{-1}Z^TY$ qui minimise le risque empirique pénalisé.

Le terme $||\beta||_2^2$ dans le risque empirique pénalisé est le terme de pénalisation, qui est minimisé proportionnellement à λ. Ce terme permet d'une certaine façon de limiter la complexité de l'estimateur associé. Ainsi, on cherche notre solution parmi des solutions de moindre norme, i.e. en un certain sens de moindre complexité.

Pour $\lambda \to 0$, on retrouve le pseudo-inverse et pour $\lambda \to +\infty$ on obtient la solution nulle.

6 Etat de l'art

	Features/Modèle	Résultats	Méthodes	Graphe	Méthode d'optimisation
Belkin et al. [3]	$f_p(x) = e_{i_p}(x)$, coordonnées aléatoires	Expression du vrai risque, non-asymptotique	Inverse de Wishart		Pseudo-inverse
Belkin et al. [3]	$f_p(\omega) = \omega^{i_p}$ sur le cercle complexe Racine de l'unité pour l'échantillon $\mathcal{D}$.	Équivalent du vrai risque, asymptotique	Transformée de Stieltjes		Pseudo-inverse
Francis Bach [1]	$\hat{y}(x) = (S^T x)^T \hat{\beta}$ et $Z = XS$ où S matrice de vecteurs sub-Gaussien.	Équivalent du risque d'excès, asymptotique	Transformée de Stieltjes		Pseudo-inverse
Kuzborskij et al. [7]	$f_p(x) = e_p(x)$	Majoration du vrai risque non-asymptotique	Inégalité de concentration		Descente de gradient
Zhenyu et al. [8]	$f_p(x) = cos(v_p^T x)$ et $f_p(x) = sin(v_p^T x)$	Limite MSE, asymptotique	Étude de la résolvente		Pseudo-inverse

7 Focus techniques

Nous allons étudier quelques théorèmes existants particulièrement intéressants pour comprendre le phénomène de double descente dans le cas du modèle linéaire. Nous noterons E le nombre de degré de liberté de la fonction cible, i.e. $\beta^\star \in \mathbb{R}^E$ où $\beta^\star$ est le paramètre associé à la fonction cible.

7.1 Théorèmes de Belkin et al.

Théorème: Belkin et al.- base canonique [3]

Dans le cas où l'on sélectionne de manière uniforme P coordonnées parmi les $E := D$ coordonnées de la base canonique comme features et que x suit une loi normal standard. Si l'on suppose $y^\star(x) = x^T\beta^\star$, avec un bruit gaussien.

- Si $P < N - 1$:

$$\mathbb{E}_{\mathcal{D}}(\mathcal{R}(\hat{y}_{\hat{\beta}})) = [(1 - \frac{P}{E}) \cdot ||\beta^\star||_2^2 + \sigma_\epsilon^2](1 + \frac{P}{N - P - 1}) \tag{1}$$

- Si $P > N + 1$:

$$\mathbb{E}_{\mathcal{D}}(\mathcal{R}(\hat{y}_{\hat{\beta}})) = ||\beta^\star||_2^2 \cdot [1 - \frac{N}{E}(2 - \frac{E - N - 1}{P - N - 1})] + \sigma_\epsilon^2(1 + \frac{N}{P - N - 1}) \tag{2}$$

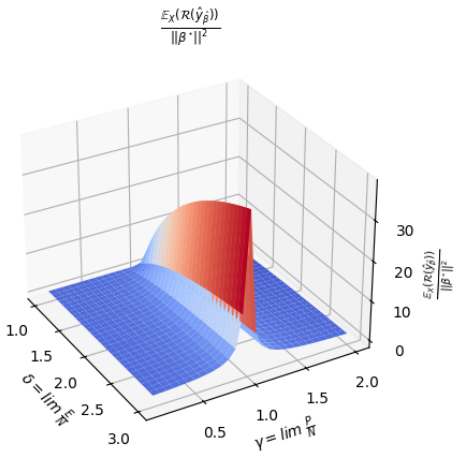
- Sinon:

$$\mathbb{E}_{\mathcal{D}}(\mathcal{R}(\hat{y}_{\hat{\beta}})) = +\infty \tag{3}$$

Éléments de preuve:
La preuve se déroule en plusieurs étapes: On note en indice (X_T, β_T) une extraction sur les lignes. On commence par poser $T \in [\![1, D]\!]^P$ variable aléatoire uniforme et $\hat{\beta}_T = X_T^{\dagger} Y$, $\hat{\beta}_{T^C} = 0$. On montre ces égalités dans l'ordre suivant: On pose $\eta_\epsilon = Y - X_T \beta_T^\star$
$\mathcal{R}(\hat{y}) = \sigma_\epsilon^2 + ||\beta_T^\star - \hat{\beta}_T||_2^2 + ||\beta_{T^C}^\star||_2^2 \quad et \quad ||\beta_T^\star - \hat{\beta}_T||_2^2 = ||\beta_T^\star||_2^2 - ||X_T^T (X_T X_T^T)^{\dagger} X_T \beta_T^\star||_2^2 + ||X_T^T (X_T X_T^T)^{\dagger} \eta_\epsilon||_2^2$
En utilisant Pythagore et les propriétés de projecteur de $X_T^T (X_T X_T^T)^{\dagger} X_T$.
$\mathbb{E}(||X_T^T (X_T X_T^T)^{\dagger} \eta_\epsilon||_2^2) = (||\beta_{T^C}||_2^2 + \sigma_\epsilon^2) \mathrm{tr}[\mathbb{E}((\mathrm{X_T X_T^T})^{\dagger})]$ On reconnaît l'inverse de Wishart.
Enfin on passe à l'espérance sur T. Et on a le résultat attendu. □

Pour $\sigma_\epsilon = 0$ et avec $N \to +\infty$, $E \backslash N \to \delta$ et $P \backslash N \to \gamma$ nous avons: (attention le modèle implique $\gamma \leq \delta$)

$$\underline{Si\ \gamma < 1 :}\ \mathbb{E}_{\mathcal{D}}(\mathcal{R}(\hat{y}_{\hat{\beta}})) \sim \frac{1}{1-\gamma}\Big[1 - \frac{\gamma}{\delta}\Big] \cdot ||\beta^\star||_2^2 \quad \underline{Si\ \gamma > 1 :}\ \mathbb{E}_{\mathcal{D}}(\mathcal{R}(\hat{y}_{\hat{\beta}})) \sim \Big[1 - \frac{2}{\delta} + \frac{1}{\delta}\frac{\delta - 1}{\gamma - 1}\Big] \cdot ||\beta^\star||_2^2$$



Ainsi, ce théorème exprime le risque moyenné sur l'échantillon.

Il montre bien le phénomène de double descente en exhibant une divergence au seuil d'interpolation i.e. $\gamma = 1$.
De plus δ permet d'estimer si oui ou non un minimum inférieur sera atteint en régime sur-paramétré.
La seule limite de ce modèle est qu'il est restreint à des lois normales, et nous en explorerons une généralisation dans la suite du rapport.

7.2 Théorème de Kuzborskij et al.

Théorème: Kuzborskij et al[7] en régime sur-paramétré

On pose $\hat{\beta}_t$ le vecteur obtenu au bout de t itérations de descente de gradient de pas α. On se place ici dans le cadre du modèle linéaire simple, en se plaçant dans l'hypothèse isotropique et sub-gaussien. On note σ_{min} la plus petite valeur singulière.

$$\mathcal{E}(\hat{y}_{\hat{\beta}_t}) \lesssim \Big[\Big(1 - \frac{2\alpha}{N}\sigma_{min}(X^T X)\Big)^{2t} + \frac{1}{\sqrt{N}}\Big] \cdot ||\beta^\star||_2^2 \tag{4}$$

Remarque:

Hypothèse isotropique: $\frac{1}{N}\mathbb{E}(X^T X) = I_D$ Hypothèse quasi-isotropique: $\frac{1}{N}\mathbb{E}(X^T X) = \sigma I_D$

Hypothèse sub-gaussien: X est sub-gaussien si et seulement si $\forall Y \in \mathbb{R}^N, X^T Y$ est sub-gaussien i.e. $\exists C, \forall t \geq 0, \mathbb{P}(|X^T Y| \geq t) \leq 2e^{-t^2 \backslash C^2}$.

L'intérêt de ce théorème est qu'il permet d'observer les éléments responsable d'une double descente dans le cas de la descente de gradient. En effet, on observe ici l'importance de la plus petite valeur singulière de Z. Et nous étudions plus en détail son importance dans la partie 7.2.2.

7.3 Théorème de Francis Bach

Théorème: Francis Bach[1]

On pose $df_i(\lambda) = tr(\Sigma^i(\Sigma + \lambda I_D)^{-i})$, $df_1(K_P) \sim P$, $df_1(K_N) \sim N$. On a ici $E := D$. On suppose ici $y^\star(x) = x^T S\beta^\star$. Ici, $S \in \mathcal{M}_{D,P}(\mathbb{R})$ est une projection aléatoire. Et $\Sigma := \frac{1}{N}\mathbb{E}(X^T X)$. On pose $\hat{\mathcal{R}}_\Sigma(\hat{y}_{\hat{\beta}}) := ||\hat{\beta} - \beta^\star||_\Sigma^2 = (\hat{\beta} - \beta^\star)^T\Sigma(\hat{\beta} - \beta^\star)$. Alors avec $\hat{\mathcal{R}}_\Sigma^{(var)}$ pour $\beta^\star = 0$ et $\hat{\mathcal{R}}_\Sigma^{(biais)}$ pour $\epsilon = 0$, et sous certaines hypothèses (vecteurs sub-Gaussiens iid) tel que $N \to +\infty$, $\frac{P}{N} \to \gamma$ et $\frac{E}{N} \to \delta$.

- Si $\gamma < 1$:

$$\mathbb{E}_\epsilon(\mathcal{R}_\Sigma^{(var)}(\hat{y}_{\hat{\beta}})) \sim \frac{\sigma_\epsilon^2\gamma}{1-\gamma}$$

$$\mathcal{R}_\Sigma^{(biais)}(\hat{\beta}) \sim \frac{K_P}{1-\gamma}||\beta^\star||^2_{\Sigma(\Sigma+K_P I)^{-1}}$$

- Si $\gamma > 1$:

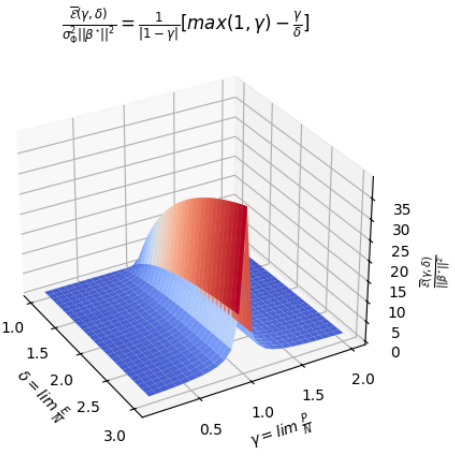
$$\mathbb{E}_\epsilon(\mathcal{R}_\Sigma^{(var)}(\hat{y}_{\hat{\beta}})) \sim \frac{\sigma_\epsilon^2}{\gamma-1} + \frac{\sigma_\epsilon^2 df_2(K_N)}{df_1(K_N) - df_2(K_N)}$$

$$\mathcal{R}_\Sigma^{(biais)}(\hat{y}_{\hat{\beta}}) \sim K_N^2\frac{df_1(K_N)}{df_1(K_N) - df_2(K_N)}||\beta^\star||^2_{\Sigma(\Sigma+K_N I)^{-2}} + \frac{K_N}{\gamma-1}||\beta^\star||^2_{\Sigma(\Sigma+K_N I)^{-1}}$$

Éléments de preuve:
On utilise la transformée de Stieltjes $\forall z \in \mathbb{C}\backslash\mathbb{R}_+$, $\hat{\Phi}(z) = \mathrm{tr}[(\mathrm{X}\mathrm{X}^\mathrm{T} + \mathrm{N}z\mathrm{I}_\mathrm{N})^{-1}]$. On utilise la limite $\lambda \to 0$ de la solution du problème pénalisé. Pour cela on pose $\hat{\Sigma} = \frac{1}{N}X^T X$ et on remarque que $\hat{\beta} = (\hat{\Sigma} + \lambda I_D)^{-1}\hat{\Sigma}S\beta_\star + \frac{1}{N}(\hat{\Sigma} + \lambda I_D)^{-1}X^T\epsilon$. Ensuite on passe de $\hat{\Sigma}$ à Σ par des propriétés de la transformée de Stieltjes. □

Application au modèle quasi-isotrope: Dans le cadre de $\Sigma = \sigma_\Sigma^2 I_D$ on trouve:

$$\overline{\mathcal{E}}(\gamma,\delta) := \mathcal{R}_{I_D}^{(biais)}(\hat{y}_{\hat{\beta}}) \sim \frac{\sigma_\Sigma^2}{|1-\gamma|}\left[max(1,\gamma) - \frac{\gamma}{\delta}\right] \cdot ||\beta^\star||_2^2 \tag{5}$$



Ainsi, ce théorème de Francis Bach permet de donner un équivalent asymptotique de l'erreur relative au bruit et à la fonction cible, et ce pour diverses distributions de $\mathbf{x}$.

De plus, il permet bien de justifier d'un phénomène de double descente 5. Enfin il décorrèle de manière singulière le nombre de paramètres de la dimension du problème à l'aide d'un projecteur.

8 Résultats théoriques

8.1 Régression linéaire

Nous avons choisi de nous intéresser à une base de features orthonormée de polynômes sur un cube I il s'agit intuitivement d'une base adaptée au problème. De plus elle rend plus aisée les expériences numériques. Enfin, elle fait apparaître expérimentalement un phénomène de double descente. Ce qui n'est pas toujours le cas avec la base canonique, voir l'expérience 4. D'où le cadre de notre étude ici.

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 $\forall\, z \in \mathbb{C}, G_P^x(z) = \sum_{p=1}^P f_p(x)z^p$.

Définition: Base de features orthonormées

On dit ici que $\mathcal{F} = \{f_0, f_1, \cdots, f_P\}$ est orthonormée par rapport à $\mathbf{X}$ si et seulement si:

- $f_0 = \tilde{\sigma_\Phi}$
- $\forall\, p,\ q \in [\![0, P]\!],\ \mathbb{E}_{\mathbf{X}}(f_p(\mathbf{X})f_q(\mathbf{X})) = \sigma_\Phi^2 \delta_{p,q}$

Remarque:

Ce modèle implique que l'on considère ici des fonctions de "moyenne" nulle i.e. $\forall\, p \geq 1, \mathbb{E}_{\mathbf{X}}(f_p(\mathbf{X})) = 0$.
En effet on a $\forall p \geq 1,\ \mathbb{E}_{\mathbf{X}}(f_p(\mathbf{X})) = \frac{1}{\sigma_\Phi}\mathbb{E}_{\mathbf{X}}(f_0(\mathbf{X})f_p(\mathbf{X})) = 0$.

Théorème: Expressions de l'estimateur

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

$$\forall\, x \in \mathcal{X},\ \hat{y}(x) = [f_1(x), \cdots, f_P(x)][(f_i, f_j)_X]^{\dagger} \begin{bmatrix} (f_1, y)_X \\ \vdots \\ (f_P, y)_X \end{bmatrix} \tag{6}$$

$$\forall\, x \in \mathcal{X},\ \hat{y}(x) = ([\{G_P^{x_i}, G_P^{x_j}\}]^{\dagger}[\{G_P^x, G_P^{x_j}\}])^T Y \tag{7}$$

Démonstration:
On a de manière général, $X^{\dagger} = (X^TX)^{\dagger}X^T = X^T(XX^T)^{\dagger}$. On applique donc cela à l'égalité $\hat{\beta} = Z^{\dagger}Y$, puis à $\hat{y}(x) = \Phi_P^T(x)\hat{\beta}$. Enfin, on remarque :

$$\{G_P^{x_i}, G_P^{x_j}\} = \frac{1}{2\pi}\int_0^{2\pi} G_P^{x_i}(e^{i\theta})\overline{G_P^{x_j}(e^{i\theta})} = \frac{1}{2\pi}\int_0^{2\pi} (\sum_{p=1}^P f_p(x_i)e^{ip\theta})(\sum_{q=1}^P f_q(x_j)e^{-iq\theta}) = \sum_{p=1}^P f_p(x_i)f_p(x_j)$$

□

Remarque:

Ces expressions permettent de visualiser la forme de l'estimateur. En effet la première expression s'approche de celle du projeté orthogonal dans $\mathcal{L}_2$ lorsque $N \to +\infty$.
Et la deuxième expression permet de constater que le risque empirique s'annule bien dans le cas du rang maximal, $rg(Z) = N$.

On observe souvent expérimentalement sur le modèle linéaire que les variations du risque suivent celles de $||\hat{\beta}||_2$, comme on peut le voir avec la figure 2. Il est donc pertinent d'étudier sa monotonie en régime sur-paramétré. On constate que $||\hat{\beta}||_2$ est décroissant à partir d'un certain rang avec le théorème suivant.

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 $rg(Z_P) = N$.

Démonstration:
On note ici Z_P la matrice Z avec P features. En supposant $rg(Z_P) = N$, nous avons ce résultat par l'étude du Lagrangien [9]. Le vecteur $\hat{\beta}_P$ est un minimum global pour la norme euclidienne dans l'espace affine $\hat{\beta}_P + Ker(Z_P) = \{\beta, Y = Z_P\beta\}$ car avec:

$$\forall \beta \in \mathbb{R}^P, \forall \gamma \in \mathbb{R}^N \mathcal{L}_P(\beta, \gamma) = ||\beta||_2^2 + \gamma^T(Y - Z_P\beta) \tag{8}$$

L'unique solution de $\nabla\mathcal{L}_P(\beta, \gamma) = 0$ i.e. $\nabla_\beta\mathcal{L}_P(\beta, \gamma) = 2\beta - Z_P^T\gamma = 0$ et $\nabla_\gamma\mathcal{L}_P(\beta, \gamma) = Y - Z_P\beta = 0$ est $\hat{\beta}_P$ car $rg(Z_P) = N$ et donc $Z_P Z_P^T$ est inversible. C'est donc l'unique solution de $Y = Z_P\beta$ de norme minimale. Et on a ${\hat{\beta}'}_{P+1} := \begin{bmatrix} \hat{\beta}_P \\ 0 \end{bmatrix}$ est aussi solution de $Y = Z_{P+1}\beta$ donc $||\hat{\beta}_{P+1}||_2 \leq ||\hat{\beta}'_{P+1}||_2 = ||\hat{\beta}_P||_2$.

□

Remarque:

Ce phénomène ($rg(Z_P) = N$) survient à partir d'un certain rang P si par exemple une seule des coordonnées dans une base est différente 2 à 2 sur l'échantillon, dans le cas de la régression polynomial à D variables. Car il suffit de pouvoir extraire une matrice de Vandermonde (quitte à changer de base). De plus, dans le cas de lois normales par exemple , si $N \neq P$ alors $rg(Z) = min(N, P)$ presque sûrement.

On pose E paramètres pour l'espace global, et on note les P premiers paramètres par P et P^C les E-P derniers paramètres parmi ces E paramètres. Ainsi: $\mathcal{F} := \{f_1, \cdots, f_P, \cdots, f_E\}$.

On note ici Z_P la matrice Z avec P features, i.e. $Z_P := \Phi_P^T(X)$, $Z_{P^C} := \Phi_{P^C}^T(X)$ et $Z_E := \left[Z_P, Z_{P^C}\right] = \Phi_E^T(X)$. Nous allons maintenant décomposer l'excès de risque en quatre termes que l'on exprimera en fonctions des matrices de features et de $\beta^\star$.

Théorème: Expression de l'excès de risque moyenné sur le bruit

On pose $\Phi_E(x) = \begin{bmatrix} \Phi_P(x) \\ \Phi_{P^C}(x) \end{bmatrix}$, $\beta^\star = \begin{bmatrix} \beta_P^\star \\ \beta_{P^C}^\star \end{bmatrix}$ et $\mathcal{E}(\hat{y}, \beta^\star) := \mathcal{R}(\hat{y}) - \mathcal{R}(y^\star)$, où l'on suppose $y^\star(x) = \Phi_E(x)^T\beta^\star$ avec $E \in \mathbb{N}$ features. On suppose ici $\mathbb{E}(\Phi_E(\mathbf{X})\Phi_E(\mathbf{X})^T) = \sigma_\Phi^2 I_E$.
On a alors:

$$\sigma_\Phi^{-2}\mathbb{E}_{\epsilon_n}(\mathcal{E}(\hat{y}_{\hat{\beta}}, \beta^\star)) = \mathbb{E}_{\epsilon_n}(||\hat{\beta} - \beta^\star||_2^2) = ||\beta^\star||_2^2 + ||Z_P^\dagger Z_{P^C}\beta_{P^C}^\star||_2^2 - \mathrm{tr}[(\mathrm{Z_P^\dagger Z_P})\beta_\mathrm{P}^\star\beta_\mathrm{P}^{\star\,\mathrm{T}}] + \sigma_\epsilon^2\mathrm{tr}[(\mathrm{Z_P Z_P^T})^\dagger] \tag{9}$$

Démonstration:
On s'inspire de la preuve de Belkin et al [3]. On note β_P pour parler des P premières composantes de β, de même pour les vecteurs colonnes des matrices. On note β_{P^C} pour les autres composantes.
$\mathcal{R}(\hat{y}) = \mathbb{E}_\mathcal{P}((\mathbf{Y} - \hat{y}(\mathbf{X}))^2) = \mathbb{E}_\mathcal{P}((\mathbf{Y} - y^\star(\mathbf{X}) + y^\star(\mathbf{X}) - \hat{y}(\mathbf{X}))^2) = \mathcal{R}(y^\star) + \mathbb{E}_\mathcal{P}((y^\star(\mathbf{X}) - \hat{y}(\mathbf{X}))^2)$
car $\mathbb{E}_\mathcal{P}(\mathbf{Y} - y^\star(\mathbf{X})|\mathbf{X}) = \mathbb{E}_\mathcal{P}(\epsilon|\mathbf{X}) = 0$ par hypothèse sur le bruit dans notre modèle.
Ainsi, on a $\mathcal{R}(\hat{y}_{\hat{\beta}}) - \mathcal{R}(y^\star) = \mathrm{tr}[\mathbb{E}_\mathbf{X}(\Phi_\mathrm{E}(\mathbf{X})\Phi_\mathrm{E}(\mathbf{X})^\mathrm{T})(\hat{\beta} - \beta^\star)(\hat{\beta} - \beta^\star)^\mathrm{T}] = \sigma_\Phi^2||\hat{\beta} - \beta^\star||_2^2$.

On pose $Y_\epsilon = \left[\epsilon_1, \cdots, \epsilon_N\right]^T$. On a $Y = Z_E\beta^\star + Y_\epsilon$ et $Z_E = \left[Z_P, Z_{P^C}\right]$. D'où $Y = Z_P\beta_P^\star + Z_{P^C}\beta_{P^C}^\star + Y_\epsilon$. On pose $\Pi_{Z_P} := Z_P^\dagger Z_P$ le projecteur orthogonale sur $Im(Z_P^\dagger)$. De plus on a $\hat{\beta}_P = Z_P^\dagger Y = Z_P^\dagger(Z_P\beta_P^\star + Z_{P^C}\beta_{P^C}^\star + Y_\epsilon)$ et $\hat{\beta}_{P^C} = 0$.
On a alors: $\mathbb{E}_{\epsilon_n}(||\hat{\beta} - \beta^\star||_2^2) - ||\beta_{P^C}^\star||_2^2 = \mathbb{E}_{\epsilon_n}(||\hat{\beta}_P - \beta_P^\star||_2^2) = \mathbb{E}_\epsilon(||Z_P^\dagger Y_\epsilon||_2^2) + ||(\Pi_{Z_P} - I_P)\beta_P^\star + Z_P^\dagger Z_{P^C}\beta_{P^C}^\star||_2^2 = \mathbb{E}_{\epsilon_n}(||Z_P^\dagger Y_\epsilon||_2^2) + ||(\Pi_{Z_P} - I_P)\beta_P^\star||_2^2 + ||Z_P^\dagger Z_{P^C}\beta_{P^C}^\star||_2^2$, par les propriétés de projecteur et car $\mathbb{E}_{\epsilon_n}(Y_\epsilon) = 0$.
De plus on a de même: $||(\Pi_{Z_P} - I_P)\beta_P^\star||_2^2 = \mathrm{tr}[\beta_\mathrm{P}^{\star\,\mathrm{T}}(\Pi_{\mathrm{Z_P}} - \mathrm{I_P})^\mathrm{T}(\Pi_{\mathrm{Z_P}} - \mathrm{I_P})\beta_\mathrm{P}^\star] = ||\beta_\mathrm{P}^\star||_2^2 - \mathrm{tr}[\Pi_{\mathrm{Z_P}}\beta_\mathrm{P}^\star\beta_\mathrm{P}^{\star\,\mathrm{T}}]$.
D'où $\sigma_\Phi^{-2}\mathbb{E}_{\epsilon_n}(\mathcal{E}(\hat{y}_{\hat{\beta}}, \beta^\star)) = \sigma_\epsilon^2\mathrm{tr}[(\mathrm{Z_P Z_P^T})^\dagger] + ||\beta^\star||_2^2 + ||\mathrm{Z_P^\dagger Z_{P^C}}\beta_{\mathrm{P^C}}^\star||_2^2 - \mathrm{tr}[\Pi_{\mathrm{Z_P}}\beta_\mathrm{P}^\star\beta_\mathrm{P}^{\star\,\mathrm{T}}]$.

□

L'excès de risque ce décompose donc en quatre termes, le premier $||\beta^\star||_2^2$ étant la norme de la fonction cible, et les deux suivant s'approchent de projections de la fonction cible. Le deuxième terme $||Z_P^\dagger Z_{P^C}\beta^\star_{P^C}||_2^2$ correspond au paramètre $\beta^\star_{P^C}$ de la fonction cible encore non observé par l'estimateur. Il est mis en facteur par un quotient de matrices de features. C'est de ce terme que provient la double descente. Le troisième terme $\mathrm{tr}[\Pi_{\mathrm{Z_P}}\beta^\star_{\mathrm{P}}\beta^{\star\,\mathrm{T}}_{\mathrm{P}}]$ correspondra a un rang de matrice. C'est un terme croissant. Le dernier terme $\sigma_\epsilon^2\mathrm{tr}[(\mathrm{Z_P Z_P^T})^\dagger]$ est lié au bruit, et contribue aussi a la double descente.

Remarque:

On peut par exemple prendre comme features une base de polynômes orthonormées sur un cube $I = \prod_{d=1}^D [a_d, b_d]$, que l'on peut obtenir par processus de Gram-Schmidt (algorithme 1). Et en prenant $\mathbf{X} \hookrightarrow \mathcal{U}(I)$ i.e. suivant une loi uniforme du cube I: alors cette base est orthonormée par rapport à $\mathbf{X}$ au sens de notre définition. Dans le cas d'une base orthonormée on a bien: $\mathbb{E}_{\mathbf{X}}(\Phi_E(\mathbf{X})\Phi_E(\mathbf{X})^T) = \sigma_\Phi^2 I_E$.

De plus on a $\forall p \geq 1$, $\mathbb{E}_{\mathbf{X}}(f_p(\mathbf{X})) = 0$. Donc $\mathbb{E}(\Phi_P(\mathbf{X})) = 0$. Ainsi, en posant $z_i = \mathbb{E}(\Phi_P(x_i)) \in \mathbb{R}^P$, les z_i sont indépendants par le lemme de coalition, de moyenne nulle et de même variance. Si l'on supposait de plus que les $f_j(x_i)$ étaient i.i.d., on se placerait alors dans le contexte de la distribution de Marchenko–Pastur [5]. Cette distribution fournirait alors des résultats théoriques pour la distribution de probabilité et la transformée de Stieltjes des matrices $\frac{1}{N} Z_P Z_P^T$ et $\frac{1}{P} Z_P^T Z_P$ asymptotiquement.

Corollaire: Expression de l'excès de risque moyen

On modélise ici $\beta^\star$ par une variable aléatoire tel que $\mathbb{E}(\beta^\star \beta^{\star T}) = \sigma^2_{\beta^\star} I_E$.
On pose $\mathcal{E}_{moyen}(\hat{y}) := \mathbb{E}_{\beta^\star, \epsilon_n}(\mathcal{E}(\hat{y}, \beta^\star))$.

$$(\sigma_\Phi \sigma_{\beta^\star})^{-2} \mathcal{E}_{moyen}(\hat{y}_{\hat{\beta}}) = E - 2rg(Z_P) + \mathrm{tr}[(\mathrm{Z_E Z_E^T} + (\sigma_\epsilon \backslash \sigma_{\beta^\star})^2 \mathrm{I_N})(\mathrm{Z_P Z_P^T})^\dagger] \tag{10}$$

Démonstration:
On a $||Z_P^\dagger Z_{P^C}\beta^\star_{P^C}||_2^2 = \mathrm{tr}[\mathrm{Z_{P^C}^T}(\mathrm{Z_P Z_P^T})^\dagger \mathrm{Z_{P^C}}\beta^\star_{\mathrm{P^C}}\beta^{\star\,\mathrm{T}}_{\mathrm{P^C}}]$ d'où:

$$\sigma_\Phi^{-2}\mathbb{E}_{\epsilon_n}(\mathcal{E}(\hat{y}_{\hat{\beta}}, \beta^\star)) = \sigma_\epsilon^2 \mathrm{tr}[(\mathrm{Z_P Z_P^T})^\dagger] + ||\beta^\star||_2^2 - \mathrm{tr}[(\mathrm{Z_P}^\dagger \mathrm{Z_P})\beta^\star_{\mathrm{P}}\beta^{\star\,\mathrm{T}}_{\mathrm{P}}] + \mathrm{tr}[\mathrm{Z_{P^C}^T}(\mathrm{Z_P Z_P^T})^\dagger \mathrm{Z_{P^C}}\beta^\star_{\mathrm{P^C}}\beta^{\star\,\mathrm{T}}_{\mathrm{P^C}}]$$

On a supposé que $\mathbb{E}(\beta^\star \beta^{\star T}) = \sigma^2_{\beta^\star} I_E$.
On a $\sigma_\Phi^{-2}\mathcal{E}_{moyen}(\hat{y}_{\hat{\beta}}) = \sigma_\epsilon^2 \mathrm{tr}[(\mathrm{Z_P Z_P^T})^\dagger] + \mathbb{E}(||\beta^\star||_2^2) - \mathrm{tr}[\mathrm{Z_P}^\dagger \mathrm{Z_P}\mathbb{E}(\beta^\star_{\mathrm{P}}\beta^{\star\,\mathrm{T}}_{\mathrm{P}})] + \mathrm{tr}[\mathrm{Z_{P^C}Z_{P^C}^T}(\mathrm{Z_P Z_P^T})^\dagger \mathbb{E}(\beta^\star_{\mathrm{P^C}}\beta^{\star\,\mathrm{T}}_{\mathrm{P^C}})]$.
D'où $(\sigma_\Phi \sigma_{\beta^\star})^{-2}\mathcal{E}_{moyen}(\hat{y}_{\hat{\beta}}) = E - \mathrm{tr}(\mathrm{Z_P}^\dagger \mathrm{Z_P}) + \mathrm{tr}[(\mathrm{Z_{P^C}Z_{P^C}^T})(\mathrm{Z_P Z_P^T})^\dagger] + (\sigma_\epsilon \backslash \sigma_{\beta^\star})^2 \mathrm{tr}[(\mathrm{Z_P Z_P^T})^\dagger]$.
Or $Z_E Z_E^T = Z_P Z_P^T + Z_{P^C} Z_{P^C}^T$, d'où:

$$\mathrm{tr}[(\mathrm{Z_{P^C}Z_{P^C}^T})(\mathrm{Z_P Z_P^T})^\dagger] = \mathrm{tr}[(\mathrm{Z_E Z_E^T})(\mathrm{Z_P Z_P^T})^\dagger] - \mathrm{tr}[\mathrm{Z_P Z_P^T}(\mathrm{Z_P Z_P^T})^\dagger] = \mathrm{tr}[(\mathrm{Z_E Z_E^T})(\mathrm{Z_P Z_P^T})^\dagger] - \mathrm{rg}(\mathrm{Z_P})$$

En effet la trace d'un projecteur est son rang, et $rg(XX^T) = rg(X)$. Donc de même $\mathrm{tr}(\Pi_{\mathrm{Z_P}}) = \mathrm{rg}(\mathrm{Z_P})$.
Ainsi $(\sigma_\Phi \sigma_{\beta^\star})^{-2}\mathcal{E}_{moyen}(\hat{y}_{\hat{\beta}}) = E - 2rg(Z_P) + \mathrm{tr}[(\mathrm{Z_E Z_E^T} + (\sigma_\epsilon \backslash \sigma_{\beta^\star})^2 \mathrm{I_N})(\mathrm{Z_P Z_P^T})^\dagger]$.

□

Considérer $\mathcal{E}_{moyen}$ est pertinent, car cela revient à moyenner l'excès de risque et le bruit de l'échantillon sur un ensemble de fonctions cibles. C'est à dire considérer $\{y_1^\star, y_2^\star, \cdots\}$, muni d'une distribution de probabilité tel que $\mathbb{E}(\beta^\star \beta^{\star T}) = \sigma^2_{\beta^\star} I_E$.

Conjecture: Limite du quotient de matrices aléatoires suivant une distribution de Marchenko-Pastur

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

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

Remarque:

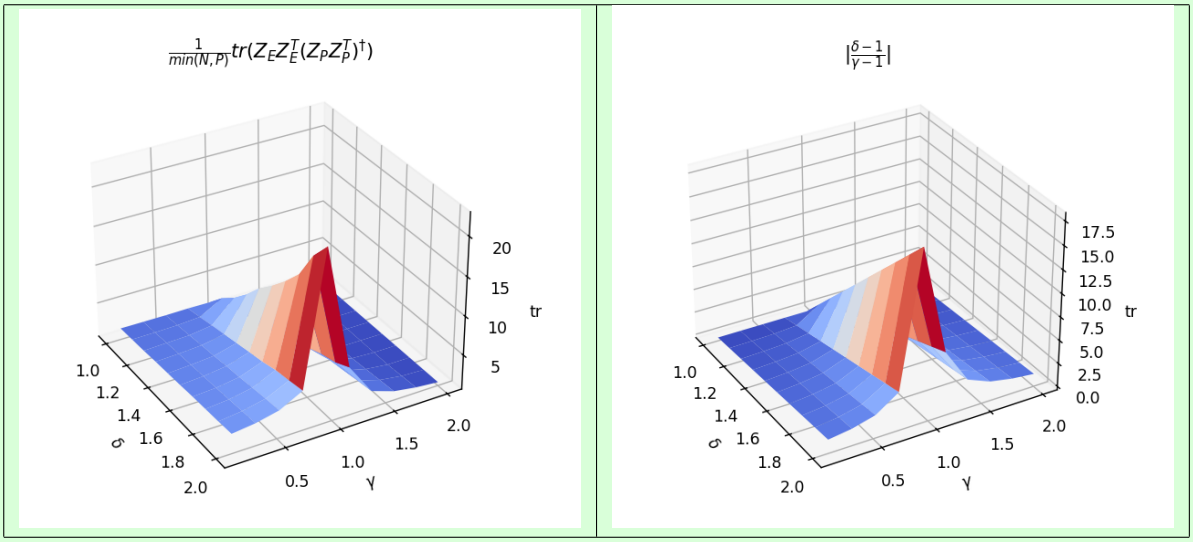
Théorie:

- Pour tenter de prouver la première conjecture, on pourrait tenter d'utiliser:
  Les lemmes de Francis Bach page 8 dans le papier [1].

  Et on peut aussi tenter d'utiliser la "distribution de Marchenko-Pastur" sur $\frac{1}{N}Z_P Z_P^T$ et sur $\frac{1}{P}Z_P^T Z_P$.
  Remarque, le quotient normalise la variance, on peut considérer $\sigma_\Phi = 1$.

- Concernant la deuxième conjecture, il faudrait d'étendre le résultat que l'on trouve dans le cas des matrices de Wishart et de la distribution de Marchenko-Pastur, a notre cas qui est un cas hybride des deux modèles.

Expérimentale: Justification expérimentale de la première conjecture: (avec une loi normale tel que $\sigma = 1$, et N = 500). Rappel: $\gamma = \lim_{N\to+\infty}\frac{P}{N}$ et $\delta = \lim_{N\to+\infty}\frac{E}{N}$.



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

En prenant pour hypothèse la conjecture précédente.
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 $\mathbf{X}$.

$$\overline{\mathcal{E}}(\gamma,\delta) := \lim_{N\to+\infty,\frac{P}{N}\to\gamma,\frac{E}{N}\to\delta} \mathcal{E}_{moyen}(\hat{y}_{\hat{\beta}}) = \sigma_\phi^2\Big[1+\frac{min(\gamma,1)}{\delta}\Big(-2+\Big|\frac{1-\delta}{1-\gamma}\Big|\Big)\Big]\cdot||\beta^\star||_2^2 + \sigma_\epsilon^2\frac{min(\gamma,1)}{|1-\gamma|} \quad (11)$$

Preuve:

On a supposé que $\frac{1}{min(P,N)}\text{tr}[\text{Z}_\text{E}\text{Z}_\text{E}^\text{T}(\text{Z}_\text{P}\text{Z}_\text{P}^\text{T})^\dagger]\Big] \sim |\frac{1-\delta}{1-\gamma}|$ et que $\text{tr}[(\text{Z}_\text{P}\text{Z}_\text{P}^\text{T})^\dagger]\Big] \sim \frac{\min(\gamma,1)}{|1-\gamma|}$.

- Régime sous-paramétré: (i.e. $\gamma < 1$)

$$\sigma_\Phi^{-2}\mathcal{E}_{moyen}(\hat{y}_{\hat{\beta}}) = \Big[1 - \frac{2P}{E} + \frac{P}{E}\frac{1}{P}\text{tr}[(\text{Z}_\text{E}\text{Z}_\text{E}^\text{T})(\text{Z}_\text{P}\text{Z}_\text{P}^\text{T})^\dagger]\Big] \cdot ||\beta^\star||_2^2 + \sigma_\epsilon^2\text{tr}[(\text{Z}_\text{P}^\text{T}\text{Z}_\text{P})^{-1}]$$

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

- Régime sur-paramétré: (i.e. $\gamma > 1$)

$$\sigma_\Phi^{-2}\mathcal{E}_{moyen}(\hat{y}_{\hat{\beta}}) = \Big[1 - \frac{2N}{E} + \frac{N}{E}\frac{1}{N}\text{tr}[(\text{Z}_\text{E}\text{Z}_\text{E}^\text{T})(\text{Z}_\text{P}\text{Z}_\text{P}^\text{T})^\dagger]\Big] \cdot ||\beta^\star||_2^2 + \sigma_\epsilon^2\text{tr}[(\text{Z}_\text{P}\text{Z}_\text{P}^\text{T})^{-1}]$$

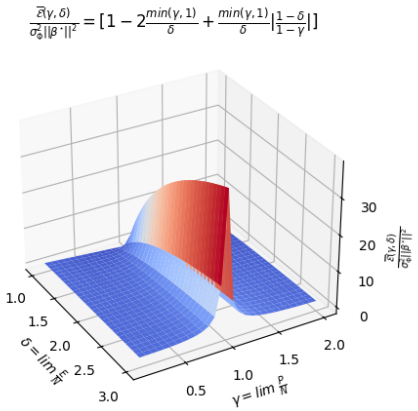
$$\sigma_\Phi^{-2}\overline{\mathcal{E}}(\gamma,\delta) = \Big[1 - 2\frac{1}{\delta} + \frac{1}{\delta}\Big|\frac{1-\delta}{1-\gamma}\Big|\Big] \cdot ||\beta^\star||_2^2 + \frac{\sigma_\epsilon^2\sigma_\phi^{-2}}{\gamma-1}$$

□

Remarque:

On se place dans le cadre $||\beta^\star||_2^2 = tr(\beta^\star\beta^{\star T}) = \sigma_{\beta^\star}^2 E =$ constante. Ce choix vient du fait que l'on imagine que l'erreur aléatoire sur les termes du développement en série de nos fonctions cibles diminue lorsque l'on prend plus de termes. Cela revient à écrire sur l'ensemble des fonctions à approximer que que $\mathbb{E}(\beta^\star) = 0$ et $\mathbb{V}(\beta^\star) = \frac{||\beta^\star||_2^2}{E} I_E$.

Sous hypothèse de cette conjecture et dans le cas $\epsilon = 0$, l'excès de risque moyen asymptotique présente cette forme : (attention, notre modèle implique $P \leq E$ d'où $\gamma \leq \delta$).



Pour $E = P$ on trouve alors $\overline{\mathcal{E}}(\gamma,\delta) = \sigma_\Phi^2 max(0,(1-\frac{1}{\gamma})) \cdot ||\beta^\star||_2^2$ et on n'a pas de double descente. Mais pour $\delta > 1$ on en a une.

On est dans un cas très similaire à celui traité par Belkin [3], les seules différences étant que nous ne sommes pas dans le cas de lois normales, c'est-à-dire que le bruit et les entrées ne suivent pas des lois normales, que nous moyennons sur les fonctions cibles, et que nous étudions une limite asymptotique. On remarque que nous trouvons la même expression dans le régime sur-paramétré que dans l'équation 2 de Belkin, lorsque $\sigma_\epsilon = 0$.

8.2 Minimisation par descente de gradient

On considère $\mathcal{D}$ fixé. On a avec $\hat{\beta}_t$ paramètre issu d'une descente de gradient à t étapes:

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

Nous avons donc trois termes à étudier, le dernier terme $||\hat{\beta} - \beta^\star||_2^2$ a déjà été étudié précédemment, car il correspond à la meilleure approximation faite par pseudo-inverse. Dans cette partie, nous allons nous intéresser au premier terme : $||\hat{\beta}_t - \hat{\beta}||_2^2$, lié à la descente de gradient. Nous évoquerons aussi une piste pour traiter le terme du milieu $\text{tr}[(\hat{\beta}_\text{t} - \hat{\beta})^\text{T}(\hat{\beta} - \beta^\star)]$, que nous considérerons nulle dans un premier temps.

Définition: **Optimisation par descente de gradient**

- Descente de gradient simple:

$$\hat{\beta}_{t+1} = \hat{\beta}_t - \alpha_t \nabla\hat{\mathcal{R}}_\mathcal{D}(\hat{y}_{\hat{\beta}_t}) \; et \; \hat{\beta}_0 \in \mathbb{R}^P$$

- Descente de gradient stochastique:

$$\hat{\beta}_{t+1} = \hat{\beta}_t - \alpha_t J_t \nabla\hat{\mathcal{R}}_{\mathcal{D}_t}(\hat{y}_{\hat{\beta}_t}) \; et \; \hat{\beta}_0 \in \mathbb{R}^P$$

  où $J_t = Diag(\delta_{1,t} \cdots \delta_{P,t})$ et $\mathcal{D}_t \in \mathcal{D}^B$. Ici $\delta_{1,t}$ sont des variables aléatoires à valeur dans $\{0, 1\}$ et $\mathcal{D}_t$ un sous-ensemble aléatoire de $\mathcal{D}$. Nous les munissons de lois uniformes.
  J_t représente le fait qu'on ne calcule qu'une partie du gradient, et $\mathcal{D}_t$ représente les batchs de taille B, on ne calcule le gradient que sur un sous-ensemble de l'échantillon.

8.2.1 Descente de gradient simple

La descente de gradient a pas constant est donnée par:

$$\boxed{\hat{\beta}_{t+1} = \hat{\beta}_t - \alpha\nabla\hat{\mathcal{R}}_\mathcal{D}(\hat{y}_{\hat{\beta}_t})}$$

On a $\hat{\mathcal{R}}_\mathcal{D}(\hat{y}) = \frac{1}{N}||Y - Z\hat{\beta}||_2^2$ et $\nabla\hat{\mathcal{R}}_\mathcal{D}(\hat{y}) = \frac{2}{N}Z^T[Z\hat{\beta} - Y]$ ainsi $\hat{\beta}_{t+1} = [I_p - \frac{2\alpha}{N}Z^TZ]\hat{\beta}_t + \frac{2\alpha}{N}Z^TY$.
On sait qu'avec des suites du type $x_{n+1} = ax_n + b$ on a $x_n = [\sum_{k=0}^{n-1} a^k]b + a^n x_0$.
D'où en posant $\alpha' = \frac{2\alpha}{N}$, on a:

$$\hat{\beta}_t = \Big[\sum_{k=0}^{t-1}(I_p - \alpha' Z^TZ)^k\Big]\alpha' Z^TY + [I_p - \alpha' Z^TZ]^t\hat{\beta}_0$$

Théorème: Approximation par descente de gradient simple

Avec $\hat{\beta}_t$ le résultat de t descente de gradient de pas α tel que $\alpha' < \sigma_{max}^{-2}$ et $\hat{\beta}_0 = 0$. Avec $\sigma_{min} > 0$ la plus petite valeur singulière non-nulle de $Z = U\Sigma V^T$, et σ_{max} sa plus grande valeur singulière, on a:

$${\sigma_{max}}^{-1}(1 - \alpha'\sigma_{max}^2)^t||Y||_2 \leq ||\hat{\beta}_t - \hat{\beta}||_2 \leq {\sigma_{min}}^{-1}(1 - \alpha'\sigma_{min}^2)^t||Y||_2 \tag{12}$$

On a :

$$\hat{\beta}_t - \hat{\beta} = V\begin{bmatrix}(I_R - \alpha'\Sigma_R^2)^t & 0\\ 0 & I_{P-R}\end{bmatrix}V^T[\hat{\beta}_0 - \hat{\beta}] \tag{13}$$

Si $\hat{\beta}_0 = 0$ on a:

$$\hat{\beta}_t - \hat{\beta} = -V\begin{bmatrix}\Sigma_R^{-1}(I_R - \alpha'\Sigma_R^2)^t & 0\\ 0 & 0\end{bmatrix}U^TY \tag{14}$$

Démonstration:

On s'attend classiquement a une convergence géométrique de la descente de gradient.
Avec $R := rg(Z)$ et $Z = U\Sigma V^T$ décomposition SVD tel que $\Sigma = \begin{bmatrix}\Sigma_R & 0\\ 0 & 0\end{bmatrix}$. On a $Z^TZ = V\begin{bmatrix}\Sigma_R^2 & 0\\ 0 & 0\end{bmatrix}V^T$.
On note $\alpha' = \frac{2\alpha}{N}$. D'où $\hat{\beta}_t = V\begin{bmatrix}(\alpha'\Sigma_R^2)^{-1}[I_R - (I_R - \alpha'\Sigma_R^2)^t] & 0\\ 0 & tI_{P-R}\end{bmatrix}\alpha'\begin{bmatrix}\Sigma_R & 0\\ 0 & 0\end{bmatrix}U^TY + V\begin{bmatrix}(I_R - \alpha'\Sigma_R^2)^t & 0\\ 0 & I_{P-R}\end{bmatrix}V^T\hat{\beta}_0$.
$\hat{\beta}_t = V\begin{bmatrix}\Sigma_R^{-1} - \Sigma_R^{-2}(I_R - \alpha'\Sigma_R^2)^t\Sigma_R & 0\\ 0 & 0\end{bmatrix}U^TY + V\begin{bmatrix}(I_R - \alpha'\Sigma_R^2)^t & 0\\ 0 & I_{P-R}\end{bmatrix}V^T\hat{\beta}_0$
Si $\hat{\beta}_0 = 0$ on a l'équation 14: $\hat{\beta}_t - \hat{\beta} = -V\begin{bmatrix}\Sigma_R^{-1}(I_R - \alpha'\Sigma_R^2)^t & 0\\ 0 & 0\end{bmatrix}U^TY$
De plus on a $V\begin{bmatrix}(I_R - \alpha'\Sigma_R^2)^t & 0\\ 0 & I_{P-R}\end{bmatrix}V^T\hat{\beta} = V\begin{bmatrix}(I_R - \alpha'\Sigma_R^2)^t & 0\\ 0 & I_{P-R}\end{bmatrix}V^TV\begin{bmatrix}\Sigma_R^{-1} & 0\\ 0 & 0\end{bmatrix}U^TY$ ce qui donne l'équation 13.
On utilise alors le lemme : $\sigma(A)_{min}||X||_2 \leq ||AX||_2 \leq \sigma(A)_{max}||X||_2$.
On peut étudier la fonction $f(\sigma) = \sigma^{-1}(1 - \alpha'\sigma^2)^t$, et on a alors deux annulations de la dérivée en $\sigma = \pm\alpha'^{-\frac{1}{2}}$. On se place donc dans la zone de décroissance de f et les valeurs propres de $\begin{bmatrix}\Sigma_R^{-1}(I_R - \alpha'\Sigma_R^2)^t & 0\\ 0 & 0\end{bmatrix}$, i.e. les $f(\sigma_r)$ sont positives, lorsque $\alpha' < \sigma_{max}^{-2}$. On a finalement le résultat attendu, l'équation 12.

□

Remarque:

On retrouve ici un résultat proche de celui de la descente de gradient classique, ici $\hat{\mathcal{R}}$ étant $\mu = \frac{2}{N}$ fortement convexe. Et on remarque que notre résultat est très proche de celui de l'inéquation 4.

Corollaire: Résultats sur la descente de gradient simple, cas du rang maximal

On suppose que $\hat{\beta}_0 = 0$ et que α le pas de la descente est constant .

- Si $P \leq N$ et rg(Z) = P:
$$\hat{\beta} - \hat{\beta}_t = (Z^TZ)^{-1}(I_P - \alpha'Z^TZ)^tZ^TY \tag{15}$$
- Si $P \geq N$ et rg(Z) = N:
$$\hat{\beta} - \hat{\beta}_t = Z^T(I_N - \alpha'ZZ^T)^t(ZZ^T)^{-1}Y \tag{16}$$

Démonstration:
On s'attend à avoir une forme simple dans le cas du rang maximal.

- Dans le régime sous-paramétré:
  On a $\hat{\beta}_t = (\sum_{k=0}^{t-1}[I_p - \alpha'Z^TZ]^k)\alpha'Z^TY = (Z^TZ)^{-1}[I_P - (I_p - \alpha'Z^TZ)^t]Z^TY = Z^\dagger Y - (Z^TZ)^{-1}(I_P - \alpha'Z^TZ)^tZ^TY$.
- Dans le régime sur-paramétré:
  On a ici $ZZ^T = U\Sigma_N^2U^T$ et $Z^\dagger Y - \hat{\beta}_t = V\begin{bmatrix}\Sigma_N^{-1}(I_N - \alpha'\Sigma_N^2)^t\\ 0\end{bmatrix}U^TY = V\begin{bmatrix}\Sigma_N\\ 0\end{bmatrix}(U^TU)(I_N - \alpha'\Sigma_N^2)^t(U^TU)\Sigma_N^{-2}U^TY = (V\Sigma^TU^T)(I_N - \alpha'U\Sigma_N^2U^T)^t(U\Sigma_N^2U^T)^{-1}Y$.
  D'où: $\hat{\beta}_t = Z^\dagger Y - Z^T(I_N - \alpha'ZZ^T)^t(ZZ^T)^{-1}Y$

□

Remarque:

Dans le régime sur-paramétré, on a pour la descente de gradient à pas constant: $\hat{\beta}_P = Z_P^{\dagger} Z_E \beta^\star$, $\mathbb{E}_{\beta^\star}(\mathrm{tr}[(\hat{\beta}_\mathrm{t} - \hat{\beta})^\mathrm{T}(\hat{\beta} - \beta^\star)]) = \sigma^2_{\beta^\star}\mathrm{tr}[(\mathrm{Z_E Z_E^T})(\mathrm{Z_P Z_P^T})^{-1}(\mathrm{I_N} - \alpha \mathrm{Z_P Z_P^T})^\mathrm{t}]$ qui semble présenter un comportement de double descente en s'inspirant du $\mathrm{tr}[(\mathrm{Z_E Z_E^T})(\mathrm{Z_P Z_P^T})^{-1}]$ de la section précédente.

8.2.2 Descente de gradient à pas variable

Dans la pratique, le pas de la descente de gradient n'est pas constant. C'est pourquoi nous allons généraliser les résultats précédents à la descente de gradient à pas variable.

La descente de gradient a pas variable est donnée par:

$$\boxed{\hat{\beta}_{t+1} = \hat{\beta}_t - \alpha_t \nabla \hat{\mathcal{R}}_{\mathcal{D}}(\hat{y}_{\hat{\beta}_t})}$$

Et donc : $\hat{\beta}_{t+1} = [I_P - \frac{2\alpha_t}{N} Z^T Z]\hat{\beta}_t + \frac{2\alpha_t}{N} Z^T Y$, on est dans le cas $u_{n+1} = a_n u_n + b_n$. Dans ce cas on a $u_n = [\prod_{k=0}^{n-1} a_k]u_0 + \sum_{k=0}^{n-1}[\prod_{i=k+1}^{n-1} a_i]b_k$.
D'où:

$$\hat{\beta}_t = \Big[\prod_{k=0}^{t-1}(I_P - \frac{2\alpha_k}{N} Z^T Z)\Big]\hat{\beta}_0 + \sum_{k=0}^{t-1}\Big[\prod_{i=k+1}^{t-1}(I_P - \frac{2\alpha_i}{N} Z^T Z)\Big]\frac{2\alpha_k}{N} Z^T Y$$

On suppose $\hat{\beta}_0 = 0$ et on pose $Z = U\Sigma V^T$ où $R = rg(Z)$, $\Sigma = \begin{bmatrix} \Sigma_R & 0 \\ 0 & 0 \end{bmatrix}$.
On va tenter de simplifier le problème en considérant un pas α_t constant par morceaux et cela permettra d'obtenir une forme explicite de l'erreur.

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 $t_{r+1} := t > t_r$:

$$\hat{\beta} - \hat{\beta}_t = V \begin{bmatrix} \Sigma_R^{-1}\prod_{j=0}^{r}(I_R - \underline{\alpha_j'}\Sigma_R^2)^{t_{j+1}-t_j} & 0 \\ 0 & 0 \end{bmatrix} U^T Y \tag{17}$$

Démonstration:
On suppose maintenant $t_1, \cdots, t_r$ changements de pas de descente de gradient. On a alors avec $t_0 = 0$ et $t_{r+1} = t-1$: $\forall j \leq r,\ \forall t \in [\![t_j, t_{j+1}[\![,\ \alpha_t' = \underline{\alpha_i'}$.
On démontre le résultat par récurrence sur le nombre de changements de pas, à l'aide du théorème sur la descente de gradient simple, en utilisant les équations 13 (hérédité) et 14 (initialisation). □

Corollaire: Approximation par descente de gradient à pas variable

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

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

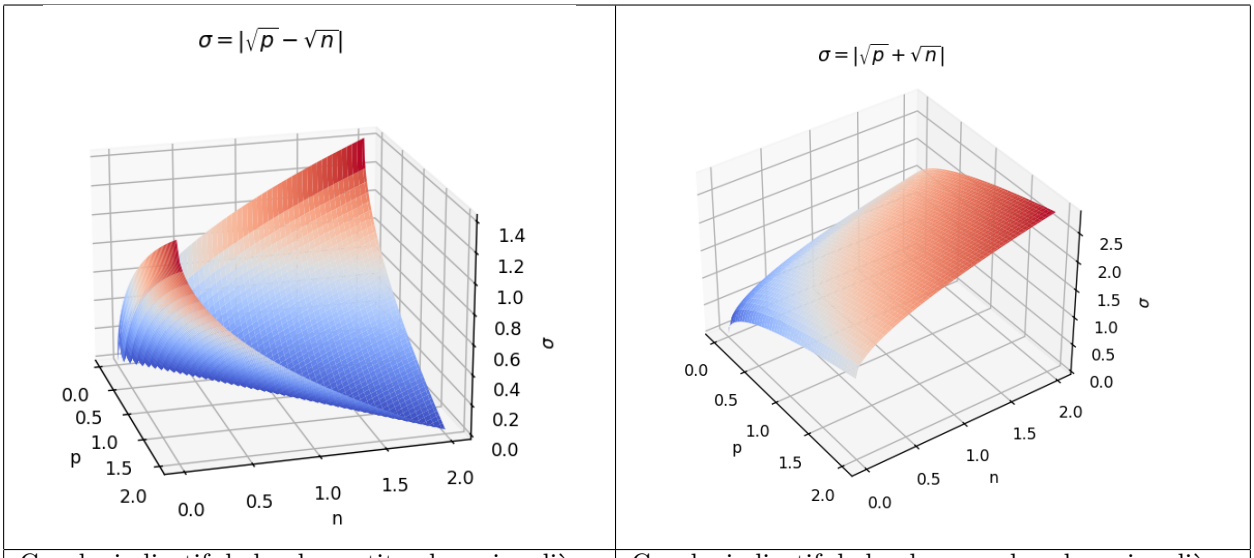
Démonstration:
On prouve cette expression à l'aide des variations de la fonction $f(\sigma) = \sigma^{-1}\Pi_{j=0}^{r}(1-\gamma_j\sigma^2)^{\Delta t_j} = \sigma^{-1}g(\sigma)$, tel que $f(\sigma) \underset{\sigma\to 0}{\sim} 1\backslash\sigma$, et tel que $f'(\sigma) = \frac{-g(\sigma)}{\sigma^2}[2\sum_{j=0}^{r}\frac{\Delta t_j\gamma_j\sigma^2}{1-\gamma_j\sigma^2} + 1] \underset{\sigma\to 0}{\sim} -1\backslash\sigma^2$ et dont les points critiques sont les solutions de $\sum_{j=0}^{r}\gamma_j\sigma^2\Delta t_j(1-\gamma_j\sigma^2)^{-1} = -1\backslash 2$ et les points $\sigma = \pm\gamma_j^{-1\backslash 2}$. Sous la condition $\max(\gamma_\mathrm{j}) \leq {\sigma_\mathrm{max}}^{-2}$, la première

équation n'a pas de solution et l'on se place dans la zone de décroissance de f, on peut donc se ramener à la plus petite et à la plus grande des valeurs singulière de Z (matrice des features) en utilisant l'équation précédente 17 et le lemme: $\sigma(A)_{min}||X||_2 \leq ||AX||_2 \leq \sigma(A)_{max}||X||_2$ □

Ainsi, on a une erreur décroissante par rapport à t et σ. Donc σ_{min} ayant une courbe en U (voir section suivante) on a bien une courbe en U inversée pour l'erreur en fonction du quotient $P\backslash N$. Ainsi l'erreur apportée par la descente de gradient présente bien un apport au phénomène de double descente.
L'inégalité obtenue est optimale au sens que la seule inégalité utilisée dans la preuve est $\sigma(A)_{min}||X||_2 \leq ||AX||_2 \leq \sigma(A)_{max}||X||_2$.
Ici, la condition sur le pas permet bien de faire tendre le pas vers 0.

8.2.3 Analyse de la plus petite et plus grande valeur singulière

Il est donc important d'étudier les valeurs singulières de la matrice Z pour comprendre l'impacte de la descente de gradient sur la double descente. Voici deux graphes qui illustrent les variations de ces valeurs singulières en fonction de P et de N.



Graphe indicatif de la plus petit valeur singulière. En effet : $\sigma_{min} \simeq \sigma_\Phi \lvert\sqrt{P} - \sqrt{N}\rvert$	Graphe indicatif de la plus grande valeur singulière. En effet : $\sigma_{max} \simeq \sigma_\Phi \lvert\sqrt{P} + \sqrt{N}\rvert$

"On the limit of the largest eigenvalue" [11] et "Marchenko–Pastur distribution" [7] 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 $\sigma_{max} \simeq \sigma_\Phi|\sqrt{P} + \sqrt{N}|$.

8.2.4 Descente de gradient stochastique

La descente de gradient stochastique sur les coordonnées du gradient est donnée par:

$$\boxed{\hat{\beta}_{t+1} = \hat{\beta}_t - \alpha_t D_t \nabla \hat{\mathcal{R}}_{\mathcal{D}_t}(\hat{y}_{\hat{\beta}_t})}$$

où $J_t = Diag(\delta_{1,t} \cdots \delta_{P,t})$ et $\mathcal{D}_t$ $subset\mathcal{D}$. Ce sont des variables aléatoires. J_t permet de sélectionner des composantes du gradient à calculer et $\mathcal{D}_t$ permet de calculer le gradient sur une partie de l'échantillon (batch). Or $\mathcal{D}_t \subset \mathcal{D}$, $\mathcal{D}_t = \{(x_{n_t(b)}, y_{n_t(b)}) \in \mathcal{X} \times \mathcal{Y}, b \in [\![1, B]\!]\}$. Et $n_t : [\![1, B]\!] \to [\![1, N]\!]$ une injection. On peut alors poser $D_t := \left[\delta_{n_t(i),j}\right] \in \mathcal{M}_{B,N}(\mathbb{R})$.

On trouve alors $\hat{\beta}_{t+1} = [I_P - \alpha_t' J_t Z^T D_t^T D_t Z]\hat{\beta}_t + \alpha_t' J_t Z^T D_t^T D_t Y$.

On pose: $P_t := D_t^T D_t = Diag(\delta_{i \in \mathrm{Im}(n_t)})$.

Théorème: Expression pour la descente de gradient stochastique

$$\hat{\beta}_t = \Big[\prod_{k=0}^{t-1}(I_P - \alpha'_k J_k Z^T P_k Z)\Big]\hat{\beta}_0 + \sum_{k=0}^{t-1}\Big[\prod_{i=k+1}^{t-1}(I_P - \alpha'_i J_i Z^T P_i Z)\Big]\alpha'_k J_k Z^T P_k Y$$

Ainsi en prenant $J \hookrightarrow Diag(\mathcal{U}([\![1,P]\!]))$ et $\mathcal{D}_t \hookrightarrow \mathcal{U}(\mathcal{D}^B)$, indépendantes deux à deux:
On a donc les mêmes résultats que précédemment, mais avec un facteur $\frac{1}{P}$ et $\frac{B}{N}$ et cette fois sur $||\mathbb{E}_J(\hat{\beta}_t) - \hat{\beta}||_2$.

$$\mathbb{E}_{\mathcal{D},J}(\hat{\beta}_t) = \Big[\prod_{k=0}^{t-1}(I_P - \alpha'_k \frac{B}{NP} Z^T Z)\Big]\hat{\beta}_0 + \sum_{k=0}^{t-1}\Big[\prod_{i=k+1}^{t-1}(I_P - \alpha'_i \frac{B}{NP} Z^T Z)\Big]\alpha'_k \frac{B}{NP} Z^T Y$$

8.3 Expressions théorique sur les MLP

On rappel que:

Définition: **Modèle MLP**

- $\hat{y}_\beta(x) = W_L \circ \sigma \circ \cdots \sigma \circ W_1(x)$, $W_l(x_l) = A_l x_l + b_l$ où $A_l \in \mathcal{M}_{S_l,E_l}(\mathbb{R})$ et $b_l \in \mathbb{R}^{S_l}$ tel que $S_l = E_{l+1}$
- $\sigma(x) = max(0,x)$ fonction ReLU.
- On optimise $\hat{\mathcal{R}}_\mathcal{D}$ par descente de gradient (stochastique) selon le paramètre $\beta = \begin{bmatrix} A_l \\ b_l \end{bmatrix}_{l\in[\![1,L]\!]}$.

Cas particulier: Dans le cas suivant, $\hat{y}_\beta(x) = W_2 \circ \sigma \circ W_1(x) = A_2\sigma(A_1 x + b_1) + b_2$, $\hat{y}_\beta : \mathbb{R}^D \to \mathbb{R}$.
$A_1 \in \mathcal{M}_{P,D}(\mathbb{R})$ et $A_2 \in \mathcal{M}_{1,P}(\mathbb{R})$. D'où avec : $\hat{\mathcal{R}}_\mathcal{D}(\hat{y}_{\hat{\beta}}) = \frac{1}{N}\sum_{i=1}^N (\hat{y}_{\hat{\beta}}(x_i) - y_i)^2$ et $\hat{y}_{\hat{\beta}}(x) = \hat{A}_2\sigma(\hat{A}_1 x + \hat{b}_1) + \hat{b}_2$ on a:

Avec $\beta = \begin{bmatrix} A_1 \\ b_1 \\ A_2 \\ b_2 \end{bmatrix}$ on a $\nabla_\beta \hat{\mathcal{R}}(\hat{y}_{\hat{\beta}}) = \begin{bmatrix} \nabla_{A_1}\hat{\mathcal{R}}(\hat{y}_{\hat{\beta}}) \\ \nabla_{b_1}\hat{\mathcal{R}}(\hat{y}_{\hat{\beta}}) \\ \nabla_{A_2}\hat{\mathcal{R}}(\hat{y}_{\hat{\beta}}) \\ \nabla_{b_2}\hat{\mathcal{R}}(\hat{y}_{\hat{\beta}}) \end{bmatrix} = \frac{2}{N}\begin{bmatrix} \Sigma_{i=1}^N x_i(\hat{A}_2 \odot \sigma'(\hat{A}_1 x_i + \hat{b}_1)^T)(\hat{y}_{\hat{\beta}}(x_i) - y_i) \\ \Sigma_{i=1}^N (\hat{A}_2 \odot \sigma'(\hat{A}_1 x_i + \hat{b}_1)^T)(\hat{y}_{\hat{\beta}}(x_i) - y_i) \\ \Sigma_{i=1}^N \sigma(\hat{A}_1 x_i + \hat{b}_1)(\hat{y}_{\hat{\beta}}(x_i) - y_i) \\ \Sigma_{i=1}^N (\hat{y}_{\hat{\beta}}(x_i) - y_i) \end{bmatrix}$

Avec $\odot$ le produit d'Hadamard.
Et on a : $\hat{\beta}_{t+1} = \hat{\beta}_t - \alpha\nabla_\beta\hat{\mathcal{R}}(\hat{y}_{\hat{\beta}_t})$. On comprend donc qu'il n'est pas aisé d'obtenir une expression manipulable analytiquement.

9 Conclusion

Ainsi, sous l'hypothèse de notre conjecture, nous avons abouti dans le cadre d'un modèle linéaire dont les features sont orthonormées et dans le cadre du pseudo-inverse comme de la descente de gradient à montrer des éléments de double descente. Nous avons ici considéré un modèle ambitieux, puisque la fonction cible se situe dans un espace plus grand que celui de l'estimateur, et avec un modèle quasi-isotropique assez général. Cependant, le résultat a été présenté sous l'hypothèse d'une conjecture sur les matrices aléatoires asymptotique. De plus, nous avons considéré une descente de gradient à pas variable, beaucoup plus réaliste que celle à pas constant.

Il reste encore de nombreuses perspectives envisageables, comme tenter de prouver notre conjecture, quitte à la modifier potentiellement en régime sous-paramétré, étudier en détails la descente de gradient stochastique en étudiant l'espérance du risque et non le risque de l'espérance, et explorer d'autres modèles comme le Perceptron Multicouche.

Afin d'aller plus loin sur ce problème, ces cours peuvent être intéressants [10], [2]. Je tiens à remercier Samuel Gruffaz pour son aide importante lors de ce stage, et Nicolas Vayatis qui nous a encadré.

10 Bibliographie

References

[1] Francis Bach. "High-dimensional analysis of double descent for linear regression with random projections". In: *SIAM Journal on Mathematics of Data Science* 6.1 (2024), pp. 26–50.

[2] Francis Bach. *Learning Theory from First Principles*. `https://www.di.ens.fr/~fbach/ltfp_book.pdf`. [Online]. 2024.

[3] 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.

[4] Mikhail Belkin et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off". In: *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854.

[5] Friedrich Götze and Alexander Tikhomirov. "Rate of convergence in probability to the Marchenko-Pastur law". In: *Bernoulli* 10.3 (2004), pp. 503–548.

[6] Emett Haddad. *Github Double Descente*. `https://github.com/EmettGabrielH/Double-descente---Emett-Haddad`. [Online]. 2024.

[7] 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.

[8] Zhenyu Liao, Romain Couillet, and Michael W Mahoney. "A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 13939–13950.

[9] Rylan Schaeffer et al. "Double descent demystified: Identifying, interpreting & ablating the sources of a deep learning puzzle". In: *arXiv preprint arXiv:2303.14151* (2023).

[10] MM Wolf. *Mathematical foundations of supervised learning (growing lecture notes)*. `https://mediatum.ub.tum.de/doc/1723378/1723378.pdf`. 2018.

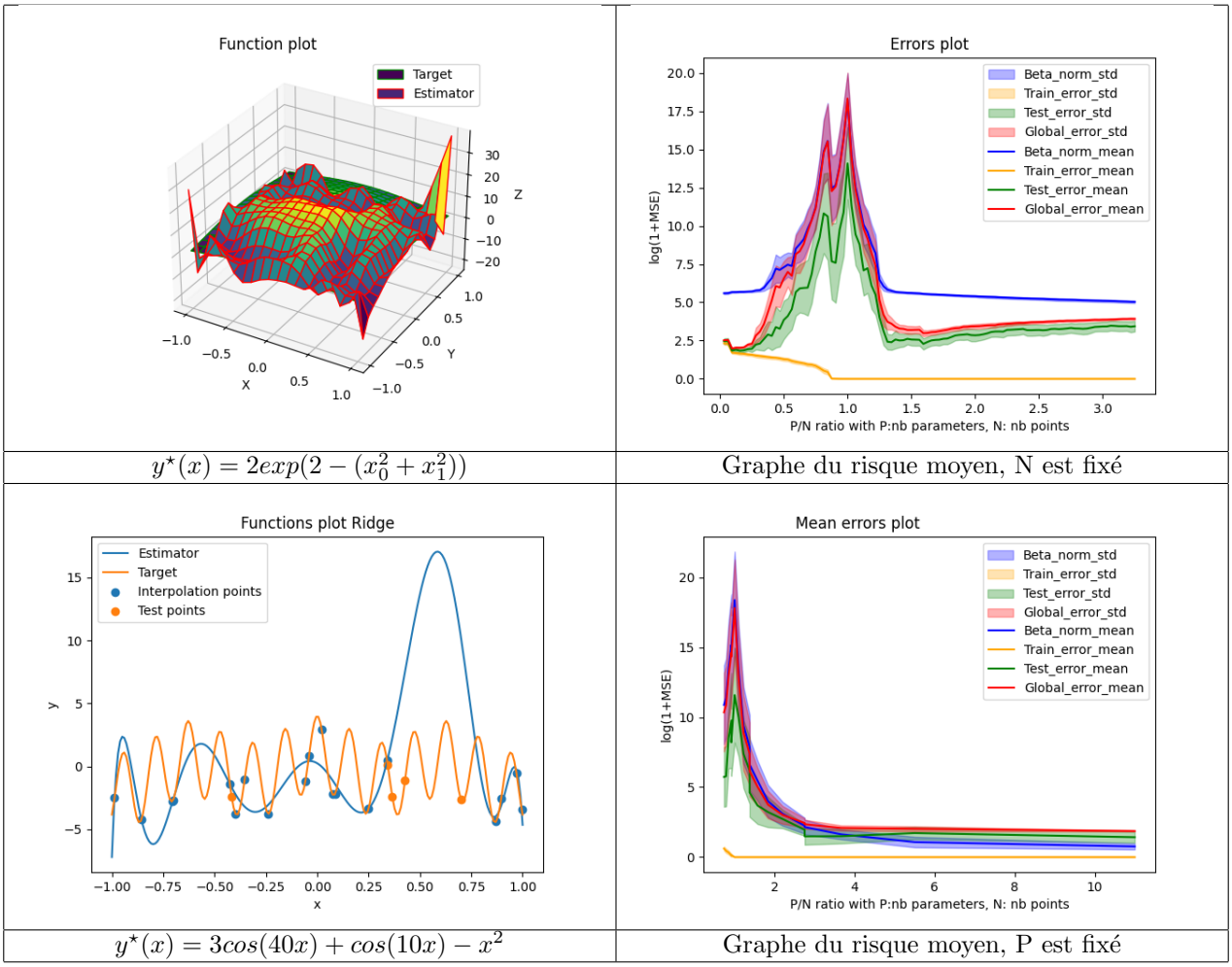
[11] Yong-Qua Yin, Zhi-Dong Bai, and Pathak R Krishnaiah. "On the limit of the largest eigenvalue of the large dimensional sample covariance matrix". In: *Probability theory and related fields* 78 (1988), pp. 509–521.

11 Appendices

A. Résultats expérimentaux

Il est important de mettre en parallèle des résultats théoriques et des expériences numériques, car cela permet d'apporter un autre point de vue à nos théorèmes. De plus, cela permet de fournir les intuitions nécessaires à la construction de la théorie de la double descente, qui est en premier lieu une observation empirique.

a. Régression polynomial



$y^\star(x) = 2exp(2 - (x_0^2 + x_1^2))$	Graphe du risque moyen, N est fixé
$y^\star(x) = 3cos(40x) + cos(10x) - x^2$	Graphe du risque moyen, P est fixé

Figure 1: 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).
Cependant, ce qui nous intéresse en pratique est bien de fixer l'échantillon $\mathcal{D}$ et donc de fixer N et de faire varier P. On utilise ici l'algorithme 3.

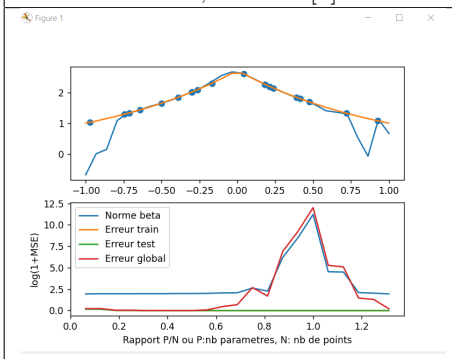
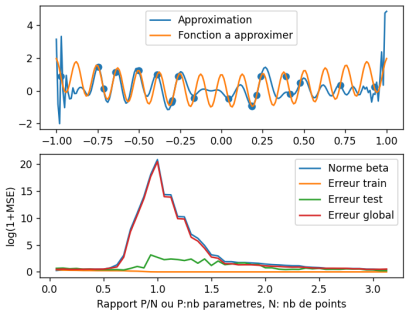
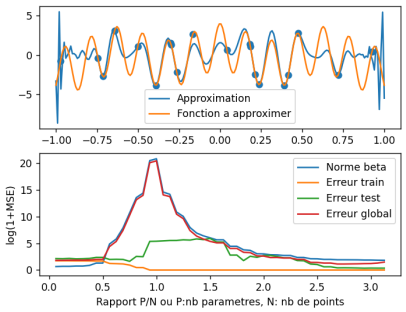
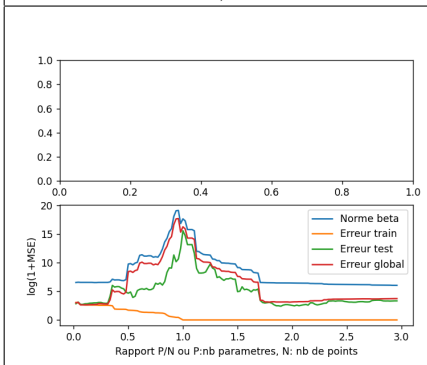
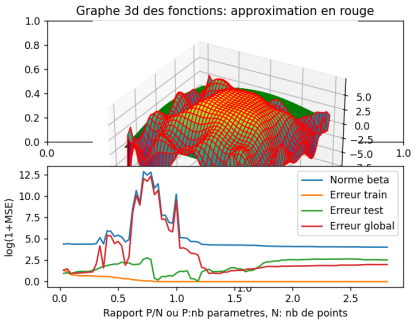
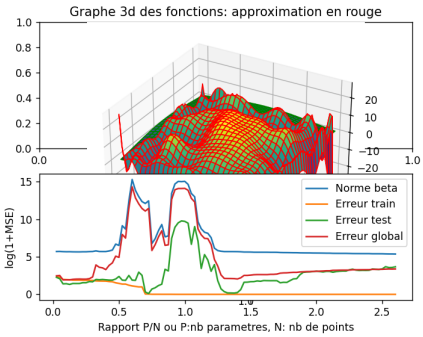
D = 1, N = 16 [6]	D = 1, N = 16	D = 1, N = 16
		
$y^\star(x) = exp(1 - \|x\|)$	$y^\star(x) = cos(50x) + x^2$	$y^\star(x) = 3cos(40x) + cos(10x) - x^2$
D = 3, N = 48	D = 2, N = 32	D = 2, N= 40
		
$y^\star(x) = exp(3 - (x_0^2 + x_1^2 + x_2^2))$	$y^\star(x) = exp(2 - (x_0^2 + x_1^2))$	$y^\star(x) = 2exp(2 - (x_0^2 + x_1^2))$

Figure 2: Phénomène de double descente en dimension 1, 2 et 3 pour des features polynomiales

On observe dans le cas de la régression polynomial simple (sans facteur de régulation) un phénomène de double descente, et ce quelque soit la dimension en entrée de la fonction cible. On observe dans chacun de ces cas un minimum inférieur de l'erreur global en régime sur-paramétré.
En effet, on constate que lorsque l'on s'approche du seuil d'interpolation, l'estimateur tend à diverger afin de pouvoir coller exactement au données d'entraînement, tout en ne disposant pas de suffisamment de degré de liberté.

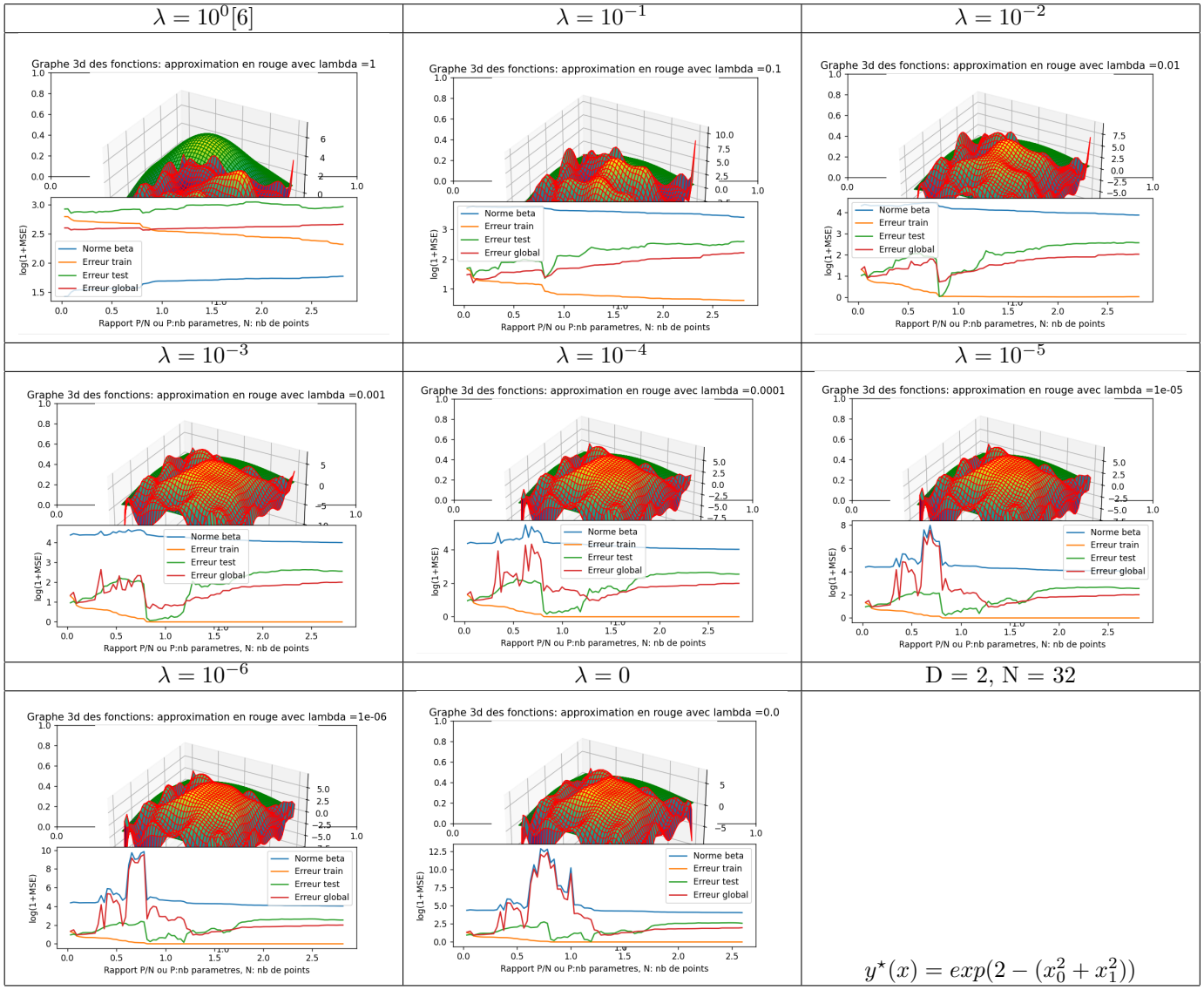


Figure 3: Influence d'un facteur de régulation sur l'apparition du phénomène de double descente.

On remarque que la régulation permet d'empêcher totalement la venue d'un phénomène de double descente par rapport au nombre de paramètre. En effet, on peut considérer que la réelle complexité réside dans la norme du paramètre β. Or, on a une descente simple par rapport à ce paramètre, comme nous l'avons remarqué avec le théorème de la décroissance de $\hat{\beta}$ 8.1. D'où le résultat.

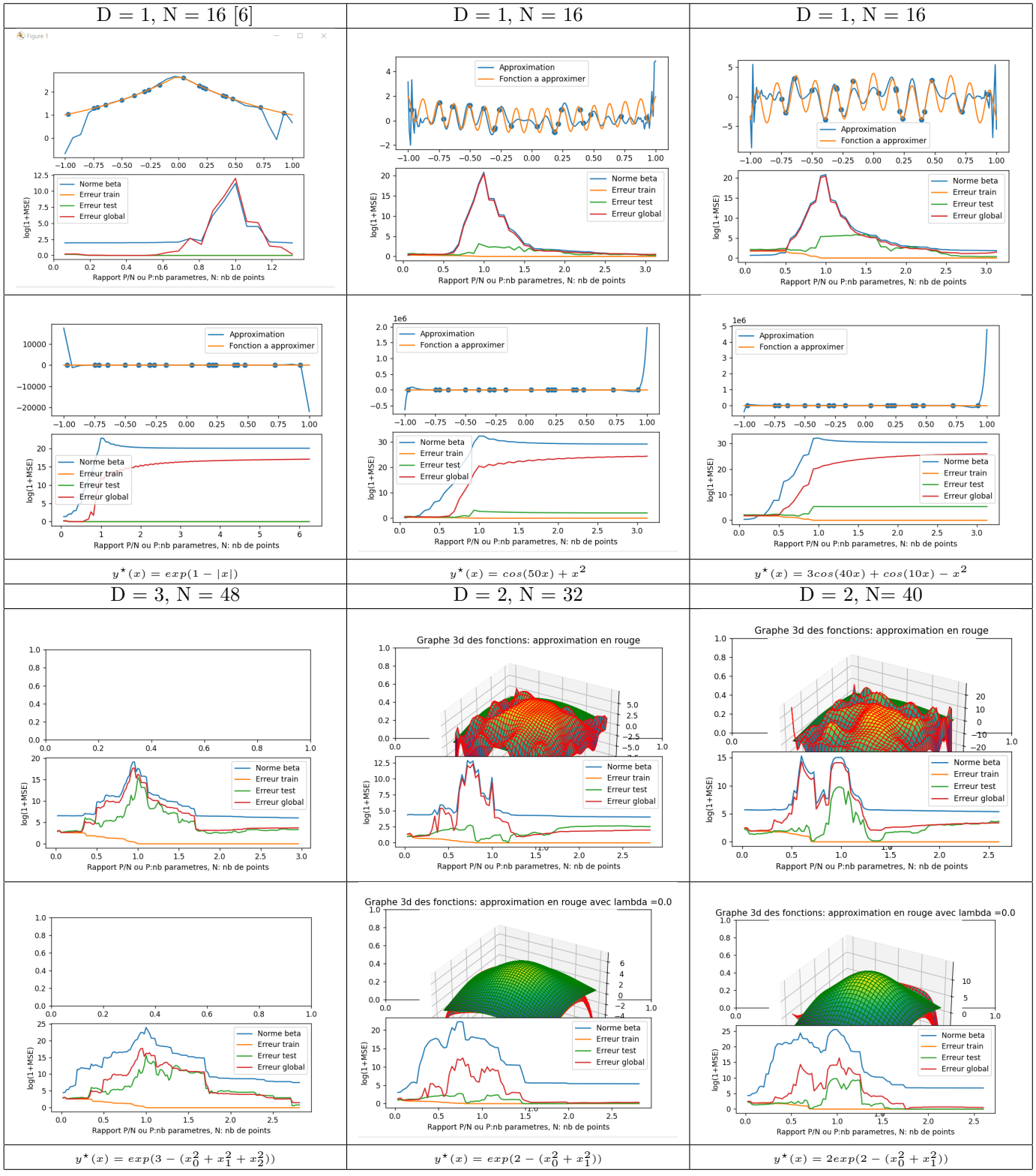
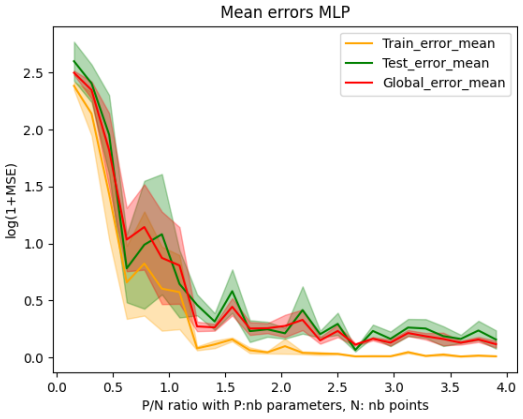


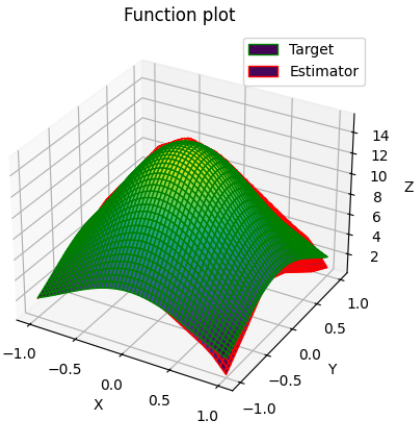
Figure 4: Influence du choix d'une base orthonormalisée/canonique sur la double descente

On remarque qu'orthonormaliser l'espace permet s'accélérer l'avenue du phénomène de double descente en faible dimension.

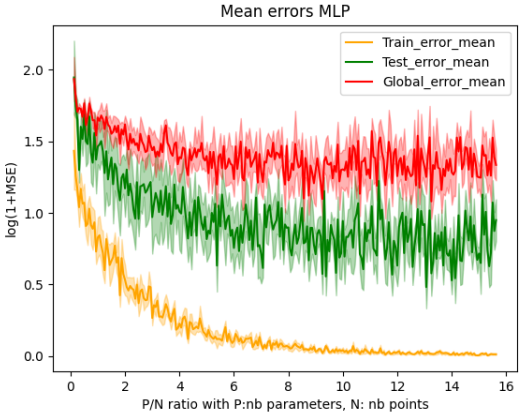
b. Perceptron Multicouche (MLP)



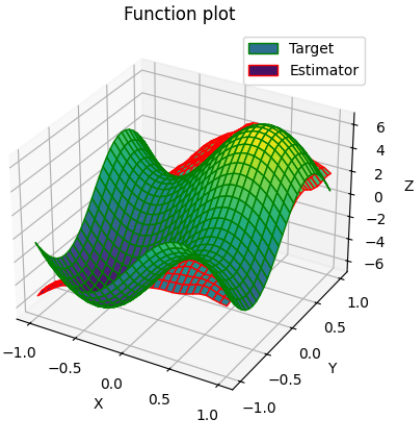
Risque moyen en fonction de $P\backslash N$, MLP



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



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



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

Figure 5: Graphe de la MSE d'un algorithme Perceptron Multicouche.

On peut constater avec le premier exemple qu'un phénomène de double descente peut apparaître en moyenne sur les MPL, cependant il peut aussi ne pas survenir, comme on peut le voir avec le deuxième exemple. On utilise ici l'algorithme 4.

B. Détails numériques

De manière général, nous avons toujours utilisé $a = 1$ la taille du cube, $r = 0.2$ le ratio entre l'échantillon test/train. Et on utilise 23334 comme random seed par défaut.

Figure 1:
Exemple 1: Base orthonormée de degrée homogène maximal Deg = 13. n = 20, M = 40, Lambda = 0,
Liste_seeds = [1000,12334,13902,29389,1233983,18938309,22387928,37289749,232382,42,3297].
Exemple 2: Base orthonormée de degrée homogène maximal Deg = 10. M_min = 2 , M_max = 20, n = 200 Lambda = 0
Liste_seeds = [1000,12334,13902,29389,1233983,18938309,22387928,37289749,232382,42,3297, 567547,378647,44478,287974,37846

Figure 2: Ici, on considère une base orthonormée de degrée homogène maximal Deg.
Exemple 1: M = 20, n = 30, Deg = 50 — Exemple 2: M = 20, n = 200, Deg = 50
Exemple 3: M = 20, n = 200, Deg = 50 — Exemple 4: M = 60, n = 30, Deg = 8
Exemple 5: M = 40, n = 40, Deg = 12 — Exemple 6: M = 50, n = 60, Deg = 13

Figure 3: Ici, on réutilise l'exemple 5 de la figure 2 (M = 40, n = 40, Deg = 12) mais avec λ facteur de régulation qui varie entre 1 et 0, en utilisant l'algorithme 3.
Figure 4: Même entrées que pour 2 mais on compare base orthonormée et base canonique.

Figure 5: (MLP)
Exemple 1: n = 100, M = 40, Epochs = 10000, alpha = 10^{-3}, P_min = 1, P_max = 25,
Liste_seeds = [1000,12334,13902,29389]
Exemple 2: n = 30, M = 20, Epochs = 5000, alpha = 10^{-4}, P_min= 2, P_max = 250
Liste_seeds = [1000,1001,1020,1234,1984]

C. Pseudo-code

a. Modèle linéaire et pénalisé

Cf: `https://ipolcore.ipol.im/demo/clientApp/demo.html?id=77777000515`

Algorithm 1: Generation of the orthonormal polynomial basis (Gram- Schmidt)

function generate_orthonormal_basis(D, D′, C)
  Input D, D', C**:** $D \in \mathbb{N}^\star$,$D' \in \mathbb{N}^\star$, $C \subset \mathbb{R}^D$
  $\text{Basis}_{\text{D}'} = [\Pi_{\text{d}=1}^{\text{D}} \text{X}_{\text{d}}^{\alpha_{\text{d}}}, \sum \alpha_{\text{d}} \leq \text{D}']$
  $\text{P} = \text{LENGHT}(\text{Basis}_{\text{D}'})$
  for $1 \leq p \leq P$ **do**
    $f'_p = \text{Basis}_{\text{D}'}[\text{p}] - \sum_{\text{i}=1}^{\text{p}-1} [\int_{\text{C}} \text{Basis\_ortho}_{\text{D}'}[\text{i}] \cdot \text{Basis}_{\text{D}'}[\text{i}]] \times \text{Basis\_ortho}_{\text{D}'}[\text{i}]$
    $\text{Basis\_ortho}_{\text{D}'}[\text{p}] = \frac{\text{f}'_{\text{p}}}{||\text{f}'_{\text{p}}||_2}$
  return $\text{Basis\_ortho}_{\text{D}'}$

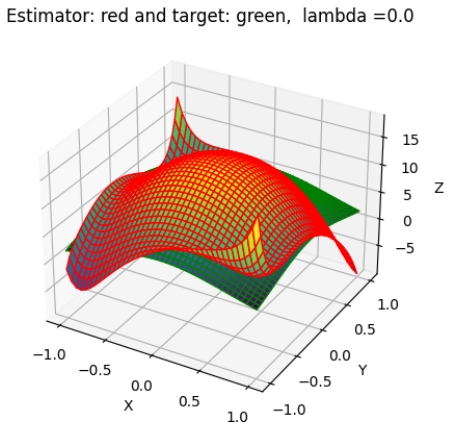
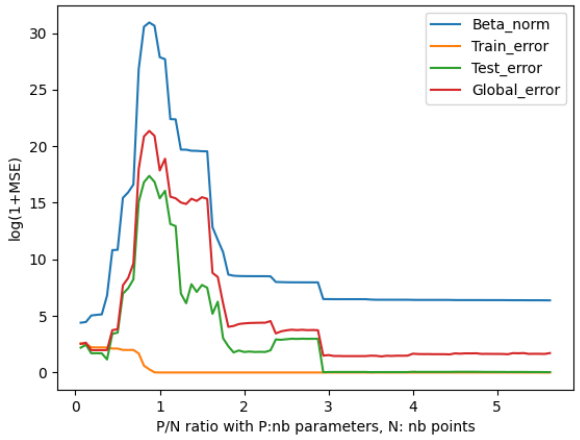
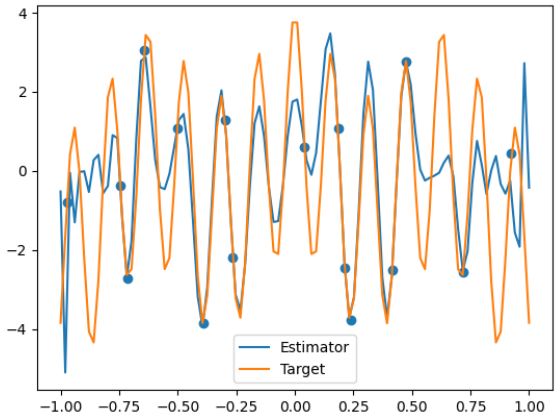
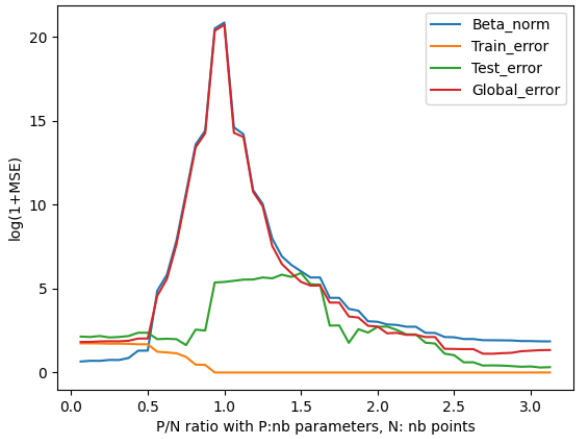
Algorithm 2: Dataset initialisation

function dataset_initialisation(f, C, M, ratio_data)
  Input y, C, M, ratio_data**:** $y : \mathbb{R}^D \to \mathbb{R}$, $C = [[a_d, b_d], 1 \leq d \leq D]$, ratio_data $\in [0, 1]$
  $U = \mathcal{U}([0,1]^{(M,D)})$
  $X = Diag(a_d) + UDiag(b_d - a_d)$ $\#$ $X \hookrightarrow \mathcal{U}(C^M)$
  $Y = y(X)$
  $X_{train}, X_{test}, Y_{train}, Y_{test} = \text{test\_split}(X, Y, \text{ratio\_data})$
  return $\text{X}_{\text{train}}, \text{X}_{\text{test}}, \text{Y}_{\text{train}}, \text{Y}_{\text{test}}$

Algorithm 3: Ridge Regression

function ridge_regression(P_{min}, P_{max}, Features, λ, Data, Data_global)
  $X_{train}, X_{test}, Y_{train}, Y_{test}$ = Data
  X_{global}, Y_{global} = Data_global
  for $P_{min} \leq p \leq P_{max}$ **do**
    $\Phi_p = [f_p, 1 \leq p \leq P]^T$ # (f_p) an orthonormal basis for p.s. on C
    $Z_p = \Phi_p^T(X_{train})$ # $Z_p \in \mathcal{M}_{N,P}(\mathbb{R})$
    $\hat{\beta}_p = \begin{bmatrix} Z_p \\ \sqrt{N\lambda} I_p \end{bmatrix}^{\dagger} \begin{bmatrix} Y_{train} \\ O_p \end{bmatrix}$
    Train_error[p] = $\log(1 + \mathrm{MSE}(Y_{train}, Z_p \hat{\beta}_p))$
    Test_error[p] = $\log(1 + \mathrm{MSE}(Y_{test}, \Phi_p^T(X_{test})\hat{\beta}_p))$
    Global_error[p] = $\log(1 + \mathrm{MSE}(Y_{global}, \Phi_p^T(X_{global})\hat{\beta}_p))$
    Beta_norm[p] = $\log(1 + ||\hat{\beta}_p||_2^2)$
  $\hat{y} = \Phi_p \cdot \hat{\beta}_{P_{max}}$
  return Train_error, Test_error, Beta_norm, Global_error, ŷ

Exemples: (random seed : 23334)

Fonction cible et estimateur	Graphe du risque
	
$y(x) = 2e^{2-(x_1^2+x_2^2)}, Basis : cano$	$M = 20, r = 0.2, n = 40, a = 1, Deg = 12, D = 2, \lambda = 0$
	
$y(x) = 3cos(40x) + cos(10x) - x^2, Basis : ortho$	$M = 20, r = 0.2, n = 100, a = 1, Deg = 50, D = 1, \lambda = 0$

b. Modèle MLP

Cf : `https://ipolcore.ipol.im/demo/clientApp/demo.html?id=77777000527`

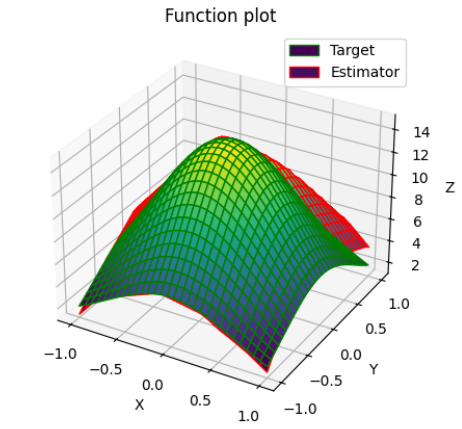
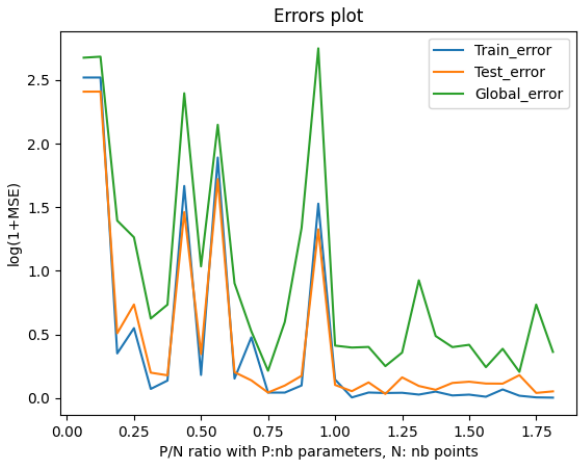
Algorithm 4: MLP Gradient Descent

function MLP_Gradient_Descent(P_{min}, P_{max}, Epochs, α, Data, Data_global)
  $X_{train}, X_{test}, Y_{train}, Y_{test}$ = Data
  X_{global}, Y_{global} = Data_global
  for $P_{min} \leq p \leq P_{max}$ **do**
    $MLP = Affine(1, P) \circ \sigma \circ Affine(P, D)$ # Creation of the MLP structure
    $MLP.fit(X_{train}, Y_{train}, Epochs, \alpha)$
    Train_error[p] = $\log(1 + \text{MSE}(Y_{train}, \text{MLP}(X_{train})))$
    Test_error[p] = $\log(1 + \text{MSE}(Y_{test}, \text{MLP}(X_{test})))$
    Global_error[p] = $\log(1 + \text{MSE}(Y_{test}, \text{MLP}(X_{global})))$
  $\hat{y} = MLP(P_{max})$
  return Train_error, Test_error, Global_error, $\hat{y}$

Algorithm 5: MLP Gradient Descent FIT

function MLP.fit(X_{train}, Y_{train}, Epochs, α)
  for $1 \leq epoch \leq Epochs$ **do**
    for $(x_{train}, y_{train}) \in \mathcal{D}$ **do**
      $score\_gradient = MLP(x_{train}) - y_{train}$ # gradient of MSE
      $backpropagation(score\_gradient)$
    $update(\alpha)$ # udpate weights

Exemples: (random seed : 23334)

Fonction cible et estimateur	Graphe du risque
	
$y(x) = 2e^{2-(x_1^2+x_2^2)}$, $Epochs = 1000$, $P_{max} = 30$	$M = 20,\ r = 0.2,\ n = 30,\ a = 1,\ D = 2,\ \alpha = 0.001$