

Apprentissage amélioré des features via régularisation : réseaux neuronaux et méthodes à noyaux

Elie LELOUCHE, Théo RENE, Jules ROBIN

Janvier 2024

Résumé

Rapport écrit réalisé dans le cadre de l'UE Apprentissage Statistique. On s'appuie sur un papier co-écrit par Francis Bach et Bertie Follain, intitulé **Enhanced Feature Learning via Regularisation : Integrating Neural Networks and Kernel Methods** [FB24].

Table des matières

1	Introduction	1
1.1	Apprentissage supervisé et Modèle linéaire	1
1.2	Kernel Ridge Regression	2
1.3	Réseaux Neuronaux	3
1.4	Le meilleur des deux Univers ?	4
2	BKerNN : un modèle unissant Méthodes à noyaux et Réseaux neuronaux	5
2.1	L'espace \mathcal{H} est un RKHS	6
2.2	Paradigme des noyaux et interprétation	7
3	Calcul de l'estimateur	9
3.1	Procédure d'optimisation	9
3.2	Robustesse de l'optimisation	10
4	Analyse statistique	11
4.1	Résultat principal	11
4.2	Raffinement : bornes sur la complexité gaussienne	14
5	Conclusion	15

1 Introduction

1.1 Apprentissage supervisé et Modèle linéaire

La Statistique Mathématique fournit de nombreux paradigmes pour l'élaboration d'algorithmes dits "d'apprentissage", mais aussi pour l'analyse mathématique de leurs qualités d'estimation, leurs facultés de convergence vers une solution. Le plus fameux, si ce n'est celui autour duquel gravite la grande majorité des techniques d'apprentissage moderne, est celui de la minimisation d'un risque empirique dont nous rappelons la définition de manière succincte ici. Soit $(x_1, y_1), \dots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$ un échantillon de variables aléatoires i.i.d. On cherche à modéliser les réponses y comme fonction de x , $f(x)$, où f appartient à une classe de fonction \mathcal{F} prédéfinie et est telle que :

$$f \in \arg \min_{f \in \mathcal{F}} \hat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i))$$

où ℓ est une fonction de perte, $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ où $\ell(y, f(x))$ mesure l'écart entre la vraie réponse y et la modélisation $f(x)$.

La fonction $\hat{\mathcal{R}}$ est une approximation empirique du risque théorique $\mathcal{R}(f) = \mathbb{E}_{X,Y}[\ell(Y, f(X))]$. Un exemple simpliste mais ayant la vertu de la pédagogie serait le modèle linéaire utilisé en régression. Cependant, s'il y a bien un élément le rendant incomplet, c'est le caractère trop restrictif des hypothèses émises. Il semble tout à fait légitime de ne pas supposer linéaire la fonction à estimer. Vient alors la question de l'aptitude du modèle à généraliser ses facultés d'estimation pour des fonctions d'intérêt de types plus larges.

Un autre défi que ce modèle seul ne saurait affronter est "le fléau de la dimension". En effet, lorsque le nombre de caractéristiques décrivant les données augmente, la vitesse de convergence, elle, diminue grandement.

On aimerait donc sélectionner des caractéristiques pertinentes, contenant en elles-mêmes une partie majeure de l'information contenue dans les données. Il est ainsi possible de penser à la régression linéaire dite LASSO, consistant en l'ajout d'une contrainte sur la norme ℓ_1 du vecteur de caractéristiques, permettant d'ajouter de la "sparsité" aux modèles, c'est-à-dire de réduire à zéro certains paramètres et de sélectionner de manière automatique les caractéristiques pertinentes.

1.2 Kernel Ridge Regression

D'autres méthodes, devenues paradigmes centraux en théorie de l'apprentissage, permettent de surpasser certaines de ces limitations : les méthodes à noyaux.

Donnons ici une description succincte d'un des modèles à noyaux : le modèle KRR ("Kernel Ridge Regression").

Définition 1. On considère $(\varphi(x_i))_{i=1}^n$ une projection des données $(x_i)_{i=1}^n$ dans un espace de Hilbert \mathcal{H} de plus grande dimension. On cherche alors à résoudre le problème :

$$\theta \in \arg \min_{\mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \varphi(x_i), \theta \rangle_{\mathcal{H}}) + \frac{\lambda}{2} \|\theta\|_2^2$$

en supposant cette fois-ci $\ell : \mathcal{Y} \times \mathbb{R} \rightarrow \mathbb{R}$.

Comme en témoigne la définition, la méthode consiste en l'application d'un modèle linéaire dans un espace de Hilbert de plus grande dimension. On choisit donc en général φ telle que la fonction $x \mapsto \langle \varphi(x), \theta \rangle_{\mathcal{H}}$ finalement obtenue soit non linéaire. La fonction φ est choisie de telle sorte à ce qu'elle appartienne à un espace fonctionnel de Hilbert \mathcal{H} particulier nommé *RKHS* ("Reproducing Kernel Hilbert Space" en anglais) dont nous rappelons la définition ici.

Définition 2. (Noyau) Une fonction $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ est appelée un noyau s'il existe un espace de Hilbert $(\mathcal{G}, \langle \cdot, \cdot \rangle)$ et une application $\phi : \mathcal{X} \rightarrow \mathcal{G}$ telle que :

$$\forall (x, x') \in \mathcal{X}^2 : k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{G}}$$

On appelle ϕ une fonction de caractéristique et \mathcal{G} un espace de caractéristique.

Définition 3. (RKHS) Soit $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ un espace de Hilbert et k un noyau. \mathcal{H} est un RKHS de noyau k (ou k est un noyau de reproduction de \mathcal{H}) si pour tout $x \in \mathcal{X}$:

- $k(\cdot, x) \in \mathcal{H}$
- $\forall f \in \mathcal{H} : \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ (propriété de reproduction)

Ce choix nous permet de faire face à un problème d'optimisation de dimension infinie en se ramenant à un problème de dimension finie. Pour se faire, on fait appel au théorème du représentant. En voici l'énoncé :

Proposition 1 (Théorème du Représentant). *Considérons une fonction de caractéristique $\varphi : \mathcal{X} \rightarrow \mathcal{H}$ où \mathcal{H} est un RKHS. Soit $(x_1, \dots, x_n) \in \mathcal{X}^n$, et supposons que la fonctionnelle $\Psi : \mathbb{R}^{n+1} \rightarrow \mathbb{R}$ soit strictement croissante par rapport à sa dernière variable. Alors, l'infimum de*

$$\Psi((\langle \theta, \varphi(x_1) \rangle), \dots, (\langle \theta, \varphi(x_n) \rangle), \|\theta\|^2)$$

peut être obtenu en restreignant θ à une forme du type

$$\theta = \sum_{i=1}^n \alpha_i \varphi(x_i),$$

avec $\alpha \in \mathbb{R}^n$.

On applique ce théorème dans un cadre d'apprentissage en prenant $\psi := \hat{\mathcal{R}}$ respectant évidemment les conditions. L'ajout d'une pénalité sur la norme ℓ_2 permet aussi de faire face au fléau de la dimension via une sélection des caractéristiques. Les méthodes à noyaux ont de même la belle propriété de fournir des bornes sur le risque, indépendantes de la dimension.

Cependant, cette méthode soulève alors la question de la fonction de projection à choisir. Quel espace pourrait le plus correspondre à la description des données ?

Aussi, le choix du noyau et de la fonction de projection imposera une description non linéaire, d'une forme unique, sur la totalité des caractéristiques, même si certaines caractéristiques peuvent n'avoir qu'une dépendance linéaire avec la caractéristique d'objectif. On se voit alors confronté à une forme plus raffinée de souci de sélection des caractéristiques. Quand bien même nous supposons que certaines caractéristiques décrivent les données d'intérêt, pourquoi les décriraient-elles toutes de manière non linéaire complexe ? Pourquoi obliger l'algorithme à supporter la complexité d'un modèle non linéaire sur l'ensemble des caractéristiques ?

Une méthode permet de surpasser certaines de ces limitations : les réseaux neuronaux.

1.3 Réseaux Neuronaux

Avant de donner une heuristique sur les réseaux de neurones, nous en donnerons la formulation mathématique, qui sera largement utile pour le reste de l'exposé.

Définition 4. Soient $w_j \in \mathbb{R}^d$, $\eta_j \in \mathbb{R}$, $b_j \in \mathbb{R}$, $j \in \{1, \dots, n\}$, $b \in \mathbb{R}$ et $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ une fonction dite d'activation. On définit alors la fonction :

$$f(x) = \sum_{j=1}^n \eta_j \sigma(w_j x_j + b_j)$$

La fonction f ainsi définie est appelée "réseau de neurones à une couche cachée". Cette fonction représente une forme élémentaire de réseaux de neurones. Son atout majeur par rapport à un modèle linéaire classique est la présence de la fonction d'activation σ . Cette dernière est en générale choisie non linéaire et ce, pour faire face à la première préoccupation énoncée plus haut : le souci de décrire les données de manière non uniquement linéaire.

Il est possible de penser aux réseaux de neurones comme un modèle linéaire de poids η_j appliquées sur les données projetées $\sigma(w_j x_j + b_j)$ on définit alors la fonction $\phi(x, w, b) = \sigma(wx + b)$ et réécrit la fonction f de la manière suivante :

$$f(x) = \sum_{j=1}^n \eta_j \phi(w_j, x_j, b)$$

Cette reformulation fait étonnamment penser à une méthode précédemment décrite : les méthodes à noyaux ! A la seule différence, et c'est en partie en ce sens que réside la pertinence et l'ingéniosité des réseaux de neurones : la dépendance du noyau ϕ en des poids w_j . Cette dépendance en les poids permettra de faire en sorte à ce que le noyau ϕ soit choisi, de manière automatique en fonction de la structure des données. La question n'est plus alors de choisir l'espace de fonction le plus adapté mais uniquement de choisir l'espace de fonction qui saura s'adapter !

En ce qui concerne le fléau de la dimension, il suffira d'ajouter un paramètre de régularisation ℓ_1 ou ℓ_2 afin d'introduire de la sparsité au modèle. On pourra aussi jouer sur le choix de la fonction

d'activation et prendre une fonction d'activation pouvant introduire du seuillage, il est ainsi naturel de penser à la fonction ReLU :

$$\sigma(x) = \mathbf{1}_{\{x>0\}}$$

Les réseaux neuronaux ont aussi la belle propriété de profiter de la relation potentiellement linéaire de certaines caractéristiques avec les paramètres d'intérêt en terme de vitesses de convergence (cf. [Bac24] §9.3.5), à contrario des méthodes à noyaux.

Se pose cependant la question de la fonction d'activation à choisir, laquelle sera la plus pertinente pour nos données ?

Une autre difficulté vient s'ajouter : celle de la résolubilité du problème d'optimisation ainsi induit. En effet, si les méthodes à noyaux présentaient jusqu'à présent des problèmes d'optimisation numériquement agréables de par l'appartenance de la fonction ϕ à un *RKHS*, rien ne dit que les fonctions d'activation appartiennent à ce type d'espace fonctionnel. Bien au contraire, il s'avère que bon nombre de problèmes d'optimisation issus de modèles neuronaux présentent certaines aspérités numériques comme la non convexité de la fonction à optimiser, fonction souvent non-lisses rendant les problèmes computationnellement ardues et complexes.

1.4 Le meilleur des deux Univers ?

Naît alors l'idée naturelle d'essayer de développer une méthode tirant du meilleur des deux mondes : celui des méthodes à noyaux et des réseaux neuronaux.

C'est bien ce qu'ont essayé de réaliser Follain et Bach dans [FB24], article sur lequel nous nous concentrerons dans cet exposé, via un modèle nommé BKerNN. L'idée fondatrice de ce modèle est la suivante : les limites computationnelles liées aux réseaux de neurones proviennent en partie de la non appartenance de la fonction d'activation à un RKHS, il serait donc pertinent de prendre une fonction d'activation appartenant à un RKHS afin d'hériter des bonnes propriétés numériques liées aux noyaux.

Follain et Bach dans [FB24] ont de même eu l'idée d'augmenter la dépendance du réseau de neurones des données en exprimant des fonctions d'activations dépendant des poids w_j et approfondir ainsi l'idée d'un réseau de neurones comme une méthode à noyaux où le noyau est appris à l'aide des données.

Dans cet exposé il sera alors question de présenter la méthode BKerNN ainsi que les résultats théoriques les plus fondamentaux, que ces derniers concernent l'optimisation de la fonction de perte ou les analyses statistiques du modèle.

2 BKerNN : un modèle unissant Méthodes à noyaux et Réseaux neuronaux

Dans cette section il s'agira de présenter le modèle. Nous insisterons sur la présentation d'heuristiques sur le modèle. Nous définirons en premier temps quelques objets mathématiques utiles à la suite de l'exposé puis nous concentrerons sur leur pertinence dans le cadre des objectifs du modèle présenté

Définition 5. Soit :

$$\mathcal{F}_\infty := \left\{ f \mid f(\cdot) = c + \int_{\mathcal{S}^{d-1}} g_w(w^T \cdot) d\mu(w) \right\}$$

où c est une constante dans \mathbb{R} , \mathcal{S}^{d-1} est la sphère unité pour une norme quelconque $\|\cdot\|$ sur \mathbb{R}^d (typiquement ℓ_2 ou ℓ_1), $\mu \in \mathcal{P}(\mathcal{S}^{d-1})$ est une mesure de probabilité sur \mathcal{S}^{d-1} , et pour tout $w \in \mathcal{S}^{d-1}$, $g_w : \mathbb{R} \rightarrow \mathbb{R}$ appartient à un espace de fonctions \mathcal{H} . On définit \mathcal{H} comme :

$$\mathcal{H} = \left\{ g : \mathbb{R} \rightarrow \mathbb{R} \mid g(0) = 0, g \text{ admet une dérivée faible } g', \int_{\mathbb{R}} (g')^2 < \infty \right\}.$$

\mathcal{H} est un espace de Hilbert et un espace de Sobolev, muni du produit scalaire défini par $\langle \tilde{g}, g \rangle_{\mathcal{H}} = \int \tilde{g}' g'$. Notons que g_w dépend du poids w .

Définition 6. Soit :

$$\mathcal{F}_m := \left\{ f \mid f(\cdot) = c + \frac{1}{m} \sum_{j=1}^m g_j(w_j \cdot), w_j \in \mathcal{S}^{d-1}, g_j \in \mathcal{H} \right\}$$

Remarque 1. Remarquons que $\mathcal{F}_m \subset \mathcal{F}_\infty$. Pour s'en convaincre il suffira de considérer la mesure de probabilité discrète :

$$\mu = \frac{1}{m} \sum_{j=1}^m \delta_{w_j}$$

On pourra penser à une fonction dans \mathcal{F}_m comme un réseau de neurones de fonctions d'activation $\sigma = g_j$ de poids de sortie $\eta_j = \frac{1}{m}$ constants et de poids d'entrée $w_j = w_j$ en reprenant les notations données en introduction. L'espace \mathcal{F}_m pourra être vu comme une "version discrète" de \mathcal{F}_∞ .

Comme nous le verrons plus loin dans l'exposé, la fonction à estimer appartiendra concrètement à \mathcal{F}_m . L'espace \mathcal{F}_∞ est utile pour des considérations d'analyses statistiques. Il sera alors question de faire appel à un concept très largement connu (notamment en physique statistique et en théorie des jeux) : la limite de champs moyens cf. [Bac24]. Il consiste en l'approximation du comportement d'un nombre fini assez large de particules en un comportement global, moyen. Dans le cadre d'un réseau de neurones lorsque la taille de l'ensemble $\{1, \dots, m\}$, m , tend vers l'infini, la somme pondérée définie en Définition 2, peut approximer l'intégrale sur la sphère unité \mathcal{S}^{d-1} contre la mesure de probabilité μ concernée. Cette approche est intéressante car, comme le notent Chizat et Bach (cf. [BC21]), sous certaines conditions (convexité des fonctions de perte et de pénalité, homogénéité de la fonction d'activation), le problème de risque empirique régularisé optimisé par descente de gradient de pas infiniment petit converge vers le minimiseur du problème correspondant avec un nombre infini de particules. Cela nous permet d'utiliser un nombre fini de particules m dans la pratique, tout en profitant des avantages théoriques dérivant du cadre continu.

Définition 7. Soit la fonction de pénalité :

$$\Omega_0(f) = \inf_{c \in \mathbb{R}, (g_w)_w \in \mathcal{H}^{\mathcal{S}^{d-1}}, \mu \in \mathcal{P}(\mathcal{S}^{d-1})} \int_{\mathcal{S}^{d-1}} \|g_w\|_{\mathcal{H}} d\mu(w),$$

tel que $f = c + \int_{\mathcal{S}^{d-1}} g_w(w^T \cdot) d\mu(w)$, et $\|g_w\|_{\mathcal{H}} := \int_{-\infty}^{+\infty} g'_w(t)^2 dt$.

On déduira la définition de la pénalité dans le cas où $f \in \mathcal{F}_m$ grâce à la Remarque 1.

La fonction objectif de l'algorithme sera alors :

$$\hat{f}_\lambda := \arg \min_{f \in \mathcal{F}} \hat{\mathcal{R}}(f) + \lambda \Omega(f) \quad (1)$$

où $\lambda > 0$ et, dans le cadre de cet exposé, Ω est Ω_0 défini précédemment. L'espace de fonction \mathcal{F} sera ou \mathcal{F}_m ou \mathcal{F}_∞ .

Ecrivons à présent l'équation dans le cas $f \in \mathcal{F}_\infty$:

$$\arg \min_{\substack{c \in \mathbb{R}, (g_w)_{w \in \mathcal{H}} \in \mathcal{H}^{\mathcal{S}^{d-1}} \\ \mu \in \mathcal{P}(\mathcal{S}^{d-1})}} \frac{1}{n} \sum_{i=1}^n \ell \left(y_i, c + \int_{\mathcal{S}^{d-1}} g_w(w^T x_i) d\mu(w) \right) + \lambda \int_{\mathcal{S}^{d-1}} \|g_w\|_{w \in \mathcal{H}} d\mu(w) \quad (2)$$

ainsi que dans le cas $f \in \mathcal{F}_m$:

$$\arg \min_{\substack{c \in \mathbb{R}, w_1, \dots, w_m \in \mathcal{S}^{d-1} \\ g_1, \dots, g_m \in \mathcal{H}}} \frac{1}{n} \sum_{i=1}^n \ell \left(y_i, c + \frac{1}{n} \sum_{j=1}^m g_j(w_j^T x_i) \right) + \lambda \frac{1}{m} \sum_{j=1}^m \|g_j\|_{\mathcal{H}} \quad (3)$$

En observant de plus près les deux équations (2) et (3) nous pouvons remarquer la dépendance des fonctions g en les poids d'entrée w_j . C'est bien en cela que réside en partie l'apport de cette méthode, les fonctions d'activation elles-même seront choisies en fonction de la structure des données, de par le fait que le risque empirique soit minimisé selon les poids w_j . La présence d'une fonction d'activation par poids d'entrée w_j est aussi notable.

2.1 L'espace \mathcal{H} est un RKHS

Dans cette sous-partie nous nous intéressons de plus près à la structure mathématique de l'espace \mathcal{H} dont nous rappelons ici la définition :

$$\mathcal{H} = \left\{ g : \mathbb{R} \rightarrow \mathbb{R} \mid g(0) = 0, \text{ où } g \text{ admet une dérivée faible } g', \int_{\mathbb{R}} (g')^2 < \infty \right\},$$

muni du produit scalaire $\langle \tilde{g}, g \rangle_{\mathcal{H}} = \int_{\mathbb{R}} \tilde{g}' g'$.

Cet espace est un *RKHS* muni du noyau de reproduction $k^{(B)}(a, b) = (|a| + |b| + |a - b|)/2 = \min(|a|, |b|) \mathbf{1}_{\{ab > 0\}}$ nommé "noyau Brownien" ("Brownian Kernel" en Anglais, justifiant les deux premières lettres du nom du modèle). L'appartenance de la fonction d'activation g_w à un *RKHS* permettra d'hériter de toutes les bonnes propriétés mathématiques des méthodes à noyaux. Permettant de se ramener, grâce au théorème du représentant à un problème d'optimisation sur un espace de dimension finie. Nous retrouvons bien à présent l'objectif initial du modèle, en faisant appartenir la fonction d'activation à un *RKHS*, il est ainsi possible de mener le problème d'optimisation avec les mêmes facilités que pour un problème de méthode à noyau, tout en conservant d'autre part les avantages d'approximation liés aux réseaux neuronaux.

Une autre particularité liée au *RKHS* \mathcal{H} est l'homogénéité positive du noyau k associé. Comme développé précédemment dans l'exposé, l'homogénéité positive de la fonction d'activation permet en partie d'hériter des propriétés numériques obtenues dans le cas continu. Ici il n'est pas question d'une homogénéité positive de la fonction d'activation mais plutôt du noyau k . Cependant, et c'est aussi en cela que nous tirons profit de l'appartenance de la fonction à un *RKHS*, la propriété de reproduction du *RKHS* nous permet d'écrire :

$$\forall x \in \mathbb{R}, g(x) = \langle g, k(\cdot, x) \rangle$$

De sorte que l'on obtienne :

$$\forall \lambda > 0, g(\lambda x) = \langle g, k(\cdot, \lambda x) \rangle \quad (4)$$

$$= \lambda \langle g, k(\cdot, x) \rangle \quad (5)$$

$$= \lambda g(x) \quad (6)$$

où en (5) nous avons utilisé la 1-homogénéité positive de k et la bilinéarité du produit scalaire.

Nous verrons, plus loin dans l'exposé que l'homogénéité positive permettra aussi de réécrire le problème d'optimisation sur les w_j dans la sphère unité, comme un problème d'optimisation sur de nouveaux coefficients, vivant eux dans \mathbb{R}^d .

2.2 Paradigme des noyaux et interprétation

Avant de se jeter sur un algorithme de minimisation on commence par modifier le problème d'optimisation. En effet ce dernier dépend toujours des variables non paramétriques $g_j \in \mathcal{H}$. Dans le lemme qui suit nous allons utiliser un outil puissant des RKHS : le théorème du représentant. Comme nous l'avons dit précédemment, grâce au paradigme des méthodes à noyaux, nous allons pouvoir nous ramener à un problème d'optimisation en dimension finie.

Lemme 1. (*Paradigme des noyaux dans le cas fini $f \in \mathcal{F}_m$*)
L'équation (3) est équivalente à :

$$\min_{w_1, \dots, w_m \in \mathbb{R}^d, c \in \mathbb{R}, \alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \ell(y_i, (K\alpha)_i + c) + \frac{\lambda}{2} \alpha^\top K \alpha + \frac{\lambda}{2} \frac{1}{m} \sum_{j=1}^m \|w_j\|, \quad (7)$$

où $K = \frac{1}{m} \sum_{j=1}^m K(w_j)$, et $K(w_j) \in \mathbb{R}^{n \times n}$ est la matrice de Gram pour le noyau $k(B)$ et les données projetées $(w_j^\top x_1, \dots, w_j^\top x_n)$, ie :

$$K(w_j)_{i,i'} = \frac{|w_j^\top x_i| + |w_j^\top x_{i'}| - |w_j^\top (x_i - x_{i'})|}{2}.$$

Remarquez que les particules $(w_j)_{j \in [m]}$ ne sont plus contraintes à appartenir à la sphère unité.

Preuve L'objectif est de transformer l'équation (7). On souhaiterait appliquer le théorème des représentants mais la fonction en $\|g_j\|$ doit être strictement croissante. Pour cela on utilise une astuce commune en optimisation sur les m particules :

$$\frac{1}{m} \sum_{j=1}^m \|g_j\|_{\mathcal{H}} = \inf_{\beta \in \mathbb{R}_+^m} \frac{1}{2m} \sum_{j=1}^m (\|g_j\|_{\mathcal{H}}^2 \beta_j + \beta_j)$$

(il suffit d'étudier le deuxième terme comme une fonction en β_j et trouver son minimum).

Fixons $(w_j)_{j \in [m]}$ et $(\beta_j)_{j \in [m]}$ dans l'équation (7), ce qui conduit au problème de minimisation suivant sur les fonctions $(g_j)_{j \in [m]}$:

$$\min_{c \in \mathbb{R}, g_1, \dots, g_m \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, c + \frac{1}{m} \sum_{j=1}^m g_j(w_j^\top x_i)) + \frac{\lambda}{2} \frac{1}{m} \sum_{j=1}^m \|g_j\|_{\mathcal{H}}^2 \beta_j \quad (8)$$

On a bien une fonction strictement croissante (fonction carrée) en $\|g_j\|$. En utilisant le théorème du représentant, nous exprimons chaque $x \mapsto g_j(w_j^\top x)$ comme

$$x \mapsto \sum_{i=1}^n \alpha_i^{(j)} k^{(B)}(w_j^\top x_i, w_j^\top x),$$

ce qui donne

$$\|g_j\|_{\mathcal{H}}^2 = \sum_{i,i'=1}^n \alpha_i^{(j)} \alpha_{i'}^{(j)} k^{(B)}(w_j^\top x_i, w_j^\top x_{i'}).$$

On réécrit la norme ainsi que l'évaluation en g_j grâce à la matrice de Gram $K(w_j)_{i,i'} = k^{(B)}(w_j^\top x_i, w_j^\top x_{i'})$ pour obtenir des formule plus compacte :

$$\begin{aligned} \|g_j\|_{\mathcal{H}}^2 &= (\alpha^{(j)})^\top K(w_j) \alpha^{(j)} \\ g_j(w_j^\top x_i) &= (K(w_j) \alpha^{(j)})_i \end{aligned}$$

Ainsi, on transforme l'Équation (8) en

$$\min_{c \in \mathbb{R}, \alpha^{(1)}, \dots, \alpha^{(m)} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \ell \left(y_i, \frac{1}{m} \sum_{j=1}^m (K(w_j) \alpha^{(j)})_i + c \right) + \frac{\lambda}{2} \frac{1}{m} \sum_{j=1}^m (\alpha^{(j)})^\top K(w_j) \alpha^{(j)} \beta_j.$$

(nous sommes toujours dans le cas où $((w_j)_{j \in [m]})$ et $((\beta_j)_{j \in [m]})$ sont fixés).

Montrons maintenant que la minimisation en $(\alpha^{(j)})_j$ est atteinte pour des vecteurs $\alpha^{(j)}$ égaux à $\beta_j \alpha$ pour un seul vecteur α . Considérons le problème convexe suivant :

$$\min_{\alpha^{(1)}, \dots, \alpha^{(m)} \in \mathbb{R}^d} \frac{1}{2m} \sum_{j=1}^m \frac{(\alpha^{(j)})^\top K^{(w_j)} \alpha^{(j)}}{\beta_j},$$

sous la contrainte

$$\frac{1}{m} \sum_{j=1}^m K^{(w_j)} \alpha^{(j)} = z \quad \text{où } z \in \mathbb{R}^d.$$

Nous définissons le lagrangien :

$$L(\alpha^{(1)}, \dots, \alpha^{(m)}, \alpha) = \frac{1}{2m} \sum_{j=1}^m \frac{(\alpha^{(j)})^\top K^{(w_j)} \alpha^{(j)}}{\beta_j} + \alpha^\top \left(z - \frac{1}{m} \sum_j K^{(w_j)} \alpha^{(j)} \right).$$

En prenant la différentielle de L par rapport à $\alpha^{(j)}$ à l'optimum, nous obtenons

$$\frac{\partial L}{\partial \alpha^{(j)}} = \frac{1}{m} K^{(w_j)} \left(\frac{\alpha^{(j)}}{\beta_j} - \alpha \right) = 0.$$

La différentielle par rapport à α montre qu'à l'optimum, la contrainte est vérifiée, c'est-à-dire $z = \frac{1}{m} \sum_j K^{(w_j)} \alpha^{(j)}$. Nous remarquons que pour $\alpha^{(j)} = \beta_j \alpha$, toutes les équations sont satisfaites, ce qui donne le résultat désiré.

Nous pouvons alors écrire l'équation (5) sous la forme suivante :

$$\min_{w_1, \dots, w_m \in \mathbb{R}^d, c \in \mathbb{R}, \beta \in \mathbb{R}_+^m, \alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \ell(y_i, (K\alpha)_i + c) + \frac{\lambda}{2} \alpha^\top K \alpha + \frac{\lambda}{2m} \sum_{j=1}^m \beta_j,$$

avec les contraintes $\forall j \in [m], w_j \in S^{d-1}$, et $K = \frac{1}{m} \sum_{j=1}^m \beta_j K(w_j)$.

Nous remarquons que $\beta_j K(w_j) = K(\beta_j w_j)$ en raison de l'homogénéité positive du noyau. Nous introduisons donc le changement de variable $\beta_j w_j = \tilde{w}_j$, ce qui donne :

$$\min_{\tilde{w}_1, \dots, \tilde{w}_m \in \mathbb{R}^d, c \in \mathbb{R}, \beta \in \mathbb{R}_+^m, \alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \ell(y_i, (K\alpha)_i + c) + \frac{\lambda}{2} \alpha^\top K \alpha + \frac{\lambda}{2m} \sum_{j=1}^m \|\tilde{w}_j\|,$$

avec $K = \frac{1}{m} \sum_{j=1}^m K(\tilde{w}_j)$, sans contrainte sur la norme de \tilde{w}_j , et ce grâce à l'homogénéité du noyau ! \square

Lemme 2. (Paradigme des noyaux dans le cas infini $f \in \mathcal{F}_\infty$)

L'équation (2) est équivalente à :

$$\min_{\nu \in \mathcal{P}(\mathbb{R}^d), c \in \mathbb{R}, \alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \ell(y_i, (K\alpha)_i + c) + \frac{\lambda}{2} \alpha^\top K \alpha + \frac{\lambda}{2} \int_{\mathbb{R}^d} \|w\| d\nu(w),$$

avec $K = \int_{\mathbb{R}^d} K(w) d\nu(w)$, et $K(w) \in \mathbb{R}^{n \times n}$ est la matrice de Gram pour le noyau $k^{(B)}$ et les données $(w^\top x_1, \dots, w^\top x_n)$.

Notez qu'il y a un changement d'espace, car ν est une distribution de probabilité sur \mathbb{R}^d , tandis que μ était une distribution de probabilité sur S^{d-1} .

La preuve suit les mêmes étapes que celle du lemme précédent en remplaçant $\frac{1}{m} \sum_{j=1}^m$ avec l'intégrale appropriée sur S^{d-1} par rapport à la mesure μ .

Remarque 2. (Interprétation)

Le lemme 1 montre que nous avons un problème de type KRR comme vu en introduction. Le noyau dépend des $(w_j)_{j \in [m]}$ et est donc appris au cours de l'optimisation.

Rappelons également que notre espace de fonctions est inspiré des réseaux neuronaux, impliquant une composante linéaire suivie par une non linéaire. Le paramètre m est équivalent au nombre de neurones de la couche cachée.

3 Calcul de l'estimateur

Dans cette partie on va partir du problème d'optimisation défini par le Lemme 1 pour aller jusqu'à la présentation de l'algorithme de calcul de notre estimateur ainsi que d'arguments en faveur de garanties de convergence.

3.1 Procédure d'optimisation

On se place ici dans le cas de la perte quadratique $\ell(y, y') = \frac{1}{2}(y - y')^2$, un choix classique qui permettra notamment d'obtenir une forme explicite pour α et c .

L'optimisation se déroule en deux temps : une première minimisation par rapport à α et c puis une seconde par rapport aux particules w_1, \dots, w_m .

Lemme 3. (*Optimisation pour des particules fixées*)

Pour des w_1, \dots, w_m fixés, et donc une matrice K fixée, définissons :

$$G(w_1, \dots, w_m) := \min_{\alpha \in \mathbb{R}^n, c \in \mathbb{R}} \frac{1}{2n} \|Y - K\alpha - c\mathbf{1}_n\|_2^2 + \frac{\lambda}{2} \alpha^\top K \alpha.$$

Le problème d'optimisation définissant G est résolu par :

$$\alpha = (\tilde{K} + n\lambda I)^{-1} \tilde{Y}, \quad c = \frac{\mathbf{1}^\top Y}{n} - \frac{\mathbf{1}^\top K \alpha}{n},$$

$$\text{où } \tilde{K} := \Pi K \Pi \text{ et } \tilde{Y} := Y - \frac{\mathbf{1}\mathbf{1}^\top Y}{n}, \quad \text{avec } \Pi = I - \frac{\mathbf{1}\mathbf{1}^\top}{n} \text{ la matrice de centrage.}$$

La valeur de l'objectif est alors :

$$G(w_1, \dots, w_m) = \frac{\lambda}{2} \tilde{Y}^\top (\tilde{K} + \lambda n I)^{-1} \tilde{Y}.$$

Le lemme 3 montre donc qu'on peut obtenir une forme explicite pour α et c .

On souhaite maintenant optimiser sur les particules w_1, \dots, w_m . Le problème de minimisation devient :

$$\min_{w_1, \dots, w_m \in \mathbb{R}^d} G(w_1, \dots, w_m) + \frac{\lambda}{2m} \sum_{j=1}^m \|w_j\|^2$$

On remarque que G est convexe en K et différentiable presque partout. Cependant le terme de pénalité n'est pas différentiable et ne nous permet pas d'utiliser une descente de gradient classique. Pour régler ce problème on emploie la descente de gradient proximal.

Rappel 1. *Descente de gradient proximal*

Soit un problème d'optimisation de la forme :

$$\min_{x \in \mathbb{R}^d} F(x) = G(x) + \Omega(x),$$

où :

- $G : \mathbb{R}^d \rightarrow \mathbb{R}$ est convexe, différentiable,
- $\Omega : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ est convexe mais potentiellement non différentiable.

La descente de gradient proximal combine deux étapes :

1. Une étape de descente en gradient sur la partie différentiable G ,
2. Une régularisation par le proximal de Ω , qui gère la partie non différentiable.

ce qui donne

$$\begin{cases} z_j^{(k)} = w_j^{(k)} - \gamma \nabla G(w_j^{(k)}) \\ w_j^{(k+1)} = \text{prox}_{\lambda \gamma \Omega}(z_j^{(k)}) \end{cases}$$

où :

- $\gamma > 0$ est le pas d'apprentissage,

— $\text{prox}_{\gamma\Omega}(z)$ est l'opérateur proximal défini par :

$$\text{prox}_{\gamma\Omega}(z) = \arg \min_{y \in \mathbb{R}^d} \left\{ \Omega(y) + \frac{1}{2\gamma} \|y - z\|^2 \right\}.$$

Étant donné une itération x^k , la mise à jour s'écrit :

$$x^{k+1} = \text{prox}_{\gamma\Omega}(x^k - \gamma \nabla G(x^k)),$$

On a donc besoin de calculer le gradient de G ainsi que l'opérateur proximal. Les formules sont obtenues dans les 2 lemmes suivants.

Lemme 4. (Gradient de G)

Soit $j \in [m]$, alors

$$\frac{\partial G}{\partial w_j} = \frac{\lambda}{4} \frac{1}{m} \sum_{i, i'}^n z_i z_{i'} \text{sign}(w_j^\top (x_i - x_{i'})) (x_i - x_{i'}),$$

$$\text{où } z = (\tilde{K} + n\lambda I)^{-1} Y.$$

La preuve est laissée au lecteur.

Lemme 5. (Opérateur proximal)

Pour la pénalité $\Omega = \frac{1}{2m} \sum_{j=1}^m \|w_j\|$, alors $\text{prox}_{\lambda\gamma\Omega}(W)_j = (1 - \frac{\lambda\gamma}{2m} \frac{1}{\|w_j\|})_+ w_j$

La preuve est laissée au lecteur.

Nous avons désormais tous les outils nécessaires pour présenter l'algorithme d'optimisation pour la perte quadratique.

Algorithm 1 Optimisation pour $w_1, \dots, w_m, c, \alpha$

```

1: Données :  $X, Y, m, \lambda, \gamma, \Omega_{\text{weights}}$ 
2: Résultat :  $w_1, \dots, w_m, c, \alpha$ 
3:  $W = (w_1, \dots, w_m) \in \mathbb{R}^{d \times m} \leftarrow (\mathcal{N}(0, 1/d))^{d \times m}$ 
4: for  $i \in [n_{\text{iter}}]$  do
5:   Calculer  $K$  {Matrice de noyau définie par  $W$ }
6:    $\alpha \leftarrow (\tilde{K} + n\lambda I)^{-1} \tilde{Y}$ 
7:    $c \leftarrow \frac{1^\top Y}{n} - \frac{1^\top}{n} K \alpha$ 
8:   Calculer  $\frac{\partial G}{\partial W}$ 
9:    $\gamma \leftarrow \gamma \times 1.5$ 
10:  while  $G(\text{prox}_{\lambda\gamma\Omega}(W - \gamma \frac{\partial G}{\partial W})) > G(W) - \gamma \frac{\partial G}{\partial W} \cdot G_\gamma(W) + \frac{\gamma}{2} \|G_\gamma(W)\|_2^2$  do
11:     $\gamma \leftarrow \gamma/2$ 
12:  end while
13:   $W \leftarrow \text{prox}_{\lambda\gamma\Omega}(W - \gamma \frac{\partial G}{\partial W})$ 
14: end for

```

Pour sélectionner le pas d'apprentissage γ on utilise la "backtracking condition" définie au niveau de la boucle **while** qui divise γ par 2 si elle n'est pas satisfaite. On pourra voir [Bec17] pour l'utilisation de cette méthode.

3.2 Robustesse de l'optimisation

Dans cette partie on tente d'aborder quelques heuristiques sur les garanties d'optimisation de notre algorithme.

On souhaite se fonder sur les travaux de [BC21] et [CB18]. Ces derniers ont montré que sous certaines hypothèses (de convexité par rapport à la distribution de probabilité et d'homogénéité d'une certaine fonction Ψ), lorsque le nombre de particules m devient infiniment grand et que la taille du pas d'apprentissage tend vers zéro, l'optimisation par descente de gradient converge vers le minimum global du problème associé à un nombre infini de particules.

A partir du lemme 2 et suivant des étapes similaires au cas des m particules, on peut montrer que le problème dans le cas continu peut s'écrire comme minimisant la quantité F sur $\mathcal{P}(\mathbb{R}^d)$ définie par

$$F(\nu) := \mathcal{Q}\left(\int_{\mathbb{R}^d} \Psi(w) d\nu(w)\right),$$

où

$$\Psi : \mathbb{R}^d \rightarrow \mathbb{R}^{n \times n} \times \mathbb{R}, \quad \Psi(w) = (K(w), \|w\|)$$

Il se trouve que \mathcal{Q} est convexe selon $\nu \in \mathcal{P}(\mathbb{R}^d)$ et on peut remarquer que Ψ est positivement 1-homogène. En effet la condition

$$\forall w \in \mathbb{R}^d, \forall \kappa > 0, \Psi(\kappa w) = \kappa \Psi(w)$$

est vérifiée grâce à l'homogénéité positive du noyau!

On ne peut pas appliquer directement les résultats utilisés par [BC21] et [CB18] car Ψ n'est pas différentiable en zéro, néanmoins nos deux paradigmes restent similaires et les hypothèses essentielles valides.

4 Analyse statistique

4.1 Résultat principal

Dans cette section, on se concentre à présent sur l'analyse statistique de l'estimateur proposé. Pour rappel, il a été défini en (1) comme :

$$\hat{f}_\lambda := \arg \min_{f \in \mathcal{F}} \hat{R}(f) + \lambda \Omega(f)$$

Nous nous appuyons toujours sur la démarche de F. Bach [FB24], avec pour objectif de déterminer une borne supérieure pour $R(\hat{f}_\lambda)$.

Sans plus attendre, énonçons le résultat principal découvert dans cet article :

Théorème 1 (Borne avec grande probabilité sur le risque généralisé). *Définissons l'estimateur suivant, qui n'est autre que (1) dans le cas $\mathcal{F} := \mathcal{F}_\infty$:*

$$\hat{f}_\lambda := \arg \min_{f \in \mathcal{F}_\infty} \hat{R}(f) + \lambda \Omega(f).$$

Faisons les hypothèses suivantes :

1. **Modèle bien spécifié** : Le minimiseur du risque généralisé existe, et on le note :

$$f^* := \arg \min_{f \in \mathcal{F}_\infty, \Omega(f) < +\infty} R(f)$$

2. **Convexité de la perte** : Pour tous $(x, y) \in \mathcal{X} \times \mathcal{Y}$, l'application $f \in \mathcal{F}_\infty \mapsto \ell(y, f(x))$ est convexe.
3. **Lipschitzianité** : La perte ℓ est L -Lipschitz en son second argument (borné). Plus précisément : pour tous $y \in \mathcal{Y}$ et $a \in \{f(x) \mid x \in \mathcal{X}, f \in \mathcal{F}_\infty, \Omega(f) \leq 2\Omega(f^*)\}$, l'application $a \mapsto \ell(y, a)$ est L -Lipschitz.
4. **Loi des données** : $(x_i, y_i)_{i \in [n]}$ est un échantillon de variables aléatoires i.i.d. de même loi que (X, Y) . On suppose en outre que $1 + \sqrt{\|X\|_*}$ sous-gaussienne, de variance proxy σ^2 ($\|X\|_*$ désigne ici la norme duale de X , associée à la norme $\|\cdot\|$).

Alors pour tout $\delta \in (0, 1)$, en posant $\lambda := 12L \left(\frac{1}{\sqrt{n}} + G_n \right) + 288 \frac{L\sigma}{\sqrt{n}} \sqrt{\log \frac{1}{\delta}}$:

$$\mathbb{P} \left[R(\hat{f}_\lambda) \leq R(f^*) + 24\Omega(f^*)L \left(\frac{1}{\sqrt{n}} + G_n + 24\sigma \sqrt{\frac{\log \frac{1}{\delta}}{n}} \right) \right] \geq 1 - \delta$$

Ce théorème nous dit, sous certaines conditions, la chose suivante : avec grande probabilité, le risque généralisé de l'estimateur converge vers le risque de l'optimal, à une vitesse explicite qui dépend de G_n . Nous discuterons de cette dépendance dans un instant, et nous donnerons des outils pour contrôler ce terme dans la prochaine sous-section.

Remarquons d'abord que les hypothèses nécessaires à l'application de ce théorème ne sont pas toutes extrêmement restrictives. Des fonctions de pertes classiques comme les pertes logistique ou quadratique satisfont en effet les points 2. et 3.. Le point 4. nécessite la sous-gaussianité de la norme et le caractère *i.i.d.* des données. C'est un cadre assez spécifique mais classique dans la littérature. On exclut ainsi par exemple les distributions à queue lourde, les anomalies (présence de données provenant d'une autre distribution), et les dépendances de données entre elles (pas de séries temporelles, donc). Un exemple de données acceptées sont les variables *i.i.d.* qui sont bornées ou à composantes gaussiennes.

Rappelons à ce propos, pour être complet, la définition des variables sous-gaussiennes :

Définition 8. (*Variable aléatoire sous-gaussienne*) Soit Z une variable aléatoire réelle. On dit que Z est sous-gaussienne, de variance proxy σ^2 , si et seulement si :

$$\forall t > 0, \max(\mathbb{P}(Z \geq t), \mathbb{P}(Z \leq -t)) \leq e^{-\frac{t^2}{2\sigma^2}}.$$

D'autre part, le choix de λ proposé ne dépend que de quantités connues. Notons qu'il peut être choisi, si ce n'est par un calcul, par cross-validation. Un autre point positif de cette approche est qu'elle est moins dépendante de la méthode de génération des données, alors que la littérature semble plus restrictive à ce sujet.

De plus, la borne obtenue dépend explicitement de la dimension des données, via la complexité gaussienne G_n . Cette notion est très utile dans la théorie de l'apprentissage statistique pour obtenir des bornes. Intuitivement, elle mesure à quel point un ensemble de fonctions peut être exploré par des combinaisons linéaires aléatoires pondérées par des coefficients gaussiens. Elle quantifie donc la richesse de cet ensemble, ce qui est crucial pour contrôler le risque d'erreur :

Définition 9 (Complexité gaussienne). La complexité gaussienne d'un ensemble de fonctions F est définie comme :

$$G_n(F) := \mathbb{E}_{\epsilon, D_n} \left[\sup_{f \in F} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right],$$

où $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, , et $D_n := (x_1, \dots, x_n)$ est un échantillon de variables aléatoires *i.i.d.* de même loi que X .

On notera par la suite $G_n := G_n(\{f \mid f(\cdot) = g(w^\top \cdot), \|g\|_{\mathcal{H}} \leq 1, w \in S^{d-1}\})$, où S^{d-1} est la sphère unitaire de dimension $d - 1$ dans l'espace euclidien de dimension d .

Les auteurs semblent penser qu'une autre preuve, reposant sur l'inégalité de McDiarmid, est possible. La constante qui en découlerait serait en réalité meilleure. Donnons tout de même les idées de la preuve originelle présentée dans [FB24] :

Preuve [du Théorème 1]

Cette preuve utilise les idées de la Proposition 4.7 de [Bac24].

Supposons les hypothèses 1, 2, 3 et 4 vérifiées.

Soit f_λ^* un minimiseur de la fonctionnelle $R_\lambda := R + \lambda\Omega$ sur \mathcal{F}_∞ . C'est-à-dire :

$$f_\lambda^* \in \arg \min_{f \in \mathcal{F}_\infty} R(f) + \lambda\Omega(f)$$

Considérons de plus les ensembles suivants :

$$C_\tau := \{f \in \mathcal{F}_\infty \mid R_\lambda(f) - R_\lambda(f_\lambda^*) \leq \tau\}$$

$$B_\tau := \{f \in \mathcal{F}_\infty \mid \Omega(f) \leq \Omega(f^*) + \tau/\lambda\}$$

pour un $\tau > 0$ qui sera fixé plus tard. Remarquons que, par l'hypothèse 2, ℓ est convexe, et donc $R(f) := \mathbb{E}[\ell(Y, f(X))]$ aussi. Comme somme d'une fonction convexe et d'une constante, $R_\lambda(f)$ est elle aussi convexe. Il en découle que C_τ est convexe, en tant que sous-niveau d'une fonction convexe.

- Nous commençons par démontrer que $C_\tau \subseteq B_\tau$.

Soit $f \in C_\tau$, alors :

$$R_\lambda(f) \leq R_\lambda(f_\lambda^*) + \tau \leq R_\lambda(f^*) + \tau \leq R(f) + \lambda\Omega(f^*) + \tau,$$

par optimalité de f^* , et définition de R_λ . En réécrivant ce que cela signifie d'appartenir à C_τ , on en déduit que :

$$R(f) + \lambda\Omega(f) \leq R(f^*) + \lambda\Omega(f^*) + \tau$$

En divisant de part et d'autre de l'égalité par λ , on obtient exactement que $f \in B_\tau$.

On a donc bien l'inclusion $C_\tau \subseteq B_\tau$.

- Posons maintenant $\tau := \lambda\Omega(f^*)$, où λ reste à choisir. Démontrons alors que $\hat{f}_\lambda \in C_\tau$ avec probabilité au moins $1 - \delta$.

Supposons pour cela que $\hat{f}_\lambda \notin C_\tau$.

Observons qu'on a au contraire $f_\lambda^* \in C_\tau$, car $\tau \geq 0$. Puisque C_τ est convexe, il existe un \tilde{f} dans le segment $[\hat{f}_\lambda, f_\lambda^*]$ qui soit précisément sur la frontière de C_τ . C'est-à-dire que

$$R_\lambda(\tilde{f}) = R_\lambda(f_\lambda^*) + \tau.$$

Notons $\hat{R}_\lambda := \hat{R} + \lambda\Omega$, la version empirique de R_λ . Puisque \hat{R} , pour la même raison que R , est lui aussi convexe, on a :

$$\hat{R}_\lambda(\tilde{f}) \leq \max(\hat{R}_\lambda(\hat{f}_\lambda), \hat{R}_\lambda(f_\lambda^*)) = \hat{R}_\lambda(f_\lambda^*) \quad \text{par optimalité de } \hat{f}_\lambda.$$

Mais alors :

$$\hat{R}(\hat{f}_\lambda) - \hat{R}(\tilde{f}) - R(\hat{f}_\lambda) + R(\tilde{f}) = \hat{R}_\lambda(\hat{f}_\lambda) - \hat{R}_\lambda(\tilde{f}) - R_\lambda(\hat{f}_\lambda) + R_\lambda(\tilde{f}) \geq -R_\lambda(\hat{f}_\lambda) + R_\lambda(\tilde{f}) = \tau \quad (\clubsuit)$$

Nous utilisons à présents les lemmes 6 et 7 pour majorer la quantité de gauche. Ils sont énoncés dans la sous-section suivante, focalisée sur la complexité gaussienne.

Remarquons que $\Omega(\tilde{f}) \leq 2\Omega(f^*)$ et $\Omega(f_\lambda^*) \leq 2\Omega(f^*)$. En utilisant les deux lemmes, pour $\delta \in (0, 1)$, avec probabilité au moins $1 - \delta$, on a, pour toute $f \in \mathcal{F}_\infty$ telle que $\Omega(f) \leq 2\Omega(f^*)$:

$$\begin{aligned} \hat{R}(f_\lambda^*) - \hat{R}(f) - R(f_\lambda^*) + R(f) &\leq \mathbb{E}_{D_n} \left[\sup_{f \in \mathcal{F}_\infty, \Omega(f) \leq 2\Omega(f^*)} (\hat{R}(f) - R(f)) + \sup_{f \in \mathcal{F}_\infty, \Omega(f) \leq 2\Omega(f^*)} (R(f) - \hat{R}(f)) \right] \\ &\quad + \Omega(f^*) \frac{96\sqrt{2e}L\sigma}{\sqrt{n}} \sqrt{\log \frac{1}{\delta}} \\ &\leq 12\Omega(f^*)L \left(\frac{1}{\sqrt{n}} + G_n \right) + \Omega(f^*) \frac{96\sqrt{2e}L\sigma}{\sqrt{n}} \sqrt{\log \frac{1}{\delta}} \end{aligned}$$

Choisissons enfin λ (et donc τ) tel que :

$$\tau = \lambda\Omega(f^*) > 12\Omega(f^*)L \left(\sqrt{\frac{1}{n}} + G_n \right) + \Omega(f^*) \frac{96\sqrt{2e}L\sigma}{\sqrt{n}} \sqrt{\log \frac{1}{\delta}}$$

Il y a alors contradiction avec l'équation $(\clubsuit)!$ Par conséquent, pour ce λ , $\mathbb{P}(\hat{f}_\lambda \in C_\tau) \geq 1 - \delta$.

- La dernière étape de cette preuve consiste en des réécritures du résultat que l'on vient d'obtenir.

En remplaçant l'événement $\{\hat{f}_\lambda \in C_\tau\}$ par sa définition :

$$\mathbb{P} \left[R_\lambda(\hat{f}_\lambda) \leq R_\lambda(f^*) + \lambda \Omega(f^*) \right] \geq 1 - \delta$$

Puis en utilisant la définition de R_λ et la positivité de Ω ,

$$\{R_\lambda(\hat{f}_\lambda) \leq R_\lambda(f^*) + \lambda \Omega(f^*)\} \subseteq \{R(\hat{f}_\lambda) \leq R(f^*) + 2\lambda \Omega(f^*)\}$$

$$\text{et donc : } \mathbb{P} \left[R(\hat{f}_\lambda) \leq R(f^*) + 2\lambda \Omega(f^*) \right] \geq 1 - \delta$$

Finalement, le choix $\lambda = 12L \left(\frac{1}{\sqrt{n}} + G_n \right) + 288 \frac{L\sigma}{\sqrt{n}} \sqrt{\log \frac{1}{\delta}}$ nous donne :

$$\mathbb{P} \left[R(\hat{f}_\lambda) \leq R(f^*) + \Omega(f^*) \left(24L \left(\frac{1}{\sqrt{n}} + G_n \right) + 576L\sigma \sqrt{\frac{\log \frac{1}{\delta}}{n}} \right) \right] \geq 1 - \delta$$

On peut démontrer que ce résultat est en fait équivalent à celui du Théorème 1, en utilisant le fait suivant : $24 \times 24 = 576$. □

4.2 Raffinement : bornes sur la complexité gaussienne

Donnons ici les deux lemmes utilisés pour la majoration de la quantité de gauche dans l'équation (♣).

Lemme 6 (Utilisation de la complexité gaussienne). *Soit $D > 0$, et l'échantillon de données $D_n = (x_i, y_i)_{i \in [n]}$ constitué d'échantillons i.i.d. de la variable aléatoire $(X, Y) \in \mathcal{X} \times \mathcal{Y}$. Supposons que la perte ℓ soit L -Lipschitzienne en son second argument (borné), c'est-à-dire, pour tout $y \in \mathcal{Y}$ et $a \in \{f(x) \mid x \in \mathcal{X}, f \in \mathcal{F}_\infty, \Omega(f) \leq D\}$, l'application $a \mapsto \ell(y, a)$ est L -Lipschitzienne. Alors, nous avons :*

$$\mathbb{E}_{D_n} \left[\sup_{f \in \mathcal{F}_\infty, \Omega(f) \leq D} (R_b(f) - R(f)) + \sup_{f \in \mathcal{F}_\infty, \Omega(f) \leq D} (R(f) - \hat{R}(f)) \right] \leq 6DL \left(\frac{1}{\sqrt{n}} + G_n \right)$$

Lemme 7 (Utilisation de l'inégalité de McDiarmid). *Soit $D > 0$ et $\delta \in (0, 1)$. Supposons que $1 + \sqrt{\|X\|_*}$ soit sous-gaussienne avec une variance proxy σ^2 et que la perte ℓ soit L -Lipschitzienne en son second argument (borné), c'est-à-dire, pour tout $y \in \mathcal{Y}$ et $a \in \{f(x) \mid x \in \mathcal{X}, f \in \mathcal{F}_\infty, \Omega(f) \leq D\}$, l'application $a \mapsto \ell(y, a)$ est L -Lipschitzienne. Alors, avec une probabilité supérieure à $1 - \delta$,*

$$\begin{aligned} & \sup_{f \in \mathcal{F}_\infty, \Omega(f) \leq D} (\hat{R}(f) - R(f)) + \sup_{f \in \mathcal{F}_\infty, \Omega(f) \leq D} (R(f) - \hat{R}(f)) \\ & \leq \mathbb{E}_{D_n} \left[\sup_{f \in \mathcal{F}_\infty, \Omega(f) \leq D} (\hat{R}(f) - R(f)) + \sup_{f \in \mathcal{F}_\infty, \Omega(f) \leq D} (R(f) - \hat{R}(f)) \right] + 48\sqrt{2e}LD\sigma \sqrt{\frac{\log \frac{1}{\delta}}{n}}. \end{aligned}$$

Enfin, [FB24] démontre des bornes supérieures sur la complexité gaussienne G_n , qui permettent d'améliorer encore celle obtenue au Théorème 1. La première dépend de la dimension d , de l'espace $\mathbb{R}^d \supseteq \mathcal{X}$:

Proposition 2 (Borne dépendante de la dimension). *On a :*

$$G_n \leq 8\sqrt{\frac{d}{n}} \cdot \sqrt{\log(n+1)} \cdot \sqrt{\mathbb{E}_X \|X\|_*},$$

où $\|\cdot\|_*$ est la norme duale de $\|\cdot\|$.

La seconde majoration de G_n ne dépend, elle, pas toujours de la dimension :

Proposition 3 (Borne indépendante de la dimension). *Si S^{d-1} est la sphère unitaire ℓ_1 ou ℓ_2 , alors :*

$$G_n \leq \frac{3}{n^{1/6}} \left((\log 2d)^{1/4} \mathbf{1}_{*= \infty} + \mathbf{1}_{*=2} \right) \left(\mathbb{E}_{D_n} \left[\max_{i \in [n]} (\|X_i\|_*)^2 \right] \right)^{1/4}$$

Notons que par exemple si l'on choisit la norme 2, le terme en d disparaît dans la borne. Remarquons que même si ce n'était pas le cas, la dépendance en d est logarithmique, l'impact de la dimension est donc moindre comparé à la Proposition 2.

En combinant les bornes sur G_n des Propositions 2 et Proposition 3, on obtient finalement :

$$G_n \leq \min \left(\frac{3}{n^{1/6}} \left((\log 2d)^{1/4} \mathbf{1}_{*= \infty} + \mathbf{1}_{*=2} \right) \left(\mathbb{E}_{D_n} \left(\max_{i \in [n]} (\|X_i\|_*)^2 \right) \right)^{1/4}, \right. \\ \left. 8\sqrt{\frac{d}{n}} \sqrt{\log(n+1)} \sqrt{\mathbb{E}_X \|X\|_*} \right)$$

ce qui permet d'affiner la borne du Théorème 1.

5 Conclusion

Pour résumer, la modèle BKerNN développé par Follain et Bach réunit bien les qualités des réseaux de neurones et des méthodes à noyaux. La définition même du modèle permet d'introduire une dépendance des fonctions d'activation du réseau aux données, il a bien été possible d'obtenir une borne sur le risque indépendante de la dimension, le problème d'optimisation a bien pu être résolu en dimension finie grâce à la structure de *RKHS* de \mathcal{H} . L'ajout d'une pénalité sur la norme fonctionnelle de la fonction f permet aussi d'introduire de la sparsité au modèle et d'obtenir par la suite une sélection de caractéristiques faisant office d'"apprentissage de caractéristiques" et de faire face ainsi au fléau de la dimension. Les expériences de Follain et Bach dans [FB24] §5.1 ont aussi montré l'avantage pratique de l'homogénéité positive du noyau, permettant l'obtention d'un risque quadratique moyen plus bas qu'avec l'utilisation d'autres noyaux non homogènes. On peut cependant se poser la question de la généralisation des résultats de convergence statistique à des données non i.i.d, conditions rarement obtenues en pratique, étudier plusieurs exemples de fonction d'activation g appartenant à l'espace \mathcal{H} , afin de pouvoir travailler par la suite sur l'aspect computationnel du modèle.

Références

- [Bac24] Francis BACH. *Learning theory from first principles*. en. London, England : MIT Press, déc. 2024.
- [BC21] Francis R. BACH et Lenaïc CHIZAT. "Gradient Descent on Infinitely Wide Neural Networks : Global Convergence and Generalization". In : *CoRR* abs/2110.08084 (2021). arXiv : 2110.08084. URL : <https://arxiv.org/abs/2110.08084>.
- [Bec17] Amir BECK. *First-Order Methods in Optimization*. Philadelphia, PA : Society for Industrial et Applied Mathematics, 2017. DOI : 10.1137/1.9781611974997. eprint : <https://epubs.siam.org/doi/pdf/10.1137/1.9781611974997>. URL : <https://epubs.siam.org/doi/abs/10.1137/1.9781611974997>.
- [CB18] Lenaïc CHIZAT et Francis BACH. *On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport*. 2018. arXiv : 1805.09545 [math.OC]. URL : <https://arxiv.org/abs/1805.09545>.
- [FB24] Bertille FOLLAIN et Francis BACH. *Enhanced Feature Learning via Regularisation : Integrating Neural Networks and Kernel Methods*. 2024. arXiv : 2407.17280 [stat.ML]. URL : <https://arxiv.org/abs/2407.17280>.