Variance function estimation in regression and its applications

Ahmed Zaoui

Université Gustave Eiffel

LAMA LABORATOIRE D'ANALYSE ET DE MATHÉMATIQUES APPLIQUÉES – UMR8050

École Doctorale Mathématiques et STIC

THÈSE

Présentée pour l'obtention du grade de DOCTEUR
de L'Université Gustave Eiffel

Spécialité

Mathématiques Appliquées

par

AHMED ZAOUI

Estimation de la fonction de variance en régression et ses applications

Soutenue le 08/12/2022 devant le jury composé de

M. CHRISTOPHE DENIS	UNIVERSITÉ GUSTAVE EIFFEL	CO-DIRECTEUR
M. ROMUALD ELIE	UNIVERSITÉ GUSTAVE EIFFEL	EXAMINATEUR
M. MOHAMED HEBIRI	UNIVERSITÉ GUSTAVE EIFFEL	DIRECTEUR
MME CLAIRE LACOUR	UNIVERSITÉ GUSTAVE EIFFEL	EXAMINATRICE
MME KATIA MEZIANI	UNIVERSITÉ PARIS DAUPHINE	EXAMINATRICE
M. BERTRAND MICHEL	ECOLE CENTRALE DE NANTES	RAPPORTEUR
M. CHRISTOPHE POUET	ECOLE CENTRALE DE MARSEILLE	RAPPORTEUR
MME ANGELINA ROCHE	UNIVERSITÉ PARIS DAUPHINE	EXAMINATRICE

Remerciements

Tout d'abord, je tiens à exprimer ma plus profonde gratitude à mes directeurs de thèse, Christophe Denis et Mohamed Hebiri. Avec leur gentillesse et leurs expériences, ils m'ont bien guidé et apporté leurs soutiens tout au long de la préparation de la thèse. Je les remercie vivement pour tout le temps qu'ils m'ont consacré, pour leurs encouragements et leurs conseils avisés. Je les remercie encore une fois pour leur patience infinie nonobstant mes trop nombreux moments de doute et passages à vide. Je remercie tout particulièrement Romuald ELIE, mon premier directeur principal, de m'avoir donné l'opportunité de faire une thèse. Je ne serais pas devant vous aujourd'hui sans sa proposition d'un sujet de thèse à la dernière minute et sa confiance qui m'a apporté.

Je suis très reconnaissant aux professeurs Bertrand Michel et Christophe Pouet qui ont accepté gentiment de rapporter cette thèse. Je tiens à remercier également Claire Lacour, Katia Meziani, Angelina Roche et Romuald Elie pour avoir accepté d'examiner mon travail. Leur présence à la soutenance de cette thèse m'honore énormément.

Je suis très reconnaissant envers le Labex MME-DII de m'avoir accordé une bourse pour mes deux années de Master à l'Institut Galilée à l'Université Sorbonne Paris Nord. J'en profite pour remercier sincèrement Monsieur Mohamed Ben Alaya, Monsieur Julien Barral et Monsieur Thomas Duyckaerts pour cette belle opportunité. Je tiens également à remercier chaleureusement tous mes professeurs.

Je remercie l'ensemble des membres du Laboratoire LAMA, et en particulier l'équipe de Probabilités et Statistiques, de m'avoir accueilli pendant ma thèse. Merci notamment à Audrey Patout, Sylvie Cach, Mariam Sidibé et Ketty Cimonard pour leurs disponibilités et leurs réactivités durant ces quatres années de thèse.

Je voudrais adresser mes chaleureux remerciements à tous les doctorants, post-doctorants et ATER du Laboratoire. Je remercie Benjamin pour son aide admistrative et son écoute. Aux anciens : Arafat, Josué et Quentin ; je vous remercie de m'avoir si bien accueilli au LAMA. Je vous remercie pour les discussions et tous les bons moments partagés qui ont été essentiels pour moi.

Merci à tous mes amis avec qui j'ai partagé énormément de choses. Merci Tatyana pour tes conseils et ton soutien infini. Marie et Nina, vous êtes deux femmes gentilles et adorables, merci pour toutes nos discussions enrichissantes.

Enfin, je remercie mes parents, Mohamed et Mounira, pour leurs encouragements, leurs bénédictions et leur soutien. Je remercie également mon frère Amir, mes soeurs : Emna et Imen, mon beau-frère Farid, ma petite princesse Lamis et mon neveu Bedis.

Production scientifique liée à la thèse

Articles publiés

- C. Denis, M. Hebiri, and A. Zaoui. Regression with reject option and application to kNN. NeurIPS, 2020 [24].
- A. Zaoui. Variance function estimation in regression model via aggregation procedures. 2022 [105]. Accept on Journal of Nonparametric Statistics.

Article Soumis

- C. Denis, M. Hebiri, and A. Zaoui. Prediction interval with fixed expected length in the Gaussian heteroscedastic regression. Preprint, 2022 [25].

Contents

Abstract

This thesis deals with the problem of estimating the variance function in the heteroscedastic regression model.

The first part is devoted to the estimation of the variance function through classical aggregation procedures. More precisely, we are interested in the estimation of the variance function by model selection (MS) and convex aggregation (C). The goal of the procedure MS is to select the best estimator among a set of predictors, that of C consists in selecting the best convex combination among the predictors. The selected predictors are then called MS and C-estimators respectively. The construction of the MS-estimator and C-estimator is based on a two-step procedure. In the first step, from a first sample, we construct estimators of the variance function by the residual-based method. In the second step, we aggregate them using a second sample. We establish the consistency of MS-estimator and C-estimator with respect to the L^2-risk and illustrate its numerical performances on simulations.

The second part is devoted to the applications where the variance function plays a crucial role. It is split into two chapters. The Chapter 3 is devoted to the study of the regression problem where we are allowed to abstain from predicting. We focus on the case where the rejection rate is fixed and derive the optimal rule which relies on thresholding the conditional variance function. We provide a semi-supervised estimation procedure of the optimal rule involving two datasets: a first labeled dataset is used to estimate both regression function and conditional variance function while a second unlabeled dataset is exploited to calibrate the desired rejection rate. The resulting predictor with reject option is shown to be almost as good as the optimal predictor with reject option both in terms of risk and rejection rate. Furthermore, we apply our approach with the k-nearest neighbor algorithm (k-NN) and establish rates of convergence for the resulting kNN predictor. Finally, a numerical study is performed to illustrate the interest in using the proposed procedure.

The Chapter 4 focuses on the construction of a prediction interval in the gaussian heteroscedastic model. The optimal interval is based on thresholding the conditional density of the Gaussian distribution. The construction of this interval is based on a semi-supervised estimation procedure in two steps (as in Chapter 3). In the first step, from the first sample of labeled dataset, we estimate the conditional density using the plug-in rule. More specifically, the conditional density estimator is based on the estimators of the regression function and the variance function. In the second step, we calibrate the threshold using a second sample of unlabeled dataset. The constructed confidence interval offers guarantees of consistency and good numerical performance. Moreover, we establish convergence rates for the confidence interval estimator based on the kNN algorithm. In addition, we run a numerical comparison to conformal prediction approaches that suggests that our method is more stable and can be preferred in real applications with small samples.

Key words: Regression, heteroscedastic model, variance function, agregation, reject option, prediction interval, plug-in rule.

Résumé

Cette thèse porte sur le problème d'estimation de la fonction de variance dans le modèle de régression hétéroscédastique.

La première partie est consacrée à l'estimation de la fonction de variance à travers les procédures d'agrégation classiques. Plus précisément, nous nous intéressons à l'estimation de la fonction de variance par agrégation de type sélection modèle (`MS`) et convexe (`C`). Le but de la procédure `MS` est de sélectionner le meilleur estimateur parmi un ensemble de prédicteurs, celui de `C` consiste à sélectionner la meilleure combinaison convexe parmi les prédicteurs. Les prédicteurs sélectionnés sont alors appelés `MS` et `C`-estimateurs respectivement. La construction de `MS` et `C`-estimateurs repose sur une procédure en deux étapes. Dans une première étape, à partir d'un premier échantillon, nous construisons des estimateurs de la fonction de variance par la méthode basée sur les erreurs résiduelles. Dans une deuxième étape, nous les agrégeons à l'aide d'un deuxième échantillon. Nous établissons la consistance de `MS`-estimateur et de `C`-estimateur vis-à-vis du risque L^2 et illustrons ses performances numériques sur simulations.

La deuxième partie est dédiée à des applications où la fonction de variance joue un rôle crucial. Elle est divisée en deux chapitres. Le chapitre 3 est consacré à l'étude du problème de la régression où l'on est autorisé à s'abstenir de prédire. Nous nous concentrons sur le cas où le taux de rejet est fixe et dérivons la règle optimale qui repose sur le seuillage de la fonction de variance conditionnelle. Nous fournissons une procédure d'estimation semi-supervisée de la règle optimale impliquant deux ensembles de données : un premier ensemble de données étiquetées est utilisé pour estimer à la fois la fonction de régression et la fonction de variance conditionnelle ; un deuxième ensemble de données non étiquetées est exploité pour calibrer le taux de rejet à celui souhaité. Le prédicteur avec option rejet résultant s'avère presque aussi bon que le prédicteur optimal avec l'option rejet, à la fois en termes de risque et de taux de rejet. En outre, nous appliquons notre approche avec l'algorithme des K-plus proches voisins (K-PPV) et établissons des vitesses de convergence pour le prédicteur K-PPV avec option rejet résultant. Enfin, une étude numérique est réalisée pour illustrer l'intérêt de l'utilisation de la procédure proposée.

Le chapitre 4 s'intéresse à la construction d'un intervalle de prédiction dans le modèle hétéroscédastique gaussien. L'intervalle optimal repose sur le seuillage de la densité conditionnelle de la loi gaussienne. La construction de cet intervalle repose sur une procédure d'estimation semi-supervisée en deux étapes (comme dans le Chapitre 3). Dans une première étape, à partir d'un premier échantillon de données étiquetées, nous estimons la densité conditionnelle en utilisant la règle plug-in. Plus précisément, l'estimateur de la densité conditionnelle est basé sur les estimateurs de la fonction de régression et de la fonction de variance. Dans une deuxième étape, nous calibrons le seuil à l'aide d'un deuxième échantillon de données non étiquetées. L'intervalle de prédiction construit offre des garanties de consistance et de bonnes performances numérique. De plus, nous établissons des vitesses de convergence pour l'intervalle de prédiction basé sur l'algorithme des K-PPV. Par ailleurs, une comparaison numérique à l'approche basée sur les *conformal predictors* suggère que notre méthode est plus stable et plus adaptée à des applications réelles où le nombre d'observations est petit.

Mots clefs: Régression, modèle hétéroscédastique, fonction de variance, agrégation, option rejet, intervalle de prédiction, règle plug-in.

Quelques notations

- $\|\cdot\|$: la distance euclidienne sur $\mathbb{R}^d$.
- $\|\cdot\|_n^2 := \frac{1}{n}\sum_{i=1}^n (\cdot)^2$.
- $a_n \lesssim b_n$: $a_n \leq cb_n$ où c est une constante indépendante de n.
- $a_n \asymp b_n$: $a_n \lesssim b_n$ et $b_n \lesssim a_n$.
- $a_n \propto b_n$: $a_n = cb_n$ où c est une constante indépendante de n.
- $\lceil\cdot\rceil$: $\lceil a\rceil$ est le plus grand entier $\leq a$.
- $[p] := \{1, \ldots, p\}$ pour tout entier $p \geq 2$.
- $\Lambda^p := \{\lambda \in \mathbb{R}^p : \lambda_j \geq 0,\ \sum_{j=1}^p \lambda_j = 1\}$.
- $\mathbb{P}_X$: la distribution marginale de $X \in \mathbb{R}^d$.
- ℓ_p-norme sup : $(\sup_{x\in\mathcal{C}} |f(x)|)^p$ où $f : \mathcal{C} \subset \mathbb{R}^d \to \mathbb{R}$.
- $\lesssim_{\log(n)}$ signifie que l'inégalité est valable jusqu'à des constantes et des facteurs logarithmiques.

Chapitre 1

Introduction

La méthode de régression a été largement utilisée en statistiques. L'objectif général de la régression est d'expliquer une variable aléatoire Y, dite réponse, en fonction d'une variable explicative X, qui peut être aléatoire ou déterministe. Le but est de prédire Y à toute nouvelle entrée X. Dans l'analyse de régression, il existe deux principaux modèles : le modèle homoscédastique dans lequel la variance des erreurs de la régression est constante, et le modèle hétéroscédastique qui correspond au cas où la variance des erreurs dépend de X. Dans ce manuscrit, nous considérons le modèle de régression hétéroscédastique. Sur le plan statistique, cette hétéroscédasticité doit être détectée et estimée car sa non prise en compte dans l'estimation invalide les conclusions de nombreux problèmes d'inférence statistique tels que les tests statistiques qui supposent que les erreurs du modèle ont toutes la même variance. Un exemple classique en statistique qui illustre l'importance de la prise en compte de l'hétéroscédasticité est apparu dans la régression linéaire hétéroscédastique. L'estimateur des moindres carrés ordinaires (MCO) est non biaisé mais il n'est pas efficace. Il existe un estimateur des moindres carrés généralisés (c'est un estimateur des MCO sur un modèle transformé dont les aléas sont homoscédastiques et non autocorrélés) qui est généralement plus efficace que l'estimateur des MCO. En outre, les intervalles de prédiction obtenus à l'aide d'une approche qui estime la variance des erreurs en fonction des variables d'entrée sont susceptibles d'être plus réalistes que ceux obtenus en supposant que la variance des erreurs est constante, puisque l'estimation de l'incertitude prédictive dépend de l'estimation de la variance de la variable de réponse. Un autre aspect important de l'estimation de l'hétéroscédasticité du modèle est que l'estimation ponctuelle de la fonction de régression est directement liée à la variance en ce point. Malgré son importance à différents niveau, l'étude statistique du problème d'estimation de la fonction de variance est moins étudiée que celle de la fonction de régression dans la littérature statistique. Enfin, il faut noter que la fonction de variance a été étudiée, pas seulement en statistiques, mais dans de nombres domaines tels que la finance, l'économie et le traitement de signal. En particulier, la volatilité de la variable étudiée est un paramètre important à considérer.

Dans ce manuscrit, nous nous intéresserons au modèle de régression hétéroscédastique et élaborerons des méthodes statistiques capables de s'adapter aux différentes structures/hypothèses que l'on imposera à ce modèle. Le premier objectif de cette thèse est de proposer et d'étudier des estimateurs de la fonction de variance basés sur l'*agrégation*. Bien que le problème d'estimation de la fonction de variance est un objectif en soi sur le plan statistique que nous considérons dans cette thèse, un autre aspect de ce manuscrit est de pointer des problèmes où l'estimation de la variance est cruciale et enfin de proposer des solutions à ces problèmes basées sur l'estimation de celle-ci. En particulier, nous nous focaliserons sur les problèmes suivants :

- **la régression avec option rejet** : nous introduisons le problème de régression avec option rejet qui vise à construire des procédures d'estimation capables de ne pas prédire lorsque le doute dans la valeur prédite est trop grand. Il apporte entre autres une meilleure compréhension des situations où il est judicieux d'avoir recours à une méthode autorisant le rejet. Le prédicteur avec option rejet résultant repose sur un seuillage de la fonction de variance. Autrement dit, le

fait de rejeter ou non dépend essentiellement de la fonction de variance : plus le bruit en une région de l'espace est fort, plus on a tendance à s'abstenir de prédire. Ainsi, en marge de l'étude du problème de rejet, nous proposons une estimation de ce prédicteur fondée sur une procédure semi-supervisée d'estimation de type plug-in et en étudiant ses propriétés statistiques.

- **Intervalle de prédiction** : l'objectif est de construire un intervalle de prédiction qui contient la variable à prédire avec grande probabilité dans le modèle gaussien hétéroscédastique. L'intervalle de prédiction optimal repose sur un seuillage de la densité conditionnelle qui dépend des fonctions de régression et de variance. L'estimation de cet intervalle repose sur une règle de type plug-in. En particulier, l'estimateur final dépend des estimateurs des fonctions de régression et de variance.

Dans ce chapitre, nous commençons par introduire le modèle de régression. Puis, nous présentons les estimateurs de la fonction de régression les plus utilisés dans la littérature, en particulier nous développons la procédure d'agrégation qui est un outil puissant utilisé dans le cadre de l'estimation de la fonction de régression. S'ensuit une présentation de l'estimation de la fonction de variance dans le cadre général. Une partie importante de l'introduction est consacrée à une présentation détaillée de l'apprentissage sous contrainte : l'option rejet et les intervalles de confiance. Enfin, nous décrivons les contributions de la thèse.

Sommaire

1.1 Modèle de régression

Le problème de régression est l'un des sujets les plus populaires et les plus classiques en statistique [9, 36, 78, 82, 86]. Les modèles de régression étudient la relation entre la variable dite de réponse et une ou plusieurs variables explicatives. Par exemple, nous pouvons nous intéresser à l'influence de la vitesse sur le nombre d'accidents sur la route. Dans ce cas, le praticien s'intéresse à expliquer une variable à travers une autre. Afin de formaliser le problème, nous considérons un couple de variables aléatoires $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ où $X \in \mathbb{R}^d$ est la variable explicative et $Y \in \mathbb{R}$ est la variable à prédire associée à l'entrée X telle que

$$Y = f^*(X) + \xi \tag{1.1}$$

où

- $f^*(x) = \mathbb{E}[Y|X = x]$, $x \in \mathbb{R}^d$, est appelée fonction de régression de Y sur X,
- ξ est le bruit dans le modèle qui satisfait les propriétés suivantes : $\mathbb{E}[\xi|X] = 0$ et $\mathbb{E}[\xi^2] < \infty$.

Dans la suite, pour tout $x \in \mathbb{R}^d$, nous définissons

$$\sigma^2(x) = \mathbb{E}\left[(Y - f^*(X))^2|X = x\right]$$

la fonction de variance conditionnelle.

Problème de régression : étant donné $X \in \mathbb{R}^d$, le but est de trouver la bonne prédiction de Y en supposant que $(X, Y) \sim \mathbb{P}$ où $\mathbb{P}$ est inconnue. Cela signifie que l'on souhaite trouver une fonction mesurable $f : \mathbb{R}^d \to \mathbb{R}$ telle que $f(X)$ approxime bien Y. Pour cela, nous introduisons le risque L^2 de f

$$\mathbb{E}\left[|Y - f(X)|^2\right]$$

que nous voulons le plus petit possible avec l'hypothèse suivante $\mathbb{E}[Y^2] < \infty$.

Pour toute fonction mesurable f, le risque L^2 de f peut être décomposé comme suit :

Lemme 1. *Pour toutes les fonctions mesurables $f : \mathbb{R}^d \to \mathbb{R}$ avec $\mathbb{E}\left[f^2(X)\right] < \infty$,*

$$\mathbb{E}\left[|Y - f(X)|^2\right] - \mathbb{E}\left[|Y - f^*(X)|^2\right] = \mathbb{E}\left[(f(X) - f^*(X))^2\right] \quad . \tag{1.2}$$

Démonstration. Il suffit de remarquer que

$$\begin{aligned}
\mathbb{E}\left[(Y - f^*(X))(f^*(X) - f(X))\right] &= \mathbb{E}[\mathbb{E}\left[(Y - f^*(X))(f^*(X) - f(X))|X\right]] \\
&= \mathbb{E}[(f^*(X) - f(X))\mathbb{E}\left[(Y - f^*(X))|X\right]] \\
&= \mathbb{E}[(f^*(X) - f(X))(\mathbb{E}\left[Y|X\right] - f^*(X))] = 0.
\end{aligned}$$

□

Le terme du membre de droite de l'Eq.(1.2) est appelé l'excès de risque ou l'erreur L^2 de f. Nous pouvons déduire du Lemme 1 que la fonction de régression f^* est bien le meilleur prédicteur de la variable de réponse Y basé sur X au sens du risque L^2 :

$$R := \mathbb{E}\left[|Y - f^*(X)|^2\right] = \inf_f \mathbb{E}\left[|Y - f(X)|^2\right]$$

où l'infimum est pris sur toutes les fonctions mesurables $f : \mathbb{R}^d \to \mathbb{R}$. La quantité R est appelé la variance résiduelle et est l'erreur quadratique moyenne minimale de tout prédicteur de Y basé sur l'observation de la variable d'entrée X.

Et après ? Typiquement, la distribution de (X, Y) et la fonction de régression f^* sont inconnues. En particulier, il est impossible de prédire Y en utilisant $f^*(X)$. Au lieu de cela, nous observons un échantillon $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ qui consiste en n copies indépendantes et identiquement distribuées (i.i.d.) de (X, Y). L'objectif de la régression est d'utiliser cet échantillon pour construire un estimateur $\hat{f}_n : \mathbb{R}^d \to \mathbb{R}$ de f^* dont le risque L^2 est faible. Dans la section suivante, nous présentons les estimateurs de la fonction de régression.

1.2 Estimations de la fonction de régression

La littérature sur les méthodes d'estimation de la fonction de régression f^* est vaste et impossible à recenser. Nous citons ici quelques références pertinentes pour la suite de l'exposé [9, 36, 78, 82, 86]. Dans tout ce manuscrit, nous plaçons dans le cadre de la régression non-paramétrique [86]. En somme, nous pouvons distinguer deux types d'estimateurs : les estimateurs linéaires et non-linéaires.

1.2.1 Estimateurs linéaires

Une méthode d'estimation de la fonction de régression f^* à partir de l'échantillon d'apprentissage $\mathcal{D}_n$ consiste à proposer un estimateur $\hat{f}_n$ défini comme suit

$$\hat{f}_n(x) = \sum_{i=1}^n W_i(x) Y_i,$$

où les poids réels $W_i(x)$ vont être des fonctions boréliennes de x et $X_1, \ldots, X_n$. Dans cette répresentation, $\hat{f}_n(x)$ se présente comme une moyenne pondérée des Y_i correspondants aux X_i situés dans

un voisinage de x. Ce voisinage est caractérisé par les poids $W_i(x)$. Cet estimateur est appelé *estimateur par moyennage locale*. Le théorème ci-dessous, initialement dû à Stone [8], fournit des conditions suffisantes sur les poids $W_i(x)$, pour assurer la consistance, dite universelle, de l'estimateur $\hat{f}_n$.

Théorème 2 (Theorem 10.1 [8]). *Supposons que les poids satisfassent les conditions suivantes quelle que soit la loi de X :*

(i) Pour toute fonction borélienne $f : \mathbb{R}^d \to \mathbb{R}$ telle que $\mathbb{E}[|f(X)|] < \infty$, il existe une constante $C > 0$ telle que

$$\mathbb{E}\left[\sum_{i=1}^{n} |W_i(X)||f(X_i)|\right] \leq C\mathbb{E}[|f(X)|] \quad \textit{pour tout} \quad n \geq 1.$$

(ii) Il existe une constante $D \geq 1$ telle que $\mathbb{P}\left(\sum_{i=1}^{n} |W_i(x)| \leq D\right) = 1$.

(iii) Pour tout $a > 0$, $\sum_{i=1}^{n} |W_i(X)| \mathbb{1}_{\{\|X_i - X\| > a\}}$ converge vers 0 en probabilité.

(iv) $\sum_{i=1}^{n} W_i(X)$ converge vers 1 en probabilité.

(v) $\max_{1 \leq i \leq n} W_i(X)$ converge vers 1 en probabilité.

Alors $\hat{f}_n$ est universellement L^2-consistante , i.e.,

$$\mathbb{E}\left[|\hat{f}_n(X) - f^*(X)|^2\right] \to 0 \quad \textit{quand} \quad n \to \infty \tag{1.3}$$

quelle que soit la loi du couple (X, Y) avec $\mathbb{E}[|Y|^2] < \infty$.

Les poids $W_i(x)$ peuvent être positifs et normalisés à 1, de telle sorte que $(W_1(x), \ldots, W_n(x))$ peut être assimiler à un vecteur de probabilités. Dans la suite de ce paragraphe, nous présentons des exemples classiques d'estimateurs linéaires de la fonction de régression qui vérifient les conditions du théorème de Stone.

• **L'algorithme des k plus proches voisins.** Les k-plus proches voisins (*k-Nearest neighbors (k-NN)* en anglais) est une idée intuitive et est un exemple d'estimateurs linéaires de la fonction de régression f^*. Pour estimer f^* en un point $x \in \mathbb{R}^d$, nous regardons les k (k un entier strictement positif compris entre 1 et n) observations X_i les plus proches de x et nous considérons comme valeur prédite en x la moyenne des Y_i correspondants. L'estimateur résultant s'appelle estimateur des k-plus proches voisins qui est un estimateur de type moyenne locale. Nous désignons la suite $(X_{(1)}(x), Y_{(1)}(x)), \ldots, (X_{(n)}(x), Y_{(n)}(x))$, la permutation de $(X_1, Y_1), \ldots, (X_n, Y_n)$ correspondante aux distances croissantes des $\|X_i - x\|$[1]. En d'autres termes,

$$\|X_{(1)} - x\| \leq \ldots \leq \|X_{(n)} - x\|.$$

Dans la suite nous supposerons que les égalités entre distances $\|X_i - x\| = \|X_j - x\|$ se produisent avec probabilité zéro, ce qui est le cas lorsque la variable aléatoire $\|X - x\|$ admet une densité par rapport à la mesure de Lebesgue. L'estimateur des k-plus proches voisins s'écrit donc

$$\hat{f}_{knn}(x) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{(i)}(x), \tag{1.4}$$

pour tout $x \in \mathbb{R}^d$. Notons que dans cette expression, $k = k_n$ est autorisé à dépendre de n. La consistance et la vitesse de convergence de l'erreur L^2 de $\hat{f}_{knn}$ ont été étudiées dans de nombreux travaux (voir par exemple [9, 36]). En particulier, nous pouvons établir le résultat suivant de vitesse de convergence.

[1] En cas d'égalité $\|X_i - x\| = \|X_j - x\|$ avec $i < j$, X_i est déclaré plus proche de x que X_j

Théorème 3 (Theorem 6.2 [36]). *Soit $d \geq 3$. Supposons que la variable X est bornée, que $\sup_{x\in\mathbb{R}^d} \sigma^2(x) \leq L_{\sigma^2}$ et que f^* est L-Lipschitz, c-à-d : il existe $L > 0$ telle que*

$$|f^*(x) - f^*(z)| \leq L\|x - z\| \ \ \forall x, z \in \mathbb{R}^d.$$

Alors pour $k_n \propto n^{2/(d+2)}$, nous avons

$$\mathbb{E}[(\hat{f}_{knn}(X) - f^*(X))^2] \lesssim n^{-2/(d+2)}.$$

Pour $d \leq 2$, la vitesse de convergence du Théorème 3 est également vérifiée sous la condition de forte densité [2] : la distribution marginale $\mathbb{P}_X$ admet une densité μ par rapport à la mesure de Lebesgue, telle que pour tout $x \in \mathbb{R}^d$, nous avons $0 < \mu_{min} \leq \mu(x) \leq \mu_{max} < \infty$.
Le paramètre k joue un rôle important dans la performance de l'estimateur $\hat{f}_{knn}$. Le Théorème ci-dessus fournit une valeur de k qui n'est pas accessible en pratique (la constante dépend de paramètres inconnus). Diverses méthodes de choix de k sont donc disponibles. En pratique, la méthode utilisée le plus souvent est la validation croisée (pour plus de détails voir [36]). Cette approche consiste à introduire un ensemble fini $\mathcal{K}_n := \{1, \dots, n\}$ qui servira de grille pour le paramètre k. Puis, pour chaque $k \in \mathcal{K}_n$, nous calculons les deux estimateurs $\hat{f}^{(k)}_{knn} := \hat{f}_{knn}$ et $\hat{f}^{(k)}_{knn,i}$ de la fonction de régression sur $\mathcal{D}_n$ et $\mathcal{D}_n \backslash \{(X_i, Y_i)\}$ respectivement. La sélection par validation croisée (nous considérons ici la version leave-one-out de la validation croisée) de k est alors donnée par

$$\hat{k}^{cv} = \underset{k\in\mathcal{K}_n}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}^{(k)}_{knn,i}(X_i))^2.$$

Enfin, l'estimateur par validation croisée est donné par $\hat{f}_{knn}(x) = \hat{f}^{(\hat{k}^{cv})}_{knn}(x)$.

• **Méthode à noyau.** Les estimations à noyau ont été étudiées à l'origine dans le cadre de l'estimation de la fonction de densité par Rosenblatt [73] et Parzen [70]. Ils ont été étendus au problème de régression par Nadaraya [65, 66] et Watson [98]. La méthode à noyau donne un estimateur linéaire.

Définition 4. *On appelle noyau une fonction $K : \mathbb{R}^d \to [0, +\infty[$ telle que $\int K = 1$ et $\int K^2 < \infty$.*

Définition 5. *Soit K un noyau et $h > 0$ un paramètre que l'on appellera fenêtre. L'estimateur de Nadaraya-Watson de f^* est donné par*

$$\hat{f}_{NW}(x) = \sum_{i=1}^{n} W_i^{NW}(x)Y_i, \ \ où \ \ W_i^{NW}(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)} \mathbb{1}_{\left\{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right) \neq 0\right\}}.$$

Le noyau K détermine la forme du voisinage autour du point x. Il est souvent pertinent de considérer des noyaux réguliers, ce qui permet d'obtenir des estimateurs réguliers à leur tour, par exemple

- le noyau gaussien : $K(x) = \frac{1}{2\pi}\exp\left(-\frac{\|x\|^2}{2}\right)$;
- le noyau d'Epanechnikov : $K(x) = \frac{3}{4}\left(1 - \|x\|^2\right)^2 \mathbb{1}_{\{\|x\|\leq 1\}}$.

La fenêtre $h := h_n$ est le paramètre de régularisation du modèle qui dépend de n en général. Elle a un rôle crucial dans la consistance de l'estimateur $\hat{f}_{NW}$ au sens du l'erreur L^2 au même titre que le paramètre k pour les k-ppv. Ainsi, les conditions suivantes

$$h_n \to 0 \ \ et \ \ nh_n \to \infty \ \ lorsque \ \ n \to \infty$$

sont minimales afin d'établir la consistance, par exemple au sens de l'erreur L^2 de l'estimateur $\hat{f}_{NW}$ (Théorème 2). Dans ce cas également, de nombreuses méthodes de calibration de h_n existent telles

que la validation croisée, plug-in ou Boostrap [69, 36]. Chaque méthode possède des avantages et des inconvénients. L'idée principale de la validation croisée est de minimiser le critère suivant sur h

$$CV(\hat{f}_{NW}) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{f}_{NW}^{(-i)}(X_i)\right)^2 \quad ,$$

où $\hat{f}_{NW}^{(-i)}$ est l'estimateur de la régression qui ne tient pas compte l'observation i. Le résultat suivant précise la vitesse de convergence de $\mathbb{E}\left[\int(\hat{f}_{NW}(x) - f^*(x))^2\mathbb{P}_X(dx)\right]$ dans le cas du noyau naïf ($K(x) = \mathbb{1}_{\{\|x\|\leq 1\}}$).

Théorème 6 (Theorem 5.2 [36]). *Supposons que la variable X est bornée, $\sup_{x\in\mathbb{R}^d}\sigma^2(x) \leq L_{\sigma^2}$ et que f^* est L-lipschitzienne. Alors pour $h_n \propto n^{-\frac{1}{d+1}}$, nous avons*

$$\mathbb{E}\left[\int(\hat{f}_{NW}(x) - f^*(x))^2\mathbb{P}_X(dx)\right] \lesssim n^{-\frac{2}{d+2}} \quad .$$

Remarque 7 (**Estimation par des polynômes locaux [36, 86]**). *L'estimateur à noyau $\hat{f}_{NW}$ peut s'écrire comme solution du problème de minimisation suivant :*

$$\hat{f}_{NW}(x) = \underset{\theta}{\operatorname{argmin}}\sum_{i=1}^{n}(Y_i - \theta)^2 K\left(\frac{X_i - x}{h_n}\right).$$

L'estimateur de NW apparaît ainsi comme un cas particulier de l'estimateur par polynômes locaux (LL) au point x

$$\hat{f}_{LL}(x) = \sum_{l=0}^{M}\hat{\theta}_l(x)\phi_l(x) \quad ,$$

où

$$(\hat{\theta}_0, \ldots, \hat{\theta}_M) = \underset{(\theta_0,\ldots,\theta_M)}{\operatorname{argmin}}\sum_{i=1}^{n}\left(Y_i - \sum_{l=0}^{M}\theta_l\phi_l(X_i)\right)^2 K\left(\frac{X_i - x}{h_n}\right),$$

et $\phi_1, \ldots, \phi_M$ sont des fonctions sur $\mathbb{R}^d$. La vitesse de convergence de $\hat{f}_{LL}$ se dégrade très rapidement avec la dimension d. Cette propriété est connue sous le nom de fléau de la dimension. L'estimateur $\hat{f}_{LL}$ est bien étudié dans le cas où $d = 1$ [86] en prenant $\phi_l = x^l$. Si $M = 0$, on revient au cadre de l'estimateur à noyau. Si $M = 1$, on parle de la régression linéaire locale.

1.2.2 Estimateurs non-linéaires

Nous pouvons citer les procédures non-linéaires populaires : les machines à vecteurs de support (support vector machines (SVM) en anglais) et les forêts aléatoires (random forest en anglais).

• **Régression par Machines à Vecteurs de Support (SVM) [10, 27, 40, 91].** Les SVM, initialement introduites dans le cadre de la classification, peuvent être étendues au problème de régression en conservant toutes les propriétés qui caractérisent l'algorithme. Dans le cas des SVM linéaires, la fonction de prédiction prend la forme

$$f(x) := f(x, \omega) =< x, \omega > +b \quad ,$$

avec $x \in \mathbb{R}^d$ est la variable d'entrée, $\omega \in \mathbb{R}^d$ est le vecteur des poids et b est une constante à déterminer. Nous cherchons la relation affine f avec la plus petite valeur de ω. Nous résolvons donc le problème d'optimisation suivant :

$$\min \frac{1}{2}\|\omega\|^2 \text{ sous la contrainte } |y_i - f(x_i)| \leq \varepsilon, i = 1, \ldots, n.$$

Ce problème n'ayant pas forcément de solution réalisable lorsque ε est trop petit. Les SVM utilisent une fonction de perte appelée ε-intensible (au fonction de seuillage) de la forme

$$|y - f(x)|_\varepsilon = \begin{cases} 0 & \text{si } |y - f(x)| \leq \varepsilon, \\ |y - f(x)| - \varepsilon & \text{sinon.} \end{cases}$$

où $\varepsilon \geq 0$. Cela nous conduit à minimiser :

$$\frac{1}{2}\|\omega\|^2 + c\sum_{i=1}^{n} |y_i - f(X_i)|_\varepsilon,$$

sous des contraintes de type $|y_i - f(X_i)|_\varepsilon = |y_i - f(X_i)| - \varepsilon$. Le problème d'optimisation précédent n'est pas quadratique. Pour palier ce problème, il est usuel de considérer le concept de marge souple en introduisant des variables d'écart ξ_i et ξ_i^* pour formuler le problème d'optimisation

$$\min \frac{1}{2}\|\omega\|^2 + c\sum_{i=1}^{n}(\xi_i + \xi_i^*) \ ,$$

sous les contraintes

$$\begin{cases} y_i - f(X_i) \leq \varepsilon + \xi_i^* \\ f(X_i) - y_i \leq \varepsilon + \xi_i \\ \xi_i, \xi_i^* \geq 0, i = 1, \ldots, n. \end{cases}$$

La constante $c > 0$ est un hyperparamètre qui permet de régler le compromis entre la quantité d'erreur autorisée et la quantité $\frac{1}{2}\|\omega\|^2$. La solution de ce problème est donnée en minimisant le Lagrangien suivant

$$\mathcal{L} = \|\omega\|^2 + c\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*) - \sum_{i=1}^{n}\alpha_i\left(\varepsilon + \xi_i - y_i + f(X_i)\right) - \sum_{i=1}^{n}\alpha_i^*\left(\varepsilon + \xi_i^* + y_i - f(X_i)\right)$$

où les η_i, η_i^*, α_i et α_i^* sont positifs et représentent les multiplicateurs de Lagrange. Nous obtenons alors la forme explicite suivante :

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) < X_i, x > + b$$

où la constante b peut se calculer grâce aux conditions de Karush-Kuhn-Tucker (KKT). La fonction de prédiction f dépend des X_i et x à travers les produits scalaires $< X_i, x >$ et peut donc s'étendre à des formes plus générales. Ceci fait le lien avec les SVM non-linéaires. Dans le cas des SVM non-linéaires, nous introduisons un espace S (appelé espace de redescription (feature space)) et un noyau $k(x_i, x) =< \Phi(x_i), \Phi(x) >$ où $\Phi : \mathbb{R}^d \to S$ une projection. Comme précédemment, nous obtenons la fonction de décision f

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)k(X_i, x) + b$$

Les noyaux les plus utilisés sont

- polynomial : $k(x, z) = (\gamma < x, z > + c_0)^d$;
- sigmoïdal : $k(x, z) = \tanh(\gamma < x, z > + c_0)$;
- linéaire : $k(x, z) =< x, z >$;
- base radiale : $k(x, z) = \exp\left(-\gamma\|x - z\|^2\right)$,

où $\gamma > 0$ et $c_0 \in \mathbb{R}$.

• **Les arbres de régression [14, 35, 34].** Les arbres de régression (aussi appelés arbres de décision) sont des méthodes qui permettent d'obtenir des modèles prédictifs. Ils sont devenus très populaires au vu de leur faible temps de calcul, de leur capacité à gérer tous types de variables et à sélectionner les plus pertinentes, ainsi que la lisibilité et la facilité d'interprétation des résultats. L'algorithme le plus utilisé est l'arbre CART (CART signifie *Classification And Regression Trees*). Il consiste à partitionner récursivement l'espace d'entrée $\mathbb{R}^d$ de façon binaire et à déterminer une sous-partition optimale pour la prédiction.

Notons que la racine de l'arbre associé à $\mathbb{R}^d$ contient toutes les observations de l'échantillon $\mathcal{D}_n$. L'algorithme commence par découper au mieux la racine en deux noeuds fils. Un découpage (split en anglais) est représentée par la forme suivante :

$$\left\{X^{(j)} < v\right\} \cup \left\{X^{(j)} \geq v\right\}.$$

où $j \in \{1, \ldots, d\}$ et $v \in \mathbb{R}$. Nous traduisons ce split : les observations avec une valeur plus petite que v sur la j-ième variable seront dans le noeud fils de gauche, et les autres avec une valeur plus grande que v dans le noeud fils de droite. À ce stade, l'arbre CART choisit le meilleur split. Autrement dit, nous trouvons le meilleur couple (j, v) qui minimise la variance intra-noeuds suite au split d'un noeud en ses deux fils. De la même façon, nous répétons cette procédure sur chaque nouveau nœud fils jusqu'à atteindre un critère d'arrêt. Un exemple de critère d'arrêt peut être de ne pas découper des nœuds qui contiennent moins d'un certain nombre d'observations. Les nœuds terminaux sont appelés les feuilles de l'arbre. Nous associons à chaque feuille de l'arbre construit une valeur constante qui correspond à la moyenne des Y_i des points qui se trouvent dans cette feuille. Nous appelons l'arbre développé jusqu'au bout un arbre maximal (le nombre de feuilles est égal au nombre d'observations). Cet arbre a un biais très faible, mais une variance très grande. A l'inverse, un arbre constitué uniquement de la racine a une très petite variance mais un biais élevé. La profondeur de l'arbre est donc un paramètre de régularisation à calibrer.

Pour se faire la seconde étape de l'arbre CART se nomme l'élagage qui permet d'éviter le surapprentissage. Elle se repose sur le choix d'un modèle parmi l'ensemble des sous-arbres élagués de l'arbre maximal. Ceci se fait par la minimisation d'un critère pénalisé où la pénalité est proportionnelle au nombre de feuilles de l'arbre.

Forêts aléatoires [13, 78, 76, 77, 34]. Les arbres de décisions ont une forte variabilité (changer une observation peut changer tout l'arbre). Pour réduire cet effet, on a recours au bagging. Les forêts aléatoires ont été introduites par Breiman [13] dont le principe consiste à agréger plusieurs arbres de décision.

Définition 8. *Soit* $\{\hat{f}(x, \Theta_1), \ldots, \hat{f}(x, \Theta_q)\}$ *une collection de prédicteurs par arbre où* $(\Theta_1, \ldots, \Theta_q)$ *est une suite de variables aléatoires* i.i.d. *. Le prédicteur des forêts aléatoires* $\hat{f}_{Rf}$ *est obtenu par agrégation de cette collection d'arbres*

$$\hat{f}_{rf}(x) = \frac{1}{q} \sum_{i=1}^{q} \hat{f}(x, \Theta_j).$$

Notons qu'une famille de forêt aléatoire se distingue parmi les autres procédures par la qualité de ses performances sur de nombreux jeux de données. Dans ce contexte, nous parlons des Random Forests-RI (forêts aléatoires à variables d'entrée aléatoires). Les forêts aléatoires sont générées comme suit :

1. Nous donnons d'abord le nombre d'arbres q et le nombre de variables candidates pour découper un noeud ($m \in \mathbb{N}^*$).
2. Nous tirons un échantillon bootstrap dans $\mathcal{D}_n$.
3. Nous construisons un arbre CART $h(x, \Theta_q)$ sur cet échantillon bootstrap : chaque coupure est sélectionnée en minimisant la fonction de coût de CART sur un ensemble de m variables choisies au hasard parmi les d variables.

4. Enfin, l'estimateur des forêts aléatoires est la moyenne des prédictions individuelles des arbres

$$\hat{f}_{rf}(x) = \frac{1}{q}\sum_{j=1}^{q} \hat{f}(x, \Theta_j).$$

Un avantage majeur de l'algorithme de forêt aléatoire est qu'il peut donner de bons résultats même en grande dimension dû à l'étape de sélection aléatoire parmi l'ensemble des variables.
De nombreux travaux sur l'analyse théorique des forêts aléatoires ont été menés comme ceux de [78, 76, 77].
Conclusion. Dans la littérature, il existe plusieurs estimateurs de la fonction de régression qui ont été bien étudiés d'un point de vue théorique et numérique : Bagging [12], Boosting [75], Réseaux de neurones [40], Smoothing spline [96], Wavelets [58], Lasso, Ridge, Elastic-Net, Fused-Lasso, Group-Lasso [40] etc.

1.2.3 Simulations

Nous étudions les performances numériques des estimateurs construits à partir des algorithmes des forêts aléatoires (rf), des k-plus proches voisins (k-PPV), des machines à vecteurs de support (svm), des arbres de régression (tree) et du noyau (NW) ($\hat{f}_{rf}$, $\hat{f}_{svm}$, $\hat{f}_{knn}$, $\hat{f}_{NW}$ et $\hat{f}_{tree}$ respectivement). Ces algorithmes seront utilisés dans la thèse. Pour les algorithmes rf, svm, knn, tree et NW, nous utilisons respectivement les packages `R`, `randomForest`, `e1071`, `FNN`, `tree` et `NonpModelCheck` avec les paramètres par défaut à l'exception de la fenêtre h et l'entier k qui sont choisis par validation croisée et $maxnode = 23$ est fixé pour rf. Nous étudions la performance de l'estimateur à noyau uniquement dans les modèles univariés car il est coûteux de les calculer en grande dimension. Les performances de l'estimateur $\hat{f}$ sont évaluées comme suit. On répète indépendamment 100 fois les étapes suivantes :

(i) Nous simulons deux ensembles de données $\mathcal{D}_n$ et $\mathcal{D}_T$ avec $n = T = 1000$.

(ii) À partir de $\mathcal{D}_n$, nous construisons les estimateurs $\hat{f}$.

(iv) Enfin, sur $\mathcal{D}_T$, nous calculons leurs erreurs empiriques L^2 : $\widehat{\mathrm{Err}}(\hat{f}) = \frac{1}{T}\sum_{i=1}^{T}(\hat{f}(X_i) - f^*(X_i))^2$.

En considérant les 100 répétitions, nous calculons la moyenne et l'écart-type de $\widehat{\mathrm{Err}}$. Pour notre étude numérique, nous considérons 4 exemples du modèle suivant

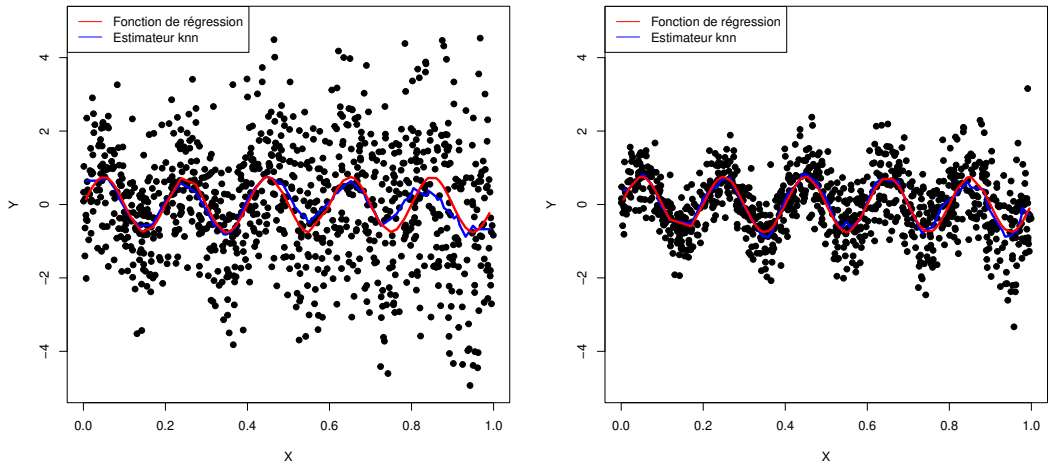
$$Y = f^*(X) + \sigma(X)\varepsilon \ , \tag{1.5}$$

où la variable d'entrée X suit la loi uniforme sur $[0,1]^d$ et ε suit la loi normale centrée réduite.

- Modèle 1 : $d = 1$, $f^*(X) = 0.75\sin(10\pi X)$, $\sigma^2(X) = 1 + 4X^2$.
- Modèle 2 : $d = 1$, $f^*(X) = 0.75\sin(10\pi X)$, $\sigma^2(X) = (1 + 4X^2)/5$.
- Modèle 3 : $d = 50$, $f^*(X) = X_1^2 + X_2^2 X_3 \exp(-X_4) + X_6 - X_8$, $\sigma^2(X) = 0.5 + \exp(-(X_1 - 0.2)^2) + 2\sin(X_2 X_3)^2 + 2X_4 X_5 X_6$.
- Modèle 4 : $d = 50$, $f^*(X) = X_1^2 + X_2^2 X_3 \exp(-X_4) + X_6 - X_8$, $\sigma^2(X) = |0.1X_1 X_2 + X_3^2 - X_4 X_7 - X_6^2 + X_8 X_{10}|^2$.

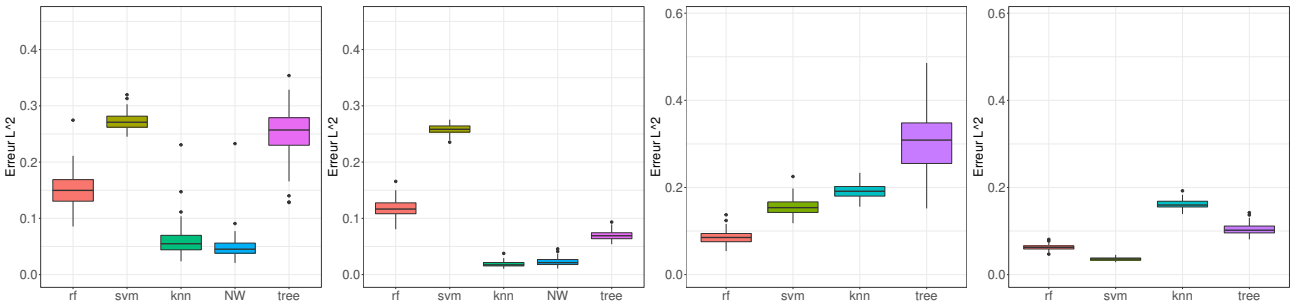
Pour chaque dimension $d \in \{1, 50\}$, les modèles ont la même fonction de régression, mais les fonctions de variance sont différentes. La fonction de variance dans les modèles 1 et 3 prend de grandes valeurs (des valeurs sont supérieures à 1). À l'opposé, elle prend des valeurs relativement modérées dans les modèles 2 et 4 (la majorité des valeurs sont inférieures à 1). Les résultats sont donnés dans la figure (1.2). Nous traçons le graphe de l'estimateur de $\hat{f}_{knn}$ dans la figure (1.1). Nous faisons deux observations. Premièrement, les meilleurs estimateurs de f^* dans chaque modèle sont $\hat{f}_{NW}$, $\hat{f}_{knn}$, $\hat{f}_{rf}$ et $\hat{f}_{svm}$ respectivement. De plus, l'estimateur $\hat{f}_{tree}$ a de bonnes performances lorsque la fonction de

variance prend des grandes valeurs. Deuxièmement, nous remarquons que les estimateurs dans les 4 modèles ont des performances similaires ce qui conduit à une concurrence entre eux. Finalement, nous concluons que plus la fonction de variance prend de grandes valeurs, plus l'estimation de f^* devient difficile.
Nous constatons que les estimateurs ont des performances proches dans chaque modèle. En général, l'existence de nombreuses méthodes d'estimation différentes conduit à des estimateurs éventuellement concurrents. Pour cela, nous allons présenter le cadre de l'agrégation d'estimateurs.



(a) Modèle 1 (b) Modèle 2

FIGURE 1.1 – Graphe de l'estimateur $\hat{f}_{knn}$ (blue) et de la fonction de régression (rouge). Les points noirs sont les données aléatoires de l'échantillon de taille 1000 montrant l'hétéroscédasticité.



(a) Modèle 1 (b) Modèle 2 (c) Modèle 3 (d) Modèle 4

FIGURE 1.2 – Boxplots de l'erreur empirique L^2 des estimateurs $\hat{f}_{rf}$, $\hat{f}_{svm}$, $\hat{f}_{knn}$, $\hat{f}_{NW}$ et $\hat{f}_{tree}$.

1.3 Agrégation d'estimateurs

La meilleure méthode d'estimation de la fonction de régression est l'agrégation. Nous présentons ci-après cette approche et ses propriétés.

1.3.1 Agrégation

L'idée de combiner des prédicteurs (estimateurs) est ancienne. Cette approche, appelée agrégation d'estimateurs, a été développée tant d'un point de vue théorique qu'algorithmique. Elle a été étudiée dans les communautés du machine learning, statistique, et traitement d'images [12, 31, 13, 8, 16, 43,

87].

Nous commençons par présenter l'intérêt de l'agrégation d'estimateurs par l'exemple de l'adaptation à la régularité inconnue d'une fonction (la fonction peut par exemple être une densité ou une fonction de régression). Il est bien connu que le choix d'un paramètre dit de régularisation est crucial pour l'adaptation à la régularité d'une fonction. Ce paramètre peut être la fenêtre pour les méthodes à noyau, le paramètre de lissage pour l'estimateur Lasso ou la constante de régularisation pour les méthodes pénalisées. Les procédures du choix de ce paramètre par validation croisée, par boostrap parmi d'autres méthodes, permettent de sélectionner le paramètre d'une façon aléatoire (dépendante des données). L'objectif est de déterminer une seule valeur utile du paramètre. Ce qui se traduit par un problème de la sélection du modèle. Nous pouvons envisager d'utiliser plusieurs estimateurs au lieu d'un seul en les agrégeant d'une manière adéquate. Nous nous intéressons dans la suite de ce chapitre aux techniques d'agrégation dans le cadre de la régression.

Dans le cadre plus développé de l'estimation de la fonction de régression, l'agrégation a pour but d'estimer la fonction de régression inconnue f^* par une combinaison d'éléments d'un ensemble connu de fonctions appelé dictionnaire. Ces fonctions sont des estimateurs tels que : la régression linéaire, les k-PPV, les noyaux, les forêts aléatoires, les SVM, le lasso, le ridge, les réseaux de neurones, le bagging, le gradient boosting, etc. Étant donné M ($M \geq 2$) estimateurs de f^* notés $\hat{f}_1, \ldots, \hat{f}_M$, le problème de l'agrégation consiste à chercher un nouvel estimateur $\tilde{f}$ en combinant les estimateurs précédents de manière appropriée. Ce nouveau estimateur $\tilde{f}$ est appelé agrégat et sa construction est appelée agrégation. Les procédures d'agrégation sont généralement basées sur la division de l'échantillon principal $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ en deux sous-échantillons $\mathcal{D}_{n_1}$ et $\mathcal{D}_{n_2}$ avec $n_1 + n_2 = n$ tels que l'échantillon $\mathcal{D}_{n_1}$ soit utilisé pour construire les éléments du dictionnaire $\hat{f}_1, \ldots, \hat{f}_M$ et que l'échantillon $\mathcal{D}_{n_2}$ serve à les agréger. En général, nous supposons que l'échantillon d'apprentissage a été fixé dans de nombreux travaux de sorte que les estimateurs $f^*, \hat{f}_1, \ldots, \hat{f}_M$ deviennent des fonctions fixées (déterministes).

1.3.2 Propriété de l'agrégat

La performance d'un estimateur $\tilde{f}$ peut être mesurée par le risque quadratique, c'est-à-dire par

$$r(\tilde{f}) = \mathbb{E}[\|\tilde{f} - f^*\|^2]$$

avec $\|f\|^2 = \int |f(x)|^2 \mathbb{P}_X(dx)$ où $\mathbb{P}_X$ est la mesure de probabilité de X. Posons $\hat{f}_\theta = \sum_{j=1}^M \theta_j \hat{f}_j$ pour un certain ensemble de paramètres $\Theta \subseteq \mathbb{R}^M$ ou $\theta \in \Theta$. Trois grands types d'agrégation sont considérés dans la littérature :

1. Agrégation par sélection de modèle (MS) : construire un estimateur agrégat $\tilde{f}$ qui est au moins aussi bon que le meilleur parmi les candidats $\hat{f}_1, \ldots, \hat{f}_M$ à un terme résiduel près, c'est-à-dire

   $$r(\tilde{f}) \leq \min_{1 \leq j \leq M} r(\hat{f}_j) + \triangle_{n,M}^{\text{MS}} \ ,$$

   où $\triangle_{n,M}^{\text{MS}}$ tend vers 0 quand n tend vers l'infini.

2. Agrégation convexe (C) : construire un estimateur agrégat $\tilde{f}$ qui est au moins aussi bon que la meilleure combinaison convexe de candidats $\hat{f}_1, \ldots, \hat{f}_M$ à un terme résiduel près, c'est-à-dire

   $$r(\tilde{f}) \leq \min_{\theta \in \Theta} r(\hat{f}_\theta) + \triangle_{n,M}^{\text{C}} \ ,$$

   où $\triangle_{n,M}^{\text{C}}$ tend vers 0 quand n tend vers l'infini, et $\Theta = \{\theta \in \mathbb{R}^M : \theta_j \geq 0, \sum_{j=1}^M \theta_j = 1\}$.

3. Agrégation linéaire (L) : construire un estimateur agrégat $\tilde{f}$ qui est au moins aussi bon que la meilleure combinaison linéaire de candidats $\hat{f}_1, \ldots, \hat{f}_M$ à un terme résiduel près, c'est-à-dire

   $$r(\tilde{f}) \leq \min_{\theta \in \mathbb{R}^M} r(\hat{f}_\theta) + \triangle_{n,M}^{\text{L}} \ ,$$

   où $\triangle_{n,M}^{\text{L}}$ tend vers 0 quand n tend vers l'infini, et $\Theta = \mathbb{R}^M$.

Les trois procédures d'agrégation ont été largement traité dans [16, 43, 87, 88, 104, 47, 48]. Les quantités $\triangle_{n,M}^{\texttt{MS}}$, $\triangle_{n,M}^{\texttt{C}}$ et $\triangle_{n,M}^{\texttt{L}}$ caractérisent la précision de procédures d'agrégation `MS`, `C` et `L` respectivement. Dans le modèle gaussien, sous certaines hypothèses, les vitesses optimales d'agrégations de type sélection de modèle, convexe et linéaire (voir [87]) sont données comme suit

$$\begin{cases} \triangle_{n,M}^{\texttt{MS}} \asymp \log(M)/n; \\ \triangle_{n,M}^{\texttt{C}} \asymp \begin{cases} M/n & \text{si } M \leq \sqrt{n}, \\ \sqrt{\log(1+M/\sqrt{n})/n} & \text{si } M > \sqrt{n}; \end{cases} \\ \triangle_{n,M}^{\texttt{L}} \asymp M/n. \end{cases}$$

En particulier, il est clair que

$$\min_{\theta\in\mathbb{R}^M} r(\hat{f}_\theta) \leq \min_{\theta\in\Theta} r(\hat{f}_\theta) \leq \min_{1\leq j\leq M} r(\hat{f}_j) \ .$$

Les deux premiers types d'agrégation sont au coeur des travaux présentés dans ce manuscrit, aussi, nous présentons ci-après les performances numériques de `MS`-estimateur et `C`-estimateur sur quelques données simulées .

1.3.3 Simulations

Nous étudions les performances du `MS`-estimateur $\hat{f}_{\texttt{MS}}$ et du `C`-estimateur $\hat{f}_{\texttt{C}}$ dans les modèles présentés dans la section 1.2.3. Leurs constructions sont plus amplement détaillées dans la partie 1.6.1 en utilisant deux échantillons indépendants. Nous introduisons un ensemble $\mathcal{F} = \{\hat{f}_s\}_{s=1}^5$ qui contient cinq estimateurs construits à partir des algorithmes des forêts aléatoires (rf), des k-plus proches voisins (K-PPV), des machines à vecteurs support (svm), des arbres de régression et des noyaux. Nous évitons la méthode à noyau dans les modèles de grande dimension. Nous répètons indépendamment 100 fois les étapes suivantes :

(i) Nous simulons trois ensembles de données $\mathcal{D}_n$, $\mathcal{D}_N$ et $\mathcal{D}_T$ avec $n = N = 500$, et $T = 1000$.

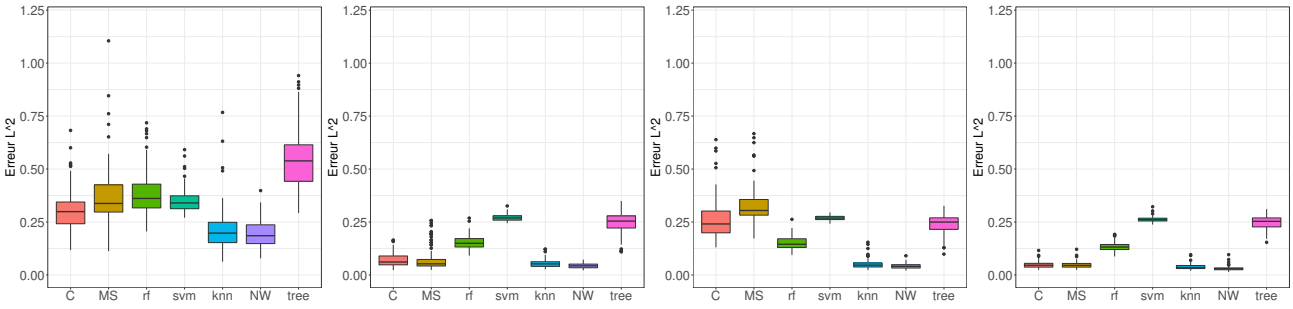
(ii) À partir de $\mathcal{D}_n$, nous construisons les estimateurs constituant $\mathcal{F}$, puis à partir de $\mathcal{D}_N$, nous calculons $\hat{f}_{\texttt{MS}}$ et $\hat{f}_{\texttt{C}}$.

(iii) Pour comparaison, à partir de $\mathcal{D}_n \cup \mathcal{D}_N$, nous calculons les estimateurs constituant $\mathcal{F}$.

(iv) Enfin, sur $\mathcal{D}_T$, nous calculons l'erreur empirique L^2 ($\widehat{\text{Err}}$) de $\hat{f}_{\texttt{MS}}$, de $\hat{f}_{\texttt{C}}$ et de tous les estimateurs de la fonction de régression f^* obtenus à l'étape (iii).

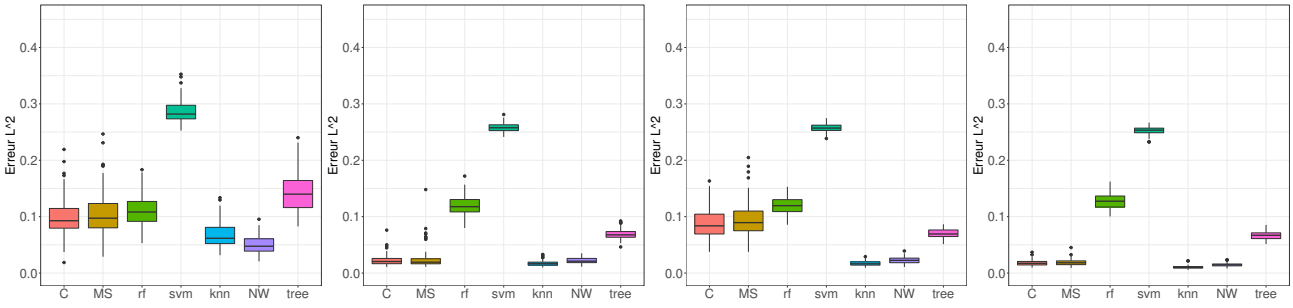
À partir des 100 répétitions, nous calculons la moyenne et l'écart-type de $\widehat{\text{Err}}$. Les résultats sont donnés dans les figures (1.3), (1.4), (1.5) et (1.6). Les courbes de MS-estimateur et de C-estimateur sont dessinées dans les figures (1.7) et (1.8). D'abord, nous observons que le MS-estimateur et le C-estimateur ont des performances similaires à celles du meilleur estimateur qui est construit à partir $\mathcal{D}_n \cup \mathcal{D}_N$ quand n et N sont assez grands. Par ailleurs, nous pouvons voir que la méthode `C` est toujours meilleure devant la méthode `MS`. Notons que l'utilisation de moins de données pour construire les $\hat{f}_j$ conduit à la perte en performance dans certains cas. Enfin, nos résultats numériques prouvent un fait important : lorsque nous divisons les données, il est avantageux de mettre plus de données dans le premier échantillon $\mathcal{D}_n$ utilisé dans l'étape d'estimation, plus précisément au stade de construction des estimateurs initiaux de f^*.

En conclusion, la qualité des estimateurs de la fonction de régression dépend fortement de la fonction de variance conditionnelle σ^2. Dans le paragraphe suivant, nous focalisons nos efforts à l'étude de la fonction σ^2.



(a) $n = N = 100$ (b) $n = 1000$, $N = 100$ (c) $n = 100$, $N = 1000$ (d) $n = N = 1000$

FIGURE 1.3 – Boxplot de l'erreur empirique L^2 des estimateurs : $\hat{f}_{\text{C}}$, $\hat{f}_{\text{MS}}$, $\hat{f}_{rf}$, $\hat{f}_{svm}$, $\hat{f}_{knn}$, $\hat{f}_{NW}$ et $\hat{f}_{tree}$ dans le modèle 1.



(a) $n = N = 100$ (b) $n = 1000$, $N = 100$ (c) $n = 100$, $N = 1000$ (d) $n = N = 1000$

FIGURE 1.4 – Boxplot de l'erreur empirique L^2 des estimateurs : $\hat{f}_{\text{C}}$, $\hat{f}_{\text{MS}}$, $\hat{f}_{rf}$, $\hat{f}_{svm}$, $\hat{f}_{knn}$, $\hat{f}_{NW}$ et $\hat{f}_{tree}$ dans le modèle 2.

1.4 Estimation de la fonction de variance conditionnelle

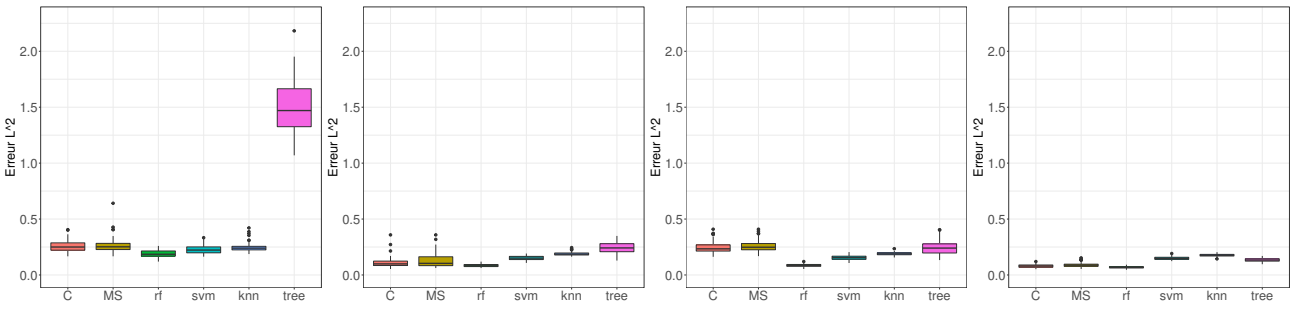
La fonction de variance conditionnelle est une mesure de l'hétéroscédasticité et joue un rôle important dans de nombreux contextes de modélisation statistique. Elle est importante dans de multiples problèmes en finance telles que la mesure de la volatilité ou du risque [1] ou les rendements boursiers à long terme [59]. En statistique, elle est utile dans la sélection de la fenêtre de lissage optimale du noyau [28], estimation du rapport signal sur bruit [93] et construction de bandes de confiance pour la fonction de régression [39].
L'estimation de σ^2 a été sérieusement envisagée dans les années 1980 [63]. Elle a été motivée par la construction des intervalles de confiance I_c pour la fonction f^* en régression gaussienne. L'intervalle I_c prend la forme habituelle suivante :

Étant donné $\alpha \in [0, 1]$,

$$I_c = \left]\hat{f}(x) - z_{\alpha/2}\sqrt{\hat{\sigma}^2(x) + s^2}, \hat{f}(x) + z_{\alpha/2}\sqrt{\hat{\sigma}^2(x) + s^2}\right[$$

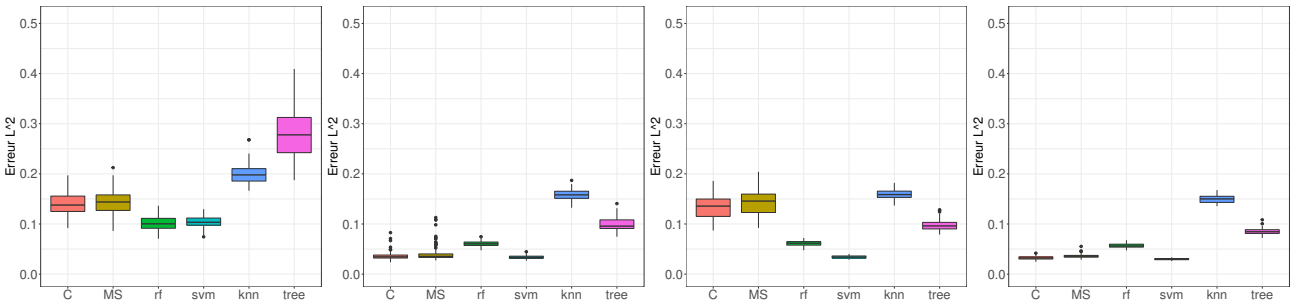
où $\hat{f}$ et $\hat{\sigma}^2$ sont les estimateurs de f^* et σ^2 respectivement. La quantité s est l'estimation de l'erreur standard de $\hat{f}$, et $z_{\alpha/2}$ peut être le quantile d'ordre $\alpha/2$ de la loi normale ou un quantile determiné par boostrap. La construction de I_c est bien détaillé par exemple dans les livres [29, 56] et l'introduction de l'article [52]. Cet intervalle dépend fortement de l'estimateur de la fonction de variance. En particulier, la longueur de I_c est égale à

$$2z_{\alpha/2}\sqrt{\hat{\sigma}^2(x) + s^2}.$$



(a) $n = N = 100$ (b) $n = 1000$, $N = 100$ (c) $n = 100$, $N = 1000$ (d) $n = N = 1000$

FIGURE 1.5 – Boxplot de l'erreur empirique L^2 des estimateurs : $\hat{f}_{\text{C}}$, $\hat{f}_{\text{MS}}$, $\hat{f}_{rf}$, $\hat{f}_{svm}$, $\hat{f}_{knn}$ et $\hat{f}_{tree}$ dans le modèle 3.



(a) $n = N = 100$ (b) $n = 1000$, $N = 100$ (c) $n = 100$, $N = 1000$ (d) $n = N = 1000$

FIGURE 1.6 – Boxplot de l'erreur empirique L^2 des estimateurs : $\hat{f}_{\text{C}}$, $\hat{f}_{\text{MS}}$, $\hat{f}_{rf}$, $\hat{f}_{svm}$, $\hat{f}_{knn}$ et $\hat{f}_{tree}$ dans le modèle 4.

Par conséquent, une mauvaise estimation de σ^2 conduit à des conclusions erronées.

Dans la suite, nous établissons un état de l'art sur le problème d'estimation de la fonction de variance. Dans la littérature, l'estimation de la fonction de variance est largement étudiée à la fois dans les cas de variables d'entrée fixées et aléatoires :

1.4.1 Design aléatoire

Dans ce manuscrit, nous nous focalisons sur le cas où les variables d'entrée sont aléatoires. Nous distinguons deux types d'approches : la méthode directe et la méthode basée sur les erreurs résiduelles. Dans la suite, nous présentons ces deux procédures en détail.

Méthode directe

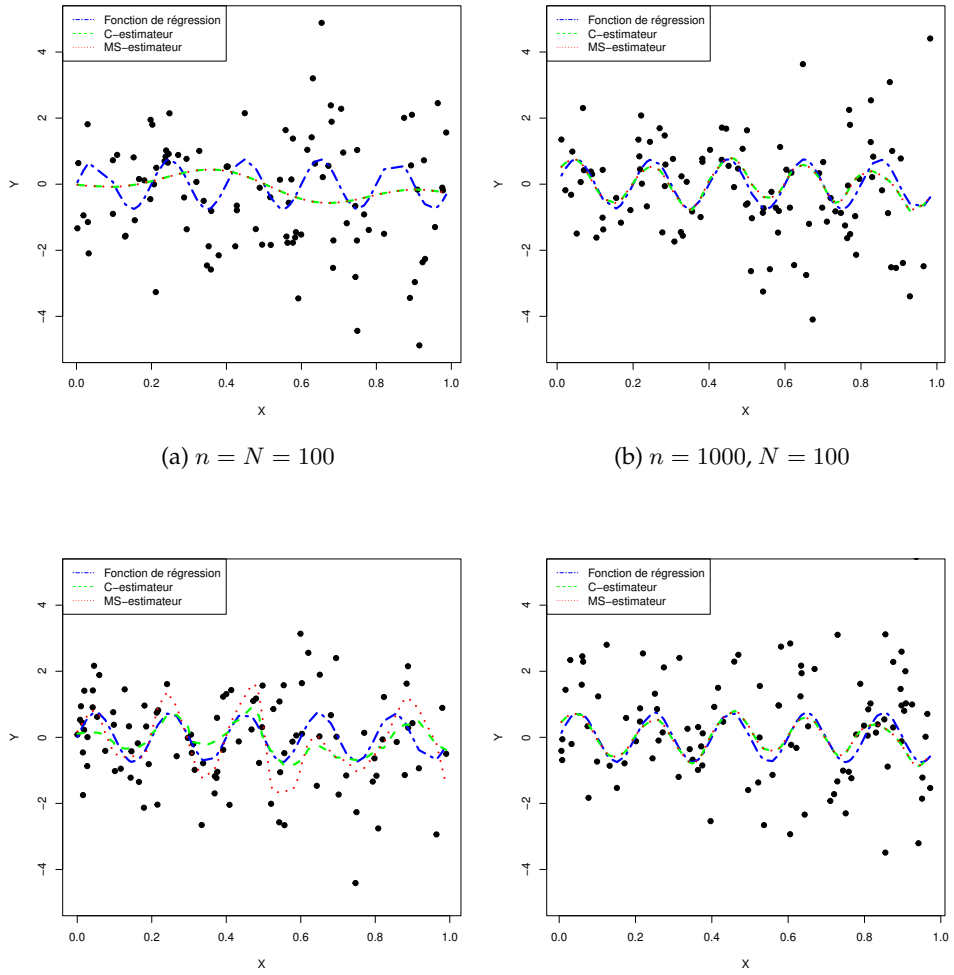
La méthode directe repose sur une décomposition de la fonction de variance conditionnelle σ^2 en moments

$$\sigma^2(x) = \mathbb{E}\left[Y^2 | X = x\right] - (f^*(x))^2 \ .$$

Elle consiste à estimer séparément les deux termes du côté droit. Pour être plus précis, l'estimateur direct de σ^2 a la forme suivante

$$\hat{\sigma}^2_d(x) = \hat{g}(x) - \hat{f}^2(x) \ , \tag{1.6}$$

où $\hat{g}$ et $\hat{f}$ sont les estimateurs de $\mathbb{E}\left[Y^2 | X = x\right]$ et f^* respectivement. Cette approche est par exemple considérée dans les articles [30, 38]. Les auteurs dans [38] ont étudié l'estimation de σ^2 dans le modèle autorégressif d'ordre 1 (en supposant que $X_i = Y_{i-1}$). Ils proposent une classe générale d'estimateurs



(a) $n = N = 100$ (b) $n = 1000$, $N = 100$

(c) $n = 100$, $N = 1000$ (d) $n = N = 1000$

FIGURE 1.7 – Graphes des estimateurs $\hat{f}_{\mathrm{C}}$, $\hat{f}_{\mathrm{MS}}$, et de la fonction de régression dans le modèle 1. Les points noirs sont les données aléatoires de l'échantillon de taille 100 montrant l'hétéroscédasticité.

de la fonction de volatilité basés sur les estimateurs par polynomiaux locaux (LP). Les définitions de $\hat{g}$ et $\hat{f}$ par la méthode LP reposent sur les problèmes de minimisation suivants :

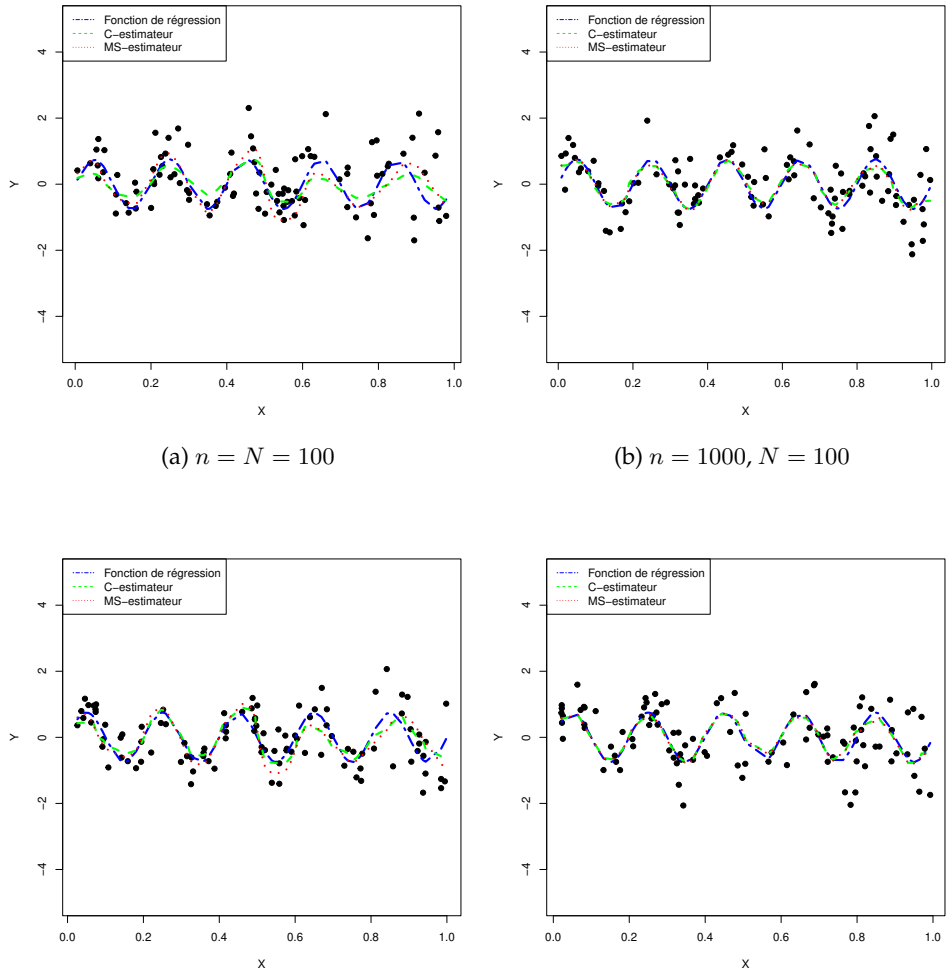
$$\bar{c}_n(x) = \underset{c\in\mathbb{R}^l}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i^2 - c^T U_{in})^2 K\left(\frac{Y_{i-1}-x}{h}\right) \quad ,$$
$$c_n(x) = \underset{c\in\mathbb{R}^l}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - c^T U_{in})^2 K\left(\frac{Y_{i-1}-x}{h}\right) \quad ,$$

où $K : \mathbb{R} \to \mathbb{R}$ est un noyau, $h > 0$ est une fenêtre et

$$U_{in} = F(u_{in}), \quad F(x) = \begin{pmatrix} 1 \\ x \\ \vdots \\ \frac{x^{l-1}}{(l-1)!} \end{pmatrix}, \quad u_{in} = \frac{Y_{i-1}-x}{h} \quad .$$

L'estimateur $\hat{\sigma}_d^2$ de σ^2 est alors défini par

$$\hat{\sigma}_n^2(x) = \bar{c}_n(x)^T F(0) - (c_n(x)^T F(0))^2 \quad .$$



(a) $n = N = 100$ (b) $n = 1000$, $N = 100$

(c) $n = 100$, $N = 1000$ (d) $n = N = 1000$

FIGURE 1.8 – Graphes des estimateurs : $\hat{f}_{\text{C}}$, $\hat{f}_{\text{MS}}$, et de la fonction de régression dans le modèle 2. Les points noirs sont les données aléatoires de l'échantillon de taille 100 montrant l'hétéroscédasticité.

La normalité asymptotique des estimateurs LP de la fonction de régression et la fonction de variance ont été établies.
La méthode directe pour estimer la fonction de variance conditionnelle est simple à mettre en œuvre. Toutefois, elle a un inconvénient majeur :

(i) l'estimateur de $\hat{\sigma}_d^2$ (1.6) n'est pas toujours positif en particulier lorsque différents paramètres de lissage sont utilisés pour établir les termes c_n et $\bar{c}_n$.

Méthode basée sur les erreurs résiduelles

La méthode basée sur les erreurs résiduelles consiste en deux étapes. Dans une première étape, nous construisons un estimateur $\hat{f}$ de la fonction de régression f^*. Dans une deuxième étape, un estimateur de σ^2 est obtenu en résolvant le problème de régression où la variable d'entrée est X et la variable à prédire est $(Y - \hat{f}(X))^2$. Ainsi quasiment toutes les méthodes classiques d'estimation peuvent être considérées. Dans la suite nous axons notre développement sur les méthodes basées sur les moyennes locales, c'est-à-dire, peuvent s'écrire sous la forme $\hat{\sigma}^2(x) = \sum_{i=1}^n w_i(x)(Y_i - \hat{f}(X_i))^2$ où chaque $w_i(x)$ est une fonction borélienne réelle de x et de $X_1, \ldots, X_n$. Bien souvent (mais pas toujours), les $w_i(x)$ sont positifs et normalisés à 1, de telle sorte que $(w_1(x), \ldots, w_n(x))$ est en fait un vecteur de probabilités. Voici deux exemples classiques d'estimateurs de type moyenne locale.

Estimateur à noyau. L'estimateur à noyau pour σ^2 est obtenu en prenant $w_i(x) = \frac{K\left(\frac{X_i-x}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i-x}{h}\right)}$ où K est un noyau sur $\mathbb{R}^d$ et $h \geq 0$. En d'autres termes, pour $x \in \mathbb{R}^d$

$$\hat{\sigma}_h^2(x) = \frac{\sum_{i=1}^n K\left(\frac{X_i-x}{h}\right)(Y_i - \hat{f}(X_i))^2}{\sum_{i=1}^n K\left(\frac{X_i-x}{h}\right)}. \tag{1.7}$$

Les auteurs dans [46] ont étudié l'estimateur (1.7) où ils ont également estimé lors de première étape la fonction de régression également par la méthode à noyau en supposant que la variable X a une variance unitaire et une densité. Ils ont établi les résultats asymptotiques de l'estimateur à noyau (1.7) dans ce cas. De plus, ils ont montré que l'estimation de la fonction de régression a une influence sur l'estimation de la fonction variance. L'inconvénient principal de cet estimateur est que l'estimation de f^* a un effet négatif sur le comportement asymptotique de l'estimateur (1.7) dans le cas où les erreurs ε_i ne sont pas indépendantes et identiquement distribuées.

Estimateur des k-plus proches voisins. En prenant les notations de la section 1.2.1, l'estimateur des k-plus proches voisins de σ^2 prend la forme suivante

$$\hat{\sigma}^2(x) = \frac{1}{k_n}\sum_{i=1}^{k_n}(Y_{(i)} - \hat{f}(X_{(i)}))^2 \ . \tag{1.8}$$

Cet estimateur, dans le cas où $\hat{f}$ est également un estimateur des k-plus proches voisins, sera étudié en détail dans la suite de ce manuscrit. Nous remettons donc à plus tard l'analyse statistique de cette méthode.
En outre, il existe différentes variantes de la méthode basée sur les erreurs résiduelles pour la deuxième étape d'estimation de σ^2 pour $d = 1$:

(i) *Application d'un estimateur localement linéaire* : Dans [30], les auteurs ont utilisé la même méthode pour estimer la fonction de régression et la fonction de variance. Plus précisément, ils ont appliqué l'estimateur localement linéaire dans les deux étapes. Soit $\hat{f}(x) = \hat{a}(x)$ l'estimateur localement linéaire qui résout le problème des moindres carrés pondérés suivant :

$$(\hat{a}, \hat{b}) = \underset{a,b\in\mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^n (Y_i - a - b(X_i - x))^2 K_1\left(\frac{X_i - x}{h_1}\right), \ \ \forall x \in \mathbb{R},$$

où K_1 est un noyau tel que $\int_{\mathbb{R}} K_1(x)dx = 1$ et $h_1 > 0$ est une fenêtre. Définissons $\hat{Z}_i = (Y_i - \hat{f}(X_i))^2$ pour tout $i = 1, \ldots, n$. L'estimateur basé sur les erreurs résiduelles $\hat{\sigma}^2(x) = \hat{\alpha}(x)$ de la fonction de variance est donné comme suit

$$(\hat{\alpha}, \hat{\beta}) = \underset{\alpha,\beta\in\mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^n (\hat{Z}_i - \alpha - \beta(X_i - x))^2 K_2\left(\frac{X_i - x}{h_2}\right) \ ,$$

où K_2 est un noyau sur $\mathbb{R}$ et $h_2 > 0$ est une fenêtre. L'estimateur précédent de σ^2 est asymptotiquement aussi bon que dans le cas où la fonction de régression f^* était donnée. Mais, il n'est pas nécessairement positif. Comme solution à cela, les auteurs dans [102] ont proposé des estimateurs basés sur une vraisemblance normale localisée de sorte que la fonction de variance estimée est toujours positive. Ils ont utilisé une forme localement linéaire pour estimer la fonction de régression, et une forme localement log-linéaire pour estimer la fonction de variance en séparant les fenêtres. Ils ont fourni une analyse asymptotique de leurs estimateurs.

(ii) *Estimateur localement exponentiel* : les auteurs dans [106] ont utilisé une approche similaire à celle dans [30] (Application d'un estimateur localement linéaire). Ils ont considéré un estimateur localement linéaire pour estimer la fonction de régression f^* (voir (i)). Cet estimateur est $\hat{f}(x) =$

$\hat{a}(x)$. L'estimateur localement exponentiel de la variance conditionnelle est obtenu à partir de la solution du problème de minimisation suivant :

$$(\hat{\alpha}_1, \hat{\beta}_1) = \underset{\alpha_1,\beta_1\in\mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n}((Y_i - \hat{f}(X_i))^2 - \exp(\alpha_1 + \beta_1(X_i - x)))^2 K\left(\frac{X_i - x}{h_2}\right) \ ,$$

où $h_2 > 0$ est une autre fenêtre. L'estimateur exponentiel est alors donné par $\hat{\sigma}_e^2(x) = \exp(\hat{\alpha}_1(x))$, et est donc toujours positif. Les auteurs ont prouvé qu'il a la même variance asymptotique (d'où la même efficacité) que l'estimateur de la fonction de variance obtenu dans (i).

(iii) *Estimateur localement constant repondéré* : étant donné l'estimateur de la fonction de régression de $\hat{f}$ dans (i), les auteurs dans [101] ont proposé un estimateur à noyau repondéré pour estimer σ^2

$$\hat{\sigma}_p^2(x) = \frac{\sum_{i=1}^n \hat{w}_i(x) K((X_i - x)/h)\hat{Z}_i}{\sum_{i=1}^n \hat{w}_i(x) K((X_i - x)/h)} \ ,$$

où les $\hat{w}_i(x)$ résolvent le problème d'optimisation suivant

$$(\hat{w}_1(x), \ldots, \hat{w}_n(x)) = \underset{(w_1(x),\ldots,w_n(x))}{\operatorname{argmin}} -2\sum_{i=1}^{n} \log(nw_i(x)) \ ,$$

sous les contraintes suivantes

$$w_i(x) \geq 0, \ \sum_{i=1}^{n} w_i(x) = 1, \ \text{ et } \ \sum_{i=1}^{n} w_i(x)(X_i - x)K_h(X_i - x) = 0 \ .$$

L'estimateur $\hat{\sigma}_p^2$ est asymptotiquement équivalent à l'estimateur de σ^2 obtenu dans (i) et est positif dans des échantillons finis.

L'existence de plusieurs estimateurs par la méthode basée sur les erreurs résiduelles permet de créer de la concurrence entre eux. En particulier, il est difficile de choisir le meilleur estimateur parmi eux. Dans ce contexte, il est naturel de poser la question suivante qui va guider nos recherches :

Question 1. Quelle est la meilleure façon d'estimer la fonction de variance ?
Nous apporterons des éléments de réponse à cette question en exploitant des arguments d'agrégation, voir section 1.6.1 pour plus de détail.

1.4.2 Design fixe

La méthode la plus populaire pour estimer la fonction de variance conditionnelle, lorsque la variable d'entrée X est déterministe, est la méthode des différences séquentielles (en anglais " *the difference-sequence method*"). Le premier estimateur basé sur les différences a été développé dans [63]. Les auteurs ont considéré le modèle (1.5) où les variables d'entrée $X_i \in [0, 1]$ pour tout $i = 1, \ldots, n$, les variables de bruit ε_i ont un quatrième moment fini et la fonction de variance σ^2 est une fonction γ-lipschitzienne continue avec $\gamma \in [0, 1]$. L'estimateur initial a la forme

$$\tilde{\sigma}^2(X_i) = \left(\sum_{j=j_1}^{j_2} w_j Y_{j+i}\right)^2 \ ,$$

où $j_1 = -\lceil m/2 \rceil$, $j_2 = \lceil m/2 - 1/4 \rceil$ avec m est un entier fixé et la séquence de différence $\{w_j\}$ satisfait les conditions

$$\sum_{j=j_1}^{j_2} w_j = 0, \ \textit{and} \ \sum_{j=j_1}^{j_2} w_j^2 = 1.$$

Les auteurs ont montré que cet estimateur n'est pas consistant et ont proposé la modification suivante de l'estimateur : ils l'ont lissé à l'aide d'un noyau K, c'est-à-dire, ils définissent l'estimateur

$$\hat{\sigma}^2(X) = \frac{1}{h}\sum_{j=1}^{n}\int_{s_j-1}^{s_j} K\left(\frac{X-u}{h}\right) du\, \tilde{\sigma}^2(X_j) \ ,$$

avec $s_j = \frac{X_j+X_{j+1}}{2}$, $1 \leq j \leq n-1$, $s_0 = 0$ et $s_n = 1$. Le paramètre h est la fenêtre de lissage qui satisfait la condition suivante : $h \to 0$ et $nh \to 0$ quand $n \to \infty$. L'estimateur $\hat{\sigma}^2$ a été montré consistant. Dans [97], les auteurs ont étudié les effets de la fonction de régression inconnue et lisse (en anglais *smooth*) sur l'estimation de la fonction de variance. Ils ont proposé un estimateur alternatif de σ^2 et ont évalué ses performances. Ils ont comparé les performances de leur estimateur avec l'estimateur basé sur les erreurs résiduelles proposé dans [30]. Leurs résultats suggèrent que l'utilisation de la méthode basée sur les erreurs résiduelles n'est adaptée que lorsque la fonction de régression est lisse. Prolongeant les travaux de [15, 63, 97] via des différences séquentielles, les auteurs de [17] ont étudié le cas général d'une fonction de variance multivariée hétéroscédastique.

1.5 Apprentissage sous contrainte

Dans la section précédente, nous avons énuméré certaines approches existantes pour estimer la fonction de variance. Ici, nous présentons deux problèmes d'apprentissage sous contrainte, l'option rejet et les intervalles de prédiction, dans lesquels la fonction de variance joue un rôle clé.

1.5.1 Option rejet en apprentissage

En général, dans un problème de prédiction, les algorithmes produisent toujours une prédiction, même lorsque celle-ci est susceptible d'être non pertinente. De nombreuses procédures d'estimation visent néanmoins à réduire les erreurs de prédiction, *e.g.*, les réseaux de neurones, les méthodes à noyaux, l'approche des K plus proches voisins (K-PPV), les moindres carrés régularisés pour n'en nommer qu'un petit nombre. Cependant, même les méthodes les plus performantes commettent des erreurs qui peuvent dans certains cas avoir de lourdes conséquences. Le domaine médical est un champ d'application notable dans lequel il est souvent souhaitable de s'abstenir plutôt que de prendre une mauvaise décision. Cette capacité à s'abstenir de faire une prédiction a plusieurs avantages. Premièrement, en ne faisant des prédictions que lorsqu'elle est fiable, nous améliorons les performances. Deuxièmement, en évitant les erreurs de prédiction, nous pouvons augmenter la confiance d'un utilisateur dans l'algorithme proposé. Troisièmement, s'abstenir à produire une prédiction peut encore entraîner des gains de temps en ne nécessitant que des interventions humaines pour prendre des décisions dans un petit nombre de cas. Introduite, pour la classification, l'utilisation de l'apprentissage avec option rejet a récemment été développé en régression. Nous développons ci-dessous les deux cas d'application pour comprendre le problème.

Classification avec option rejet

L'idée de classification avec une option rejet remonte aux articles [18, 19]. L'auteur a étudié à la fois la règle de décision de rejet optimale et le compromis entre le taux de rejet et l'erreur. En outre, cela se fait en particulier sous la fonction de perte (le coût) 0-1, en supposant que la distribution sous-jacente est complètement connue. La classification avec une option rejet consiste à permettre à un prédicteur de refuser de répondre à une question en présence d'incertitude. Les instances doivent être rejetées chaque fois qu'aucune des probabilités a posteriori n'est pas suffisamment forte. La classification avec option a été développée dans le cadre de la théorie de l'apprentissage statistique [19, 42, 103, 23, 21, 99, 41, 33]. Elle est basée sur la classification binaire. Dans ce cas, une règle de classification avec l'option de rejet prend ses valeurs dans $\{0, 1, re\}$ où re désigne le refus de répondre. Mathématiquement parlant, rappelons rapidement le problème de classification binaire. Soit $(X, Y) \in \mathcal{X} \times \{0, 1\}$ un couple admettant une loi de probabilité inconnue $\mathbb{P}$ sur $\mathcal{X} \times \{0, 1\}$ où $\mathcal{X}$ est un

espace des entrées (*features* en anglais). Le vecteur $X \in \mathcal{X}$ est appelé vecteur des entrées et $Y \in \{0,1\}$ est l'étiquette associée à X. Une observation x peut être difficile à classer si la probabilité conditionnelle $\eta(x) = \mathbb{P}(Y = 1|X = x)$ est proche de $1/2$. Soit $g : \mathcal{X} \to \{0, 1, re\}$ un classifieur avec option rejet où re signifie rejeter. Ce classifieur a deux caractéristiques ; l'erreur $\mathbb{P}(g(X) \neq Y, g(X) \neq re)$ et le taux de rejet fixe $\mathbb{P}(g(X) = re)$. Pour réaliser un compromis entre ces deux mesures de qualité de l'estimateur de g, nous considérons le risque associée suivant

$$\mathbb{P}(g(X) \neq Y, g(X) \neq re) + d\mathbb{P}(g(X) = re) \ .$$

où $d \in]0, 1/2[$ est un paramètre de régularisation qui permet de réaliser ce compromis. Il contrôle le poids que l'on souhaite donner à l'emploi de l'option rejet. Ce paramètre prend une valeur faible. Donc, on aura facilement recours au rejet. La règle optimale minimisant le risque précédent est donnée dans [19]

$$g^*(x) = \begin{cases} 0 & \text{si } 1 - \eta(x) > \eta(x) \text{ et } 1 - \eta(x) > 1 - d, \\ 1 & \text{si } \eta(x) > 1 - \eta(x) \text{ et } \eta(x) > 1 - d, \\ re & \text{si } \max(1 - \eta(x); \eta(x)) \leq 1 - d. \end{cases}$$

Les caractéristiques du classifieur de Bayes avec option rejet sont établies dans [19, 42, 103, 49, 23]. Par exemple, [42] a étudié la vitesse de convergence des estimateurs plug-in et des minimiseurs du risque empirique. En outre, nous pouvons définir deux quantités : la probabilité de rejet et la probabilité de mauvaise classification. [49] se focalise à minimiser la probabilité de rejet à condition d'avoir une borne fixe sur la probabilité de mauvaise classification. À l'inverse, étant donné une borne fixe sur la probabilité de rejet, [23] a étudié le problème minimisation de la probabilité de mauvaise classification. Dans cette thèse, nous nous intéressons au problème de régression avec option rejet.

Régression avec option rejet

Le rejet en régression a rarement été considéré dans la littérature. L'article [100] a analysé la possibilité d'une option rejet dans le contexte de la régression des moindres carrés. Nous considérons un jeu de données étiquetées $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$. Nous notons l'ensemble de toutes les fonctions de prédiction $\mathcal{F}$. Étant donné une fonction de perte, $\ell : \mathbb{R} \times \mathbb{R} \to [0, \infty)$, nous quantifions la qualité de prédiction de tout f par le risque

$$R(f) = \mathbb{E}\left[\ell(f(X), Y)\right] = \int_{\mathbb{R}^d \times \mathbb{R}} \ell(f(x), y) d\mathbb{P}(x, y).$$

La meilleure fonction de prédiction est la fonction de $\mathcal{F}$ minimisant le risque R :

$$f^* \in \operatorname*{argmin}_{f \in \mathcal{F}} R(f).$$

Le risque empirique d'un prédicteur f sur $\mathcal{D}_n$ est défini par

$$\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)).$$

Étant donné un échantillon $\mathcal{D}_n$, le minimiseur du risque empirique (ERM) sur $\mathcal{F}$ est $\hat{f} = \inf_{f \in \mathcal{F}} \hat{R}(f)$.

Un prédicteur avec option rejet est une fonction mesurable de $\mathbb{R}^d$ dans $\mathbb{R} \cup \emptyset$. Dans notre méthodologie, la définition d'un prédicteur avec option rejet s'appuie sur une fonction de prédiction. Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction de prédiction. Un *prédicteur avec option rejet* $\Gamma_f : \mathbb{R}^d \to \mathbb{R} \cup \emptyset$ est une application qui à tout $x \in \mathbb{R}^d$ associe $\Gamma_f(x) \in \{\emptyset, \{f(x)\}\}$. L'utilisation d'un prédicteur avec option rejet offre donc deux possibilités : soit $\Gamma_f(x) = \{f(x)\}$ et on prédit $f(x)$ pour la variable x ; soit $\Gamma_f(x) = \emptyset$ et l'option rejet a été utilisée. Notons que dans ce dernier cas, le cardinal de $\Gamma_f(x)$, noté $|\Gamma_f(x)|$, est nul (et $|\Gamma_f(x)| = 1$ dans le cas de la prédiction). Les auteurs dans [100] définissent la notion de ε-point par point optimal qui demande la condition suivante

$$\forall x \in \{x \in \mathbb{R}^d, |\Gamma_f(x)| = 1\}, \ \ |f(x) - f^*(x)| \leq \varepsilon.$$

Basé sur ERM, les auteurs ont analysé l'option rejet du point de vue de l'ε-optimalité, et ont garanti que la prédiction est à l'intérieur d'une boule de rayon ε autour du fonction de régression avec grande probabilité tout en rejetant seulement une partie limitée du domaine. L'inconvénient de cette approche est que leur méthodologie est associée seulement à des procédures de minimisation du risque empirique. De plus, la calibration de ε est très délicate. Il est naturel de poser la question suivante :

Question 2. Est-il possible de généraliser le problème de régression avec option rejet en utilisant n'importe quelle procédure ?
Nous apporterons les éléments de réponse dans la section 1.6.2.

1.5.2 Intervalles de prédiction

Les intervalles de prédiction (IP) jouent un rôle central en statistique inférentielle. La construction d'IP nécessite une estimation fiable des objets inconnus f^* et σ^2. En pratique, il est important de disposer de mesures précises de l'incertitude des prévisions d'un modèle pour éviter les conclusions erronées.
L'intervalle de prédiction est une plage de valeurs susceptibles d'inclure le paramètre d'intérêt avec un certain degré de confiance. L'estimation en tout point n'est plus une valeur unique, mais une région de prédiction/tolérance.

Bien que la construction d'IP est très ancienne, de nouvelles approches ont vu le jour. Ainsi, la prédiction conforme a fait récemment l'objet de nombreux efforts de recherche en statistique avec d'intéressants développements méthodologiques, informatiques et théoriques depuis le début des années 2000. Elle a été initialement formulée pour la tâche de classification [94, 79, 95] puis elle a été utilisée en régression [50, 51, 52, 84]. Étant donné l'échantillon $\mathcal{D}_n$, l'objectif est de prédire le label Y_{n+1} associée à la nouvelle entrée X_{n+1}. Plus précisément, étant donné un niveau de confiance $\alpha \in [0,1]$, la prédiction conforme permet de fournir un ensemble de confiance
$\mathbf{C}_n(X_{n+1}) = \mathbf{C}_n(X_1, Y_1, \ldots, X_n, Y_n; X_{n+1}) \subseteq \mathbb{R}$ qui satisfait

$$\mathbf{P}\left(Y_{n+1} \in \mathbf{C}_n(X_{n+1})\right) \geq 1 - \alpha, \tag{1.9}$$

où $\mathbf{P} = \mathbb{P}^{n+1}$ est la distribution jointe de $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$. Nous présentons les différentes notions de validité dans [52] :
Validité marginale. C'est la notion la plus classique pour les IP. L'équation (1.9) signifie que l'ensemble $\mathbf{C}_n$ a la garantie de validité marginale. Soit ν la mesure de Lebesgue. L'ensemble de prédiction optimal de niveau $1-\alpha$ est

$$\mathbf{C}^\alpha = \underset{\mathbf{C}:\mathbb{P}((X,Y)\in\mathbf{C})\geq 1-\alpha}{\operatorname{argmin}} \nu(\mathbf{C}) := \{(x,y) : p(x,y) \geq t_\alpha\},$$

où t_α satisfait $\mathbb{P}\left(\mathbf{C}^\alpha\right) = 1-\alpha$ et $p(X,Y)$ est la densité jointe de (X,Y). La validité marginale n'est pas conditionnée par X_{n+1}. Dans ce cas, il est possible d'avoir des garanties plus solides par
Validité conditionnelle. La validité conditionnelle est donnée par :

$$\mathbf{P}\left(Y_{n+1} \in \mathbf{C}_n(x) | X_{n+1} = x\right) \geq 1 - \alpha \text{ pour tout } \mathbb{P} \text{ et presque } x.$$

Elle implique la validité marginale mais pas l'inverse (voir [52, 55] pour plus de détails). Un ensemble de prédiction est défini par

$$C^*_{\mathbb{P}}(x) = \inf_x \mathbb{P}\left(Y \in \mathbf{C}(x) | X = x\right) \geq 1 - \alpha,$$

Si la loi jointe de (X,Y) est connue, son expression est donnée comme suit

$$C^*_{\mathbb{P}}(x) = \{y : p(y|x) \geq t_\alpha(x)\},$$

où $t_\alpha(x)$ est choisi de cette manière : $\int_{\mathbb{R}} p(y|x)\mathbb{1}_{\{p(y|x)\geq t_\alpha(x)\}}dy = 1-\alpha$. L'ensemble $C_{\mathbb{P}} = \{C^*_{\mathbb{P}}(x)\}$ est appelé la bande oracle conditionnelle qui minimise $\nu(\mathbf{C}(x))$ pour tout x parmi toutes les bandes $C^*_{\mathbb{P}}$.

En général, la validité conditionnelle est obtenue que sur des hypothèses fortes sur la distribution de (X, Y). La méthode proposée pour construire un ensemble de confiance est
Validation locale [52, 55]. Elle consiste à interpoler entre la validité marginale et la validité conditionnelle. Étant donné une partition $\mathcal{A} = \{A_k, 1 \leq k \leq m\}$ de $\mathbb{R}^d$ de sorte que $\cup_{k=1}^m A_k = \mathbb{R}^d$, un ensemble de confiance vérifie la validation locale si

$$\mathbf{P}\left(Y_{n+1} \in \mathbf{C}_n(x) | X_{n+1} \in A_k\right) \geq 1 - \alpha \text{ pour tout } \mathbb{P} \text{ et } 1 \leq k \leq m.$$

La prédiction conforme est bien détaillée dans les articles [94, 79, 52]. Nous notons $Z_{n+1} = (X_{n+1}, Y_{n+1})$ et $z = (x, y)$. Nous formons un jeu de données augmenté $\mathcal{D}_{n+1,z}$ correspondant à $\mathcal{D}_{n+1}$ avec $Z_{n+1} = z$. Nous définissons un score de conformité $r_i(z) = g(Z_i, \mathcal{D}_{n+1,z})$ pour $i = 1, \ldots, n$ où g est une fonction mesurable basé sur $\mathcal{D}_{n+1,z}$. Puis, nous faisons l'hypothèse $H_0 : Y_{n+1} = y$ en utilisant la fonction

$$\pi_n(x, y) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{1}_{\{r_i(z) \leq r_{n+1}(z)\}} \ .$$

Enfin, le prédicteur de confiance est

$$\mathbf{C}_n(x) = \{y : (n+1)\pi_n(x, y) \geq \lfloor (n+1)\alpha \rfloor\}.$$

Dans [52], les auteurs ont étudié ce prédicteur en prenant g comme un estimateur à noyau de la jointe (X, Y). L'IP résultant est marginalement valide. Mais, il est coûteux en temps de calcul : nous devons trouver le score de conformité $r(z)$ pour chaque z. Les auteurs ont construit des bandes de prédiction avec une validation locale. Ils ont proposé un estimateur appelé COPS (en anglais *conformal optimized prediction set*) qui a toujours une garantie d'échantillon fini. Sous des conditions de régularité, cet estimateur converge vers une bande oracle à une vitesse optimale au sens minimax d'ordre $(\log(n)/n)^{\beta/(\beta(d+2)+1)}$, où β est la régularité de la classe de Holder. En grande dimension ($d >> n$), [50] propose une stratégie pour construire un prédicteur de confiance pour la variable de réponse en utilisant n'importe quel estimateur de la fonction de régression.

1.6 Contributions

Les travaux présentés dans ce manuscrit sont regroupés en trois chapitres. Le premier chapitre donne une possible réponse à la **Question 1**. Il introduit une stratégie d'agrégation pour l'estimation de la fonction de variance. Le deuxième chapitre représente le problème de régression avec option rejet où l'on est autorisé à s'abstenir de prédire. Il décrit une réponse de la **Question 2**. Le troisième chapitre détaille la construction d'intervalle de prédiction en prédéterminant sa longueur. Tous ces travaux concernent le modèle de régression hétéroscédastique.

1.6.1 Estimation de la fonction de variance

Nous commençons cette section par une contribution à l'étude théorique de l'estimateur de K-PPV.

Estimation de la fonction de variance par K-PPV

Nous avons décrit dans les sections précédentes l'estimation des fonctions de régression et de variance par l'algorithme de K-PPV. Ici, nous présentons un nouveau résultat sur les vitesses de convergence de l'estimateur (1.4) et de l'estimateur (1.8) par rapport à la ℓ_p-norme sup.
Nous faisons tout d'abord les hypothèses suivantes :

Hypothèse 1. *Les fonctions f^* et σ^2 sont lipschitziennes.*

Hypothèse 2 (Hypothèse de forte densité). *La distribution marginale $\mathbb{P}_X$ satisfait l'hypothèse de forte densité si*

- *$\mathbb{P}_X$ est supporté sur un ensemble régulier compact $\mathcal{C} \subset \mathbb{R}^d$,*
- *$\mathbb{P}_X$ admet une densité μ w.r.t. à la mesure de Lebesgue telle que pour tout $x \in \mathcal{C}$, nous avons $0 < \mu_{\min} \leq \mu(x) \leq \mu_{\max} < \infty$.*

Hypothèse 3. *Les variables $Y - f^*(X)$ et $(Y - f^*(X))^2 - \sigma^2(X)$ satisfont une condition de bruit uniforme : il existe $c_0 > 0$ tel que*

$$\sup_{x \in \mathcal{C}} \mathbb{E}\left[\exp(\lambda\,(Y - f^*(X))) \mid X = x\right] \leq \exp(c_0^2 \lambda^2), \quad \text{pour} \quad |\lambda| \leq \frac{1}{c_0} \ .$$

Ces trois hypothèses sont assez classiques [36, 86]. L'hypothèse de forte densité a été introduite dans le contexte de la classification binaire par exemple dans [2]. L'hypothèse 3 exige que le conditionnel sur X soit sous-exponentielle uniformément sur $\mathcal{C}$.

Théorème 1. *Supposons que les trois hypothèses précédentes sont satisfaites. Soit $p \geq 1$ et $k_n \propto n^{2/(d+2)}$, alors*

$$\mathbb{E}\left[\left(\sup_{x \in \mathcal{C}} \left|\hat{f}_{knn}(x) - f^*(x)\right|\right)^p\right] \lesssim \log(n)^p n^{-p/(d+2)}, \ \ \mathbb{E}\left[\left(\sup_{x \in \mathcal{C}} \left|\hat{\sigma}^2(x) - \sigma^2(x)\right|\right)^p\right] \lesssim \log(n)^p n^{-p/(d+2)} \ .$$

Ce résultat présente les vitesses de convergence pour les estimateurs des fonctions de régression et de variance par K-PPV par rapport à la ℓ_p-norme sup. Nous remarquons que les deux estimateurs ont la même vitesse de convergence qui est d'ordre $\log(n)^p n^{-p/(d+2)}$ et qui dépend de la dimension d et du paramètre p. En effet, la borne supérieure de la ℓ_p-norme sup de l'estimateur $\hat{\sigma}^2$ dépend de la borne supérieure de la ℓ_p-norme sup de l'estimateur $\hat{f}_{knn}$.

Estimation de la fonction de variance par les procédures d'aggrégation

Nous rappelons que l'agrégation est une approche populaire en apprentissage statistique pour estimer f^* dans le modèle de régression (Section 1.3). Dans ce sens, nous développons l'agrégation pour estimer la fonction de variance σ^2 dans laquelle les estimateurs initiaux sont construits par la méthode basée sur les erreurs résiduelles (Section 1.4). Nous introduisons d'abord deux échantillons d'apprentissage indépendants : $\mathcal{D}_n = \{(X_i, Y_i), i = 1, \ldots, n\}$ et $\mathcal{D}_N = \{(X_i, Y_i), i = n+1, \ldots, n+N\}$ qui consistent respectivement en n et N copies i.i.d. de (X, Y). Nous décrivons l'algorithme d'estimation de la fonction de variance σ^2 en utilisant deux types d'agrégation : agrégation de type sélection de modèle et l'agrégation convexe. Les estimateurs résultants sont appelés `MS`-estimateur et `C`-estimateur respectivement.

Agrégation de type sélection modèle. La méthode que nous proposons est en deux étapes et est décrite dans l'Agorithme 1.

Algorithme 1 : *MS-estimateur de la fonction de variance*

Étape 1 : Estimons f^* par MS

1. Soit $M_1 \geq 2$ fixé. Nous considérons M_1 estimateurs de la fonction de régression $\hat{f}_1, \ldots, \hat{f}_{M_1}$ basée sur $\mathcal{D}_n$.

2. Nous sélectionnons l'indice optimal $\hat{s}$

$$\hat{s} \in \underset{s\in[M_1]}{\operatorname{argmin}} \hat{\mathcal{R}}_N(\hat{f}_s), \text{ avec } \hat{\mathcal{R}}_N(\hat{f}_s) = \frac{1}{N}\sum_{i=n+1}^{n+N} |Y_i - \hat{f}_s(X_i)|^2 \ .$$

3. Le MS-estimateur de la fonction de régression, noté $\hat{f}_{\text{MS}}$, est donné comme suit

$$\hat{f}_{\text{MS}} := \hat{f}_{\hat{s}}.$$

Étape 2 : Estimons σ^2 par MS

4. Étant donné l'estimateur $\hat{f}_{\text{MS}}$ construit sur $\mathcal{D}_N$, nous construisons M_2 estimateurs de la fonction de variance $\hat{\sigma}^2_{\hat{s},1}, \ldots, \hat{\sigma}^2_{\hat{s},M_2}$, construits à partir de $\mathcal{D}_n$ par **la méthode basée sur les erreurs résiduelles** avec $2 \leq M_2 < \infty$.

5. Sur la base de $\mathcal{D}_N$, nous sélectionnons l'indice optimal $\hat{m}$ comme suit

$$\hat{m} \in \underset{m\in[M_2]}{\operatorname{argmin}} \hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) \text{ où } \hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) = \frac{1}{N}\sum_{i=n+1}^{n+N} | \left(Y_i - \hat{f}_{\text{MS}}(X_i)\right)^2 - \hat{\sigma}^2_{\hat{s},m}(X_i)|^2 \ .$$

6. Sortie le MS-estimateur de la fonction de variance : $\hat{\sigma}^2_{\text{MS}} := \hat{\sigma}^2_{\hat{s},\hat{m}}$.

Agrégation convexe. Notre méthode se déroule toujours en deux étapes : une pour l'estimation de la fonction de régression f^* et la seconde pour l'estimation de la fonction de variance σ^2 (*c.f.*, Algorithme 2). Les différences avec l'agrégation C résident dans les étapes 2-3 et 5-6 définissant les poids pour les fonctions à agréger.

Résultats principaux. Dans la suite, nous décrivons la consistance de nos approches.

Soit $\mathcal{R}(\hat{f}_s) = \mathbb{E}\left[|Y - \hat{f}_s(X)|^2\right]$ le risque quadratique pour $\hat{f}_s$ pour tout $s \in [M_1]$. Nous définissons s^* comme suit

$$s^* \in \underset{s\in[M_1]}{\operatorname{argmin}} \mathcal{R}(\hat{f}_s) \ .$$

Nous introduisons également les hypothèses suivantes :

Hypothèse 4. *Les fonctions f^* et σ^2 sont bornées.*

Hypothèse 5. *Pour tout $s \in [M_1]$ et tout $m \in [M_2]$, $\hat{f}_s$ et $\hat{\sigma}^2_{s,m}$ sont bornés pour presque tout $\mathcal{D}_n$.*

Hypothèse 6 (Hypothèse de séparabilité)**.** *Il existe $\delta_0 > 0$ telle que*

$$\delta^*(\mathcal{D}_n) = \min_{s\neq s^*}\{|\mathcal{R}(\hat{f}_s) - \mathcal{R}(\hat{f}_{s^*})|\} > \delta_0 \ .$$

Hypothèse 7. *Y est borné ou Y provient du modèle gaussien*

$$Y = f^*(X) + \sigma(X)\xi,$$

où $\xi \sim \mathcal{N}(0,1)$ est indépendante de X.

Hypothèse 8. *Pour tout $i \in [M_1]$, tout $m \in [M_2]$ et tout $\lambda \in \Lambda^{M_1}$, $\hat{f}_i$ et $\hat{\sigma}^2_{\lambda,m}$ sont bornés pour presque tout $\mathcal{D}_n$.*

Hypothèse 9. *Il existe une constante $K \geq 0$ telle que pour chaque $j \in [M_2]$*

$$\mathbb{E}\left[|\hat{\sigma}^2_{\lambda_1,j}(X) - \hat{\sigma}^2_{\lambda_2,j}(X)|\right] \leq K\|\lambda_1 - \lambda_2\|_{1,M_1}, \quad \forall \lambda_1, \lambda_2 \in \Lambda^{M_1} \text{ presque sûrement } \mathcal{D}_n.$$

Les hypothèses 1 jusqu'à 9 jouent un rôle crucial sur l'étude de la consistance de $\hat{\sigma}^2_{\text{MS}}$ et de $\hat{\sigma}^2_{\text{C}}$ respectivement. Les hypothèses 5 et 8 décrivent la bornitude des estimateurs initiaux qui sont construits sur l'échantillon $\mathcal{D}_n$. Les bornes de ces estimateurs ne sont pas nécessairement connues. Elles jouent un rôle clé dans la détermination de la vitesse de convergence de l'erreur d'estimation (la variance) de $\hat{\sigma}^2_{\text{MS}}$ et de $\hat{\sigma}^2_{\text{C}}$. L'hypothèse 6 permet de contrôler la différence entre la variable aléatoire $\hat{s}$ et le sélecteur déterministe s^* avec grande probabilité. L'hypothèse 9 est une hypothèse forte. Elle est vérifiée par exemple pour les estimateurs de moyenne locale.

Algorithme 2 *C-estimateur de la fonction de variance*

Étape 1 : Estimons f^* par C

1. Soit $M_1 \geq 2$ fixé. Nous considérons M_1 estimateurs de la fonction de régression $\hat{f}_1, \ldots, \hat{f}_{M_1}$ basée sur $\mathcal{D}_n$.
2. Pour tout $\lambda \in \Lambda^{M_1}$, nous définissons la combinaison linéaire $\hat{f}_\lambda$: $\hat{f}_\lambda = \sum_{j=1}^{M_1} \lambda_j \hat{f}_j$. Basée sur $\mathcal{D}_N$, nous déterminons
$$\hat{\lambda} \in \underset{\lambda \in \Lambda^{M_1}}{\operatorname{argmin}} \hat{\mathcal{R}}_N(\hat{f}_\lambda).$$
3. Le C-estimateur de la fonction de régression est donné comme suit
$$\hat{f}_{\text{C}} := \hat{f}_{\hat{\lambda}} = \sum_{j=1}^{M_1} \hat{\lambda}_j \hat{f}_j \ ,$$

Étape 2 : Estimons σ^2 par C

4. Étant donné l'estimateur $\hat{f}_{\text{C}}$ construit sur $\mathcal{D}_N$, nous construisons M_2 estimateurs de la fonction de variance $\hat{\sigma}^2_{\hat{\lambda},1}, \ldots, \hat{\sigma}^2_{\hat{\lambda},M_2}$ construits à partir de $\mathcal{D}_n$ par la méthode basée sur les erreurs résiduelles avec $M_2 \geq 2$ fixé.
5. Pour tout $\beta \in \Lambda^{M_2}$ nous définissons $\hat{\sigma}^2_{\hat{\lambda},\beta}$: $\hat{\sigma}^2_{\hat{\lambda},\beta} = \sum_{j=1}^{M_2} \beta_j \hat{\sigma}^2_{\hat{\lambda},j}$. Sur $\mathcal{D}_N$, nous calculons
$$\hat{\beta} \in \underset{\beta \in \Lambda^{M_2}}{\operatorname{argmin}} \hat{R}_{\mathbf{N}}(\hat{\sigma}^2_{\hat{\lambda},\beta}), \text{ où } \hat{R}_N(\hat{\sigma}^2_{\hat{\lambda},\beta}) = \frac{1}{N} \sum_{i=n+1}^{n+N} |(Y_i - \hat{f}_{\text{C}}(X_i))^2 - \hat{\sigma}^2_{\hat{\lambda},\beta}(X_i)|^2 \ .$$
6. Sortie le C-estimateur de la fonction de variance : $\hat{\sigma}^2_{\text{C}} := \hat{\sigma}^2_{\hat{\lambda},\hat{\beta}}$.

Nous obtenons le premier résultat :

Théorème 2. *Soit $\hat{f}_{MS}$ et $\hat{\sigma}^2_{MS}$ deux MS-estimateurs de f^* et de σ^2 respectivement. Alors, il existe deux constantes absolues $C_1 > 0$ et $C_2 > 0$ tel que*

$$\mathbb{E}\left[|\hat{\sigma}^2_{MS}(X) - \sigma^2(X)|^2\right] \leq \mathbb{E}\left[\min_{m \in [M_2]} \mathbb{E}_X\left[|\hat{\sigma}^2_{s^*,m}(X) - \sigma^2(X)|^2\right]\right] + C_1 \left\{\min_{s \in [M_1]} \mathbb{E}\left[\|\hat{f}_s - f^*\|^2_N\right]\right\}^{1/2p} + C_2\left(\frac{\log(M_1)}{N}\right)^{1/4p},$$

avec $p = 1$ si Y est borné et $p = 2$ si Y provient du modèle gaussien.

Ce Théorème donne la borne supérieure de l'erreur quadratique de $\hat{\sigma}^2_{\texttt{MS}}$. En résumé, le MS-estimateur est asymptotiquement aussi bon que le meilleur estimateur $\hat{\sigma}^2_{s^*,m}$ qui est construit sur le premier échantillon $\mathcal{D}_n$ au sens de l'erreur quadratique L^2. La borne se compose de deux parties : la première partie est le biais du MS-estimateur $\hat{\sigma}^2_{\texttt{MS}}$ et dépend du sélecteur déterministe s^* ; la seconde partie est composée des deux termes restants et correspond à l'erreur d'estimation. Le premier terme est le terme de biais de $\hat{f}_{\texttt{MS}}$ exprimé en fonction de la norme empirique, et le deuxième terme caractérise le prix à payer pour l'agrégation MS qui est d'ordre ; $(\log(M_1)/N)^{1/4}$ dans le cas où Y est borné et $(\log(M_1)/N)^{1/8}$ dans le cas où Y provient du modèle gaussien. Cette vitesse lente est du au fait que l'estimation de la fonction de variance repose sur f^* que l'on doit également estimer.

Le deuxième résultat décrit la consistance de $\hat{\sigma}^2_{\texttt{C}}$.

Théorème 3. *Soit $\hat{f}_C$ et $\hat{\sigma}^2_C$ deux C-estimateurs de f^* et de σ^2 respectivement. Alors, il existe deux constantes absolues $C_1 > 0$ et $C_2 > 0$ tel que*

$$\mathbb{E}\left[|\hat{\sigma}^2_C(X) - \sigma^2(X)|^2\right] \leq \mathbb{E}\left[\inf_{\beta\in\Lambda^{M_2}} \mathbb{E}_X\left[|\hat{\sigma}^2_{\hat{\lambda},\beta}(X) - \sigma^2(X)|^2\right]\right] + C_1\left\{\inf_{\lambda\in\Lambda^{M_1}} \mathbb{E}\left[\|\hat{f}_\lambda - f^*\|^2_N\right]\right\}^{1/2p} + C_2\left(\frac{\log(M_1)}{N}\right)^{1/4p},$$

avec $p = 1$ si Y est borné et $p = 2$ si Y provient du modèle gaussien.

Ce résultat décrit la borne supérieure de l'erreur quadratique de $\hat{\sigma}^2_{\texttt{C}}$. La borne se décompose en trois termes. Le premier terme représente le biais de $\hat{\sigma}^2_{\texttt{C}}$ qui dépend de la variable aléatoire $\hat{\lambda}$. Le deuxième et le troisième terme sont une borne du terme de variance de $\hat{\sigma}^2_{\texttt{C}}$. Le deuxième terme est le terme de biais de $\hat{f}_{\texttt{C}}$ par rapport à la norme empirique. Le troisième terme est le prix à payer pour l'agrégation C.

Les deux procédures MS et C ont la même vitesse de convergence. En effet, les termes de la variance de $\hat{\sigma}^2_{\texttt{MS}}$ et $\hat{\sigma}^2_{\texttt{C}}$ sont basés sur la borne supérieure de l'erreur quadratique empirique de $\hat{f}_{\texttt{MS}}$ et $\hat{f}_{\texttt{C}}$. De plus, $\hat{f}_{\texttt{MS}}$ et $\hat{f}_{\texttt{C}}$ ont la même vitesse de convergence (voir [88] pour plus de détails). En comparaison aux vitesses de convergence de $\hat{f}_{\texttt{MS}}$ et $\hat{f}_{\texttt{C}}$ au sens de l'erreur quadratique, nos vitesses sont lentes à cause de la double agrégation que nous devons effectuer pour l'estimation de la fonction de variance conditionnelle. L'optimalité de ces résultats reste un problème ouvert.

1.6.2 Régression avec option rejet et application au K-PPV

Dans le Chapitre 3, nous décrivons une première application de la fonction de variance conditionnelle. Nous introduisons le problème de régression avec option rejet qui vise à construire des procédures d'estimation capables de ne pas prédire lorsque le doute dans la valeur prédite est trop grand. En particulier, nous nous intéressons au cadre de travail dans lequel le taux de rejet est fixe, laissant ainsi la possibilité à l'humain d'agir sur une proportion des données jugées, par l'algorithme, trop délicate à traiter par l'algorithme lui-même. Cette décision de rejeter et donc de ne pas traiter certaines données par la machine est prise par l'algorithme et vise à rentabiliser l'effort de l'humain.

Nous avons présenté dans la Section 1.5.1 la définition d'un prédicteur avec option rejet Γ_f. Nous rappelons que ce prédicteur est une fonction mesurable de $\mathbb{R}^d$ dans $\mathbb{R} \cup \emptyset$. Il a deux caractéristiques :

- le taux de rejet : $r(\Gamma_f) = \mathbb{P}(|\Gamma_f(X)| = 0)$;
- l'erreur de prédiction : $\mathrm{Err}(\Gamma_f) = \mathbb{E}\left[(Y - f(X))^2 \mid |\Gamma_f(X)| = 1\right]$.

Pour mesurer la performance de Γ_f, nous considèrons un compromis entre ces deux quantités. Dans ce cas, le risque de Γ_f est donné par

$$\mathcal{R}_\lambda(\Gamma_f) = \mathbb{E}\left[(Y - f(X))^2 \mathbb{1}_{\{|\Gamma_f(X))|=1\}}\right] + \lambda\,\mathbb{P}(|\Gamma_f(X)| = 0),$$

où $\mathbb{1}_{\{\cdot\}}$ est la fonction indicatrice et $\lambda > 0$ est un paramètre de régularisation. En particulier, une forte valeur de λ impose une contrainte importante sur l'utilisation de l'option rejet.

Inspiré de la classification avec option rejet [23], nous présentons dans cette section le cadre d'étude où le prédicteur avec option rejet a un taux de rejet fixé à l'avance. Ceci traduit typiquement des situations où nous disposons de ressources humaines limitées pour traiter toutes les tâches. Pour ce faire, nous ajoutons une contrainte sur le taux de rejet du prédicteur. Plus précisément, soit $\varepsilon \in (0, 1]$. Le *prédicteur optimal à taux de rejet* ε, ou le *ε-prédicteur optimal* pour raccourcir est défini par :

$$\Gamma_\varepsilon^* \in \arg\min\{\mathrm{Err}(\Gamma_f) \;:\; r(\Gamma_f) \leq \varepsilon\}\ .$$

Notamment, sous l'hypothèse de continuité de la fonction de répartition F_{σ^2} de la variable aléatoire $\sigma^2(X)$, le problème d'optimisation précédent admet une solution explicite

$$\Gamma_\varepsilon^*(X) = \begin{cases} \{f^*(X)\} & \text{if } F_{\sigma^2}(\sigma^2(X)) \leq 1 - \varepsilon\ , \\ \emptyset & \text{sinon}\ . \end{cases}$$

Le ε-prédicteur optimal repose donc sur un seuillage de la fonction de variance. Nous pouvons montrer par ailleurs qu'il a exactement un taux de rejet égal à ε. Autrement dit, le fait de rejeter ou non dépend essentiellement de la fonction de variance. La nature de l'oracle suggère une approche semi-supervisée de type plug-in qui repose sur l'estimation de f^* et σ^2, ainsi que F_{σ^2}. L'algorithme 3 détaille la construction du *prédicteur plug-in de niveau de rejet* ε.

Algorithme 3 *Prédicteur plug-in de niveau ε*

1. Nous introduisons deux échantillons indépendants : $\mathcal{D}_n$ et $\mathcal{D}_M = \{X_i\}_{i=n+1}^{n+M}$ qui contient M observations indépendantes, et *non étiquetées*, de même loi que X.
2. À partir de $\mathcal{D}_n$, nous construisons des estimateurs $\hat{f}$ et $\hat{\sigma}^2$ de f^* et σ^2.
3. Sur $\mathcal{D}_M$, nous estimons F_{σ^2} :
   (a) Randomisation :
   $$\hat{\hat{\sigma}}^2(X_i) = \hat{\sigma}^2(X_i) + \zeta_i, \qquad \zeta_i \sim \mathcal{U}([0, u]) \text{ avec } u > 0,$$
   (b) Fonction de répartition empirique de $\hat{\hat{\sigma}}^2(X_i)$: $\hat{F}_{\hat{\hat{\sigma}}^2}(\cdot) = \frac{1}{M}\sum_{i=n+1}^{n+M} \mathbb{1}_{\{\hat{\hat{\sigma}}^2(X_i) \leq \cdot\}}$.
4. Sortie du *prédicteur plug-in de niveau ε*

$$\hat{\Gamma}_\varepsilon(x) = \begin{cases} \left\{\hat{f}(x)\right\} & \text{if } \hat{F}_{\hat{\hat{\sigma}}^2}(\hat{\sigma}^2(x) + \zeta) \leq 1 - \varepsilon\ , \\ \emptyset & \text{sinon}\ . \end{cases}$$

L'étape 3(a) de randomisation n'a pas d'effet sur la qualité de prédiction. Elle a pour objet d'assurer la continuité de la fonction de distribution de $\hat{\hat{\sigma}}^2$ et donc d'obtenir un taux de rejet souhaité. De plus, la consistance de $\hat{\sigma}^2$ implique la consistance $\hat{\hat{\sigma}}^2$ si u tend vers 0.
L'excès de risque de tout prédicteur avec option rejet est défini comme la distance suivante :

$$\mathcal{E}_{\lambda_\varepsilon}(\Gamma) = \mathcal{R}_{\lambda_\varepsilon}(\Gamma) - \mathcal{R}_{\lambda_\varepsilon}(\Gamma_\varepsilon^*).$$

Le résultat suivant est le central de ce chapitre et décrit la consistance de notre approche.

Théorème 4. *Soit $\varepsilon \in (0,1)$ et $\lambda_\varepsilon = F^{-1}_{\sigma^2}(1-\varepsilon)$. Supposons que σ^2 est borné, $\hat{f}$ est un estimateur consistant de f^* par rapport au risque L_2 et $\hat{\sigma}^2$ est un estimateur consistant de σ^2 par rapport au risque L_1. Sous des hypothèses de regularité, nous avons*

$$\mathbb{E}\left[\mathcal{E}_{\lambda_\varepsilon}\left(\hat{\Gamma}_\varepsilon\right)\right] \underset{n,M\to+\infty}{\longrightarrow} 0, \quad et \quad \left|\mathbb{E}\left[r(\hat{\Gamma}_\varepsilon)\right] - \varepsilon\right| \lesssim M^{-1/2} \ .$$

Ce résultat prouve que le prédicteur $\hat{\Gamma}_\varepsilon$ est asymptotiquement aussi bon que le prédicteur optimal et de niveau de rejet ε. La convergence de l'excès de risque du $\hat{\Gamma}_\varepsilon$ nécessite que deux estimateurs consistants $\hat{f}$ et $\hat{\sigma}^2$. En particulier, le théorème montre que le taux de rejet du $\hat{\Gamma}_\varepsilon$ est de niveau de rejet ε jusqu'à un terme d'ordre $M^{-1/2}$. Dans la suite, nous considérons le cas où ces estimateurs sont obtenus en utilisant l'algorithme des K-plus proches voisins (K-PPV). Nous déterminons la vitesse de convergence de l'excès de risque de cet estimateur sous des hypothèses techniques (voir Chapitre 3) comme par exemple la condition α-marge qui garantit que la variable aléatoire $\sigma^2(X)$ ne puisse pas être trop proche du seuil λ_ε.

Théorème 5. *Soit $\varepsilon \in (0,1)$ et $\alpha > 0$. Si $k_n \propto n^{2/(d+2)}$ et $u \leq n^{-1/(d+2)}$, alors le ε-prédicteur basé sur K-PPV $\hat{\Gamma}_\varepsilon$ satisfait*

$$\mathbb{E}\left[\mathcal{E}_{\lambda_\varepsilon}\left(\hat{\Gamma}_\varepsilon\right)\right] \lesssim n^{-2/(d+2)} + \log(n)^{(\alpha+1)} n^{-(\alpha+1)/(d+2)} + M^{-1/2} \ .$$

Ce résultat porte plusieurs enseignements. Chaque terme de la vitesse de convergence ci-dessus décrit une caractéristique donnée du problème. La première repose sur l'erreur L^2 d'estimation de la fonction de régression. La seconde, qui dépend du paramètre α, est due à l'erreur d'estimation en norme sup de la fonction de variance. Le dernier terme est directement lié à l'estimation du seuil λ_ε. Dans le cas $\alpha = 1$, le deuxième terme est du même ordre que le terme correspondant à l'estimation de f^*. Enfin, pour $\alpha > 1$ et si la taille de l'échantillon des données non étiquetées M est suffisamment grande, alors la vitesse de l'excès de risque est la même que la vitesse de convergence de $\hat{f}$ par rapport à l'erreur L^2. Cette dernière est alors la meilleure situation à laquelle nous pouvons nous attendre pour le paramètre de rejet.

1.6.3 Intervalle de prédiction

Dans le dernier chapitre, nous présentons une deuxième application de l'estimation de la fonction de variance. Nous abordons ici une nouvelle approche pour construire un intervalle de prédiction en donnant sa longueur a priori dans le modèle gaussien hétéroscédastique.

Nous supposons que le bruit ε suit une loi gaussienne centrée réduite. En particulier, la variable $Y|X$ suit une loi gaussienne de moyenne $f^*(X)$ et de variance $\sigma^2(X)$. Sa densité est donnée comme suit :

$$p(y|X) = \frac{1}{\sqrt{2\pi\sigma^2(X)}} \exp\left(-\frac{(y-f^*(X))^2}{2\sigma^2(X)}\right) \ .$$

Nous faisons les hypothèses suivantes :

Hypothèse 10. *Il existe $0 < \sigma_0 < \sigma_1 < \infty$ tel que pour tout $x \in \mathbb{R}^d$: $\sigma_0 \leq \sigma(x) \leq \sigma_1$.*

Cette condition est faible et garantit que la densité conditionnelle $p(y|X)$ est bornée. Nous supposons de plus que $p(y|X)$ est sans atome, c'est-à-dire

Hypothèse 11. *Pour tout $y \in \mathbb{R}$, l'application $t \mapsto \mathbb{P}_X(p(y|X) \geq t)$ est continue sur $\mathbb{R}_+^*$.*

Étant donné une observation $X \in \mathbb{R}^d$, notre objectif est de produire une plage de valeurs la plus précise possible où se trouve l'étiquette correspondante $Y \in \mathbb{R}$. Soit $\Gamma : \mathbb{R}^d \to \mathbb{R}$ un intervalle de prédiction. Il a deux caractéristiques :

- l'erreur : $\mathbb{P}\left(Y \notin \Gamma(X)\right)$;

- le taux de la longueur : $\mathcal{L}(\Gamma) := \mathbb{E}_X\left[L(\Gamma(X)\right] = \mathbb{E}_X\left[\int_{\mathbb{R}} \mathbb{1}_{\{y\in\Gamma(X)\}}\mathrm{d}y\right]$ où $L(\cdot)$ représente la longueur de l'intervalle.

Nous nous intéressons à minimiser l'erreur de l'intervalle de prédiction en ajoutant une contrainte sur le taux de la longueur du $\Gamma(X)$. Nous fixons $\ell > 0$. Le ℓ-intervalle de prédiction optimal est :

$$\Gamma^*_\ell \in \arg\min\{\mathbb{P}\left(Y \notin \Gamma(X)\right) \ : \ \Gamma \text{ tel que } \mathcal{L}(\Gamma) \leq \ell\} \ .$$

Concernant ce problème d'optimisation : Est-ce que Γ^*_ℓ admet une expression explicite ? La réponse est oui. D'abord, nous introduisons la quantité suivante : $\forall t > 0$

$$G(t) := \int_{\mathbb{R}} \mathbb{P}_X(p(y|X) \geq t)dy.$$

La fonction G est décroissante sur $\mathbb{R}^*_+$. Elle est continue grâce à l'hypothèse 11. Enfin, la formule explicite de Γ^*_ℓ est donnée comme suit

$$\Gamma^*_\ell(X) = \{y : p(y|X) \geq \lambda_\ell\}, \ \text{ avec } \ \lambda^*_\ell := G^{-1}(\ell)$$

où G^{-1} est la pseudo-inverse de G[2]. Le ℓ-intervalle de prédiction optimal repose sur un seuillage de la densité conditionnelle. Son taux de la longueur est égal ℓ. Pour mesurer la performance d'intervalle Γ, nous considérons la mesure de risque suivante

$$R_\ell(\Gamma) = \mathbb{P}\left(Y \notin \Gamma(X)\right) + \lambda^*_\ell \mathcal{L}(\Gamma) \ ,$$

où le paramétre λ^*_ℓ contrôle l'équilibre entre les deux termes.
Comme $p(y|X)$ et λ^*_ℓ sont inconnues, nous estimons Γ^*_ℓ par algorithme semi-supervisé de type plug-in : Algorithme 4.

Algorithme 4 *Intervalle de prédiction plug-in de longueur ℓ*

1. Nous introduisons deux échantillons indépendants : $\mathcal{D}_n$ et $\mathcal{D}_M$
2. À partir de $\mathcal{D}_n$, nous construisons un estimateur $\tilde{p}(y|x)$ de $p(y|x)$ par la règle plug-in, c'est-à-dire, nous construisons des estimateurs $\hat{f}$ et $\hat{\sigma}^2$ de f^* et de σ^2 respectivement

$$\tilde{p}(y|x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2(x)}} \exp\left(-\frac{(y-\hat{f}(x))^2}{2\hat{\sigma}^2(x)}\right) \ .$$

3. Sur $\mathcal{D}_M$, nous estimons λ^*_ℓ, en particulier, nous calculons la fonction empirique de G :

$$\tilde{G}(t) = \int_{\mathbb{R}} \frac{1}{M}\sum_{i=1}^{M} \mathbb{1}_{\{\tilde{p}(y|X_i)\geq t\}}dy \ .$$

4. Sortie d'*intervalle de prédiction plug-in de longueur ℓ*

$$\tilde{\Gamma}_\ell(x) = \{y \in \mathbb{R} : \tilde{p}(y|x) \geq \tilde{G}^{-1}(\ell)\} \ .$$

Comme la variable Y n'est pas bornée, l'étude des propriétés théoriques de l'estimateur $\tilde{\Gamma}_\ell$ pourrait être difficile. Dans ce contexte, nous faisons les modifications suivantes sans perte de généralité :

- **Seuillage.** Soit $s > 0$. Nous seuillons l'estimateur $\tilde{p}$

$$\hat{p}(y|x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2(x)}} \exp\left(-\frac{(y-\hat{f}(x))^2}{2\hat{\sigma}^2(x)}\right) \mathbb{1}_{\{|y|\leq s\}} \ .$$

[2]Si $t = 0$, nous avons $G(0) = +\infty$ et nous utiliserons la convention $G^{-1}(+\infty) = 0$.

- **Randomisation.** Nous introduisons une perturbation aléatoire ζ distribuée selon une distribution uniforme sur $[0, u]$, pour $u > 0$ et indépendant de (X, Y) et nous considèrons

$$\hat{p}\left(y|X,\zeta\right) = \hat{p}(y|X) + \zeta \mathbb{1}_{\{|y|\leq s\}} \ .$$

  Cette étape assure la continuité de l'application $t \mapsto \mathbb{P}_X(\hat{p}(y|X) \geq t)$ sur $\mathbb{R}_+^*$ pour tout $|y| \leq s$.

- **Discrétisation.** Nous approximons $\tilde{G}$ par la somme de Riemann basée sur la grille régulière $\mathcal{G} = \{y_1, \ldots, y_A\}$ de $[-s, s]$ pour un certain $A \geq 1$. Soit $(\zeta_1, \ldots, \zeta_M)$ M copies i.i.d. de ζ. Nous définissons

$$\hat{G}(t) = \frac{2s}{MA} \sum_{k=1}^{A} \sum_{i=1}^{M} \mathbb{1}_{\{\hat{p}(y_k|X_{n+i},\zeta_i)\geq t\}} \ .$$

Nous sommes en train de déterminer un intervalle de prédiction empirique dont nous pouvons analyser à la fois d'un point de vue numérique et théorique. Cet estimateur est défini comme suit

$$\hat{\Gamma}\left(X,\zeta\right) = \{y \in \mathbb{R} : \hat{p}\left(y|X,\zeta\right) \geq \hat{G}^{-1}(\ell)\} \ .$$

Le résultat suivant décrit que l'estimateur $\hat{\Gamma}$ a une longueur moyenne égale à la valeur demandée ℓ avec une vitesse de convergence d'ordre $s/\sqrt{M}$ sans aucune hypothèse.

Proposition 1. *Supposons que $A \geq 4\sqrt{M}$, alors*

$$\mathbb{E}\left[\left|\mathcal{L}(\hat{\Gamma}) - \ell\right|\right] \lesssim s/\sqrt{M} \ .$$

Le théorème suivant donne la consistance de $\hat{\Gamma}$ par rapport à l'excès de risque :

Théorème 6. *Supposons les hypothèses 10- 11. Considérons que $s = \log(\min(n, N))$ et $A = 4s\sqrt{M}$. Supposons que $\mathbb{E}[|f^*(X)|] \leq \infty$, $u = u_n \to 0$,*

$$\sqrt{s}\mathbb{E}\left[(\hat{f}(X) - f^*(X))^2\right] \underset{n\to+\infty}{\longrightarrow} 0, \quad \textit{et que} \quad s^{5/2}\mathbb{E}\left[|\hat{\sigma}^2(X) - \sigma^2(X)|\right] \underset{n\to+\infty}{\longrightarrow} 0,$$

alors

$$\mathbb{E}\left[\mathcal{E}_{\lambda_\ell}\left(\hat{\Gamma}_\ell\right)\right] \underset{n,N\to+\infty}{\longrightarrow} 0 \ .$$

Nous appliquons notre méthodologie à l'algorithme des K-PPV en ajoutant la condition α-marge basée sur la variable $p(Y|X)$.

Théorème 7. *Supposons que les hypothèses 1- 2 et 10-11 sont satisfaites. Soit $k_n \propto n^{2/(d+2)}$. Pour $s = \log\left(\min(n, M)\right)$, $A = 4s\sqrt{N}$ et $u_n = 1/n$, nous avons alors*

$$\mathbb{E}\left[\mathcal{E}_{\lambda_\ell}\left(\hat{\Gamma}_\ell\right)\right] \lesssim_{\log(n)} n^{-(1+\alpha)/(d+2)} + \min(n, N)^{-(1+\alpha)} + N^{-1/2} \ .$$

1.7 Conclusion et perspectives de recherche

Conclusion. Cette thèse a proposé des procédures d'estimation de la fonction de variance dans le modèle de régression hétéroscédastique - l'algorithme des K-PPV - et les procédures d'agrégation classiques "agrégation de type sélection modèle et convexe" (`MS` et `C` respectivement). Nous avons utilisé la méthode basée sur les erreurs résiduelles. Sous des hypothèses techniques, nous avons déterminé la vitesse de convergence de ces méthodes. La vitesse de convergence de l'estimateur de la fonction de variance par K-PPV par rapport à la ℓ_p-norme sup est d'ordre $\log(n)^p n^{-p/d+2}$. De plus, nous avons utilisé deux échantillons indépendants pour construire les estimateurs de la fonction de variance par `MS` et `C`. Nous avons montré que les deux approches d'agrégation ont la même vitesse de convergence qui est d'ordre $(\log(M_1)/N)^{-1/4}$ dans le cas où Y est borné et d'ordre $(\log(M_1)/N)^{-1/8}$ dans le cas gaussien. Ces vitesses lentes sont dues au fait que l'estimation de la fonction de variance repose sur la fonction de régression que l'on doit également estimer par les méthodes d'agrégation. Cette thèse a développé deux nouvelles applications de la fonction de variance : le problème de la régression avec option rejet et les intervalles de confiance en régression.

- **Problème de la régression avec option rejet.** Nous avons développé la prédiction avec option rejet dans le cadre de la régression. Nous nous sommes concentrés sur le cas où le taux de rejet est fixé. Nous avons fourni une règle optimale. Elle repose sur le seuillage de la fonction de variance. En utilisant deux échantillons indépendants (un de données étiquetées et un de données non étiquetées), nous avons construit un algorithme semi-supervisé basé sur le principe de type plug-in qui peut être appliqué à tout estimateur standard des fonctions de régression et de variance. Nous avons prouvé que notre prédicteur de type plug-in a un taux de rejet $\varepsilon \in (0, 1)$ à un terme d'ordre $M^{-1/2}$ près où M est la taille de l'échantillon des données non étiquetées. Nous avons obtenu un résultat de consistance de manière générale sur l'excès de risque qui dépend de la consistance des estimateurs des fonctions de régression et de variance par rapport à l'erreur L^2. Nous avons également appliqué notre méthodologie à l'algorithme des K-PPV, c'est-à-dire nous avons estimé f^* et σ^2 par K-PPV. Nous avons établi la vitesse de convergence pour le prédicteur de type plug-in basé sur K-PPV en utilisant la condition α-marge pour la variable $\sigma^2(X)$ autour du seuil λ_ε. Nous avons remarqué que si $\alpha > 1$ et si M est suffisamment grande, alors la vitesse de l'excès de risque est la même que la vitesse de convergence de $\hat{f}$ par rapport à l'erreur L^2.

- **Intervalle de prédiction en régression.** Nous avons construit un intervalle de prédiction avec une longueur moyenne fixé dans le modèle gaussien hétéroscédastique. Nous avons dérivé une règle optimale qui repose sur le seuillage de la densité conditionnelle, qui elle-même dépend des fonctions de régression et de variance. Cette règle permet d'interpréter la variable de sortie. Comme dans le problème d'option rejet, nous avons proposé une approche de type plug-in pour l'intervalle de prédiction optimal et pour l'estimation de la densité conditionnelle. Cette approche peut être adaptée à n'importe quel estimateur des fonctions de régression et de variance. Nous avons étudié la consistance de notre prédicteur résultant en termes d'excès de risque et de taux de longueur. Nous avons prouvé que notre estimateur a la même longueur que l'intervalle de prédiction optimal à un terme d'ordre $s/\sqrt{M}$ près, où le paramètre $s \geq 1$. Nous avons déterminé la vitesse de convergence en se basant sur l'algorithme des K-PPV.

Enfin, les procédures proposées dans cette thèse sont facilement implémentables. Les résultats obtenus dans ce manuscrit ouvrent des pistes pour des futurs travaux

Perspectives de recherche. Les travaux en cours portent sur plusieurs directions.

1. **Estimation de la fonction de variance.** Dans le deuxième chapitre, nous avons obtenu des vitesses de convergence de nos méthodes d'agrégation. Ce serait une motivation pour établir des bornes inférieures afin d'étudier l'optimalité des vitesses. De plus, une question intéressante serait d'étudier l'estimation de la fonction variance en grande dimension puisque travailler dans des espaces en grande dimension apporte souvent des difficultés théoriques qui viennent s'ajouter aux problèmes numériques de temps de calcul.

2. **Intervalle de prédiction.** Nous avons déterminé la borne supérieure de l'excès de risque de notre intervalle de prédiction empirique. Dans ce contexte, nous voudrions aborder la question des vitesses minimax en cas général. De plus, nous avons construit un intervalle de prédiction dans le cas où la variable $Y|X$ admet une densité. Il pourrait être intéressant de proposer une construction dans un cadre général où cette hypothèse n'est pas nécessairement vérifiée.

Chapter 2

Variance function estimation in regression model via aggregation procedures

Abstract: In the regression problem, we consider the problem of estimating the variance function by the means of aggregation methods. We focus on two particular aggregation setting: Model Selection aggregation (MS) and Convex aggregation (C) where the goal is to select the best candidate and to build the best convex combination of candidates respectively among a collection of candidates. In both cases, the construction of the estimator relies on a two-step procedure and requires two independent samples. The first step exploits the first sample to build the candidate estimators for the variance function by the residual-based method and then the second dataset is used to perform the aggregation step. We show the consistency of the proposed method with respect to the L^2-error both for MS and C aggregations. We evaluate the performance of these two methods in the heteroscedastic model and illustrate their interest in the quantile regression.

Contents

2.1 Introduction

Building efficient estimation of the level of noise is highly important for real applications and statistical analysis. In the heteroscedastic regression, which corresponds to the case where the variance of the errors depends on input variables, the heteroscedasticity must be detected and estimated. Indeed, not taking it into account in the estimation invalidates the conclusions of many statistical inference problems such as statistical tests which assume that the errors of the model all have the same variance. In addition, when using an approach that estimates the error variance as a function of input variables, the prediction intervals we obtain are likely to be more realistic than those obtained by assuming that the error variance is constant since the predictive uncertainty estimate depends on the estimate of the variance of the response variable. In general, testing and confidence intervals are two historical statistical problems where a bad calibration of the noise may lead to bad conclusions. Another important aspect of estimating the heteroskedasticity of the model is that the point estimate of the regression function is directly related to the variance function. The range of use of the variance structure in the data is even wider such as in selection the optimal kernel bandwidth [28], estimation correlation structure of the heteroscedastic spatial data [68], estimation of the quantile regression [44, 80], estimation of the signal-to-noise ratio [93], or choosing optimal design [63] and finding important applications, for instance, in finance with the problems of measuring volatility or risk [1] or long-term stock returns [59]. In our case, we highlight the interest in providing an efficient estimation of the variance function in the problem of regression with regret option where the good calibration of the rejection rule is highly dictated by the efficiency of the estimator of the noise level [24]. We focus on the regression problem: we denote by $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ the couple of random variables where X is the feature vector and Y is the response variable such that

$$Y = f^*(X) + W \ .$$

Here W is the noise and is such that $\mathbb{E}[W|X] = 0$ and $\mathbb{E}[W^2] < \infty$. In particular for any $x \in \mathbb{R}^d$ we write $f^*(x) = \mathbb{E}[Y|X = x]$ to denote the regression function and $\sigma^2(x) = \mathrm{Var}(Y|X = x) = \mathbb{E}\left[(Y - f^*(X))^2|X = x\right]$ to denote the conditional variance function.

Despite the popularity of the problem of estimating the noise level, there remains much to do. In particular, we study in the present chapter this problem from the aggregation perspective and build estimators of the conditional variance function based on Model Selection (MS) and Convex (C) aggregations. We study their consistency properties as well as their numerical performances.

2.1.1 Related work

Our literature review consists of three related fields:

Conditional variance estimation: The problem of estimating the regression function is classical and widely studied, see for example [9, 36, 78, 82, 86] and references therein.

Even though the problem of estimation of the conditional variance function is less studied, it has been considered in several works that can be cast into two groups according to the nature of the design (fixed or random).

When the design is fixed, the estimation of σ^2 has been studied mainly via residual-based methods [37, 38, 74] and difference-based methods [15, 63, 97]. Difference-based estimators do not require the estimation of the regression function f^*. The first difference-based estimators have been developed by [63]. They considered an initial variance estimates which are squared weighted sums of m observations neighbouring the fixed point where the variance function is to be estimated. The authors showed that the proposed initial variance estimates are not consistent. To solve this problem, they smoothed them with a kernel estimate. [15] presented a class of non-parametric variance estimators based on different sequences and local polynomial estimation and established asymptotic normality. [97] were interested in the effect of the unknown mean on the estimation of the variance function and proved numerically that the residual-based method performs better than the first-order-difference-based estimator when the unknown regression function f^* is very smooth.

In this work, we rather focus on random design. Less methods have been proposed to estimate the conditional variance function in this case. Most classical methods are the direct and the residual-based:

1. **The direct method**: a simple decomposition the conditional variance function σ^2 is rewritten as the difference of the first two conditional moments, $\sigma^2(x) = \mathbb{E}[Y^2|X = x] - (\mathbb{E}[Y|X = x])^2$. The direct method consists in estimating the two terms in the right side separately, see for example [30, 38]. To be more specific, the direct estimator of σ^2 has the following form

$$\hat{\sigma}_d^2(x) = \hat{g}(x) - (\hat{f}(x))^2 \ ,$$

where $\hat{g}$ and $\hat{f}$ are estimators of $\mathbb{E}[Y^2|X = x]$ and f^*, respectively. The main drawback of this approach is that it is not always nonnegative, for example, if different smoothing parameters are used in estimating those terms and adaptation to the unknown regression function f^* is still not available. [38] proposed a local polynomial regression estimates of those terms using the same bandwidth and the same kernel function. They established the asymptotic normality of local polynomial estimators of the regression function and the variance function.

2. **The residual-based method**: this approach consists of two steps. First, one estimates the regression function and computes the squared residuals $\hat{r} = (Y - \hat{f}(X))^2$ where $\hat{f}$ is the estimator of f^*. Second, we estimate the variance function by solving the regression problem when the input is the feature X and the output variable is the residuals $\hat{r}$. For more details, see [30, 67, 74]. It exists many ways to study this method. For instance, [30] applied a local linear regression in both steps and showed that their estimator is fully regression-adaptive to the unknown regression function. Using the local polynomial regression can be negative when the bandwidths are not selected appropriately. As a solution to this, [102] proposed estimators based on a localised normal likelihood, using a standard local linear form for estimating the mean and a local log-linear form for estimating the variance. [106] introduced an exponential estimator of the conditional variance in the second step to ensure the nonnegativity of the estimator of σ^2. [101] used a reweighted local constant estimator (kernel estimator) based on maximizing the empirical likelihood subject to a bias-reducing moment restriction. Moreover, such estimators have the form $\hat{\sigma}^2(X) = \sum_i \omega_i(X)(Y_i - \hat{f}(X_i))^2$ where $\omega_i(X)$ are weight functions ([24, 46]). Recently, [24] used the previous estimator and focused on estimating the regression function and the variance function respectively, by kNN. Under mild assumptions, they provided the rate of convergence of the kNN estimator of the conditional variance function in supremum norm. The residual-based method can be regarded as a generalized difference-based estimator. For more details, see [30].

In this chapter, we focus on the residual-based method to estimate the variance function since they appear more tractable. In particular, we develop an aggregation procedures for this task.

Aggregation methods: Aggregation is a popular approach in statistics and machine learning. This technique is well known to estimate the regression function in the homoscedastic or heteroscedastic model. We refer the reader to the baseline articles [4, 16, 43, 87, 88, 104]. Given a set of estimators of regression function f^*, the aggregation constructs a new estimator, called the aggregate, which mimics in a certain sense, the behavior of the best estimator in a class of estimates. There are several popular types of aggregation and we focus on two: the model selection aggregation (MS) which allows to select the best estimator from the set; the convex aggregation (C) where the goal is to select the best convex combination of functions in the set. In general, the aggregation procedures are based on sample splitting, that is, the original data set $\mathcal{D}_N$ is split into two independent data sequences $\mathcal{D}_k$ and $\mathcal{D}_l$ with $N = l + k \geq 1$. The first subsample $\mathcal{D}_k$ is exploited to build $M > 1$ competing estimators of the regression function f^* and $\mathcal{D}_l$ is used to aggregate them. Most of the work has focused on fixing the first sample, resulting in fixed estimators (the estimators are then seen as fixed functions). Under mild assumptions, the authors in [87] showed that the optimal rates for MS and C aggregation *w.r.t.* L^2-error in gaussian regression model are of order $\frac{\log(M)}{N}$, and $\frac{M}{N}$ if $M \leq \sqrt{N}$, respectively, $\sqrt{\frac{1}{N}\log\left(\frac{M}{\sqrt{N}}+1\right)}$ if $M \geq \sqrt{N}$ in both cases.

In this chapter, we consider aggregation methods for the conditional variance estimation. Up to our knowledge, such approaches have not been considered yet for this problem.

2.1.2 Main contribution

We develop the notions of model selection aggregation and convex aggregation to estimate the conditional variance function. To our knowledge, this work is the first to deal with this setting. We consider two independent datasets: the first will be used to build the initial estimators of the variance function and the second to aggregate them. We call these estimators the MS-estimator and C-estimator. We consider the residual-based method to build the initial estimators which is based on estimating the regression function in the first step. We focus on estimating the regression function by model selection aggregation and convex aggregation. In this chapter, the major part is then devoted to show the upper-bounds of L^2-error of the MS-estimator and C-estimator when the initial estimators can be arbitrary or verify very weak conditions such that boundedness. We establish that the rate of convergence for MS and C procedures is of order $O((\log(M_1)/N)^{1/8})$ when Y is unbounded and is of order $O((\log(M_1)/N)^{1/4})$ when Y is bounded. Finally, we obtain numerical results which show the performance of our procedures.

2.1.3 Outline

The chapter is organized as follows. In the next section, the aggregation problems, the model selection and convex aggregations, are described in detail. Section 2.3 is focused on investigating the upper-bounds for the L^2-error of our procedures. Finally, Section 2.4 presents a numerical comparison of the proposed method w.r.t. the heteroscedastic model as well as a direct application to the quantile regression.

Notations. We introduce some notation that is used throughout this chapter. Let $p \geq 2$ be an integer, the set of integers $\{1, \dots, p\}$ is denoted $[p]$. Let n and N be integers. For any function $f : \mathbb{R}^d \to \mathbb{R}$, we define the empirical norm $\|f\|_N^2 = \frac{1}{N}\sum_{i=n+1}^{n+N} |f(X_i)|^2$ and the supremum norm $\|f\|_\infty = \sup_{x\in\mathbb{R}^d} |f(x)|$. Moreover, we denote by $\Lambda^p := \{\lambda \in \mathbb{R}^p : \lambda_j \geq 0, \sum_{j=1}^p \lambda_j = 1\}$ the simplex. Let $\|\cdot\|_{1,p}$ denote the ℓ_1 norm on $\mathbb{R}^p$, that is, $\|\lambda\|_{1,p} := \sum_{j=1}^p |\lambda_j|$. For the sake of simplicity, let $Z = (Y - f^*(X))^2$.

2.2 Aggregation estimators

In this section, we describe an estimation algorithm of the variance function σ^2 by aggregation. In particular, we focus on two types of aggregations: the model selection aggregation (MS), and the convex aggregation (C). These aggregation problems, (MS) and (C), have been considered to estimate the unknow regression function in the regression model. The objective is to estimate f^* by a combination of elements of a known set called dictionary made up of deterministic functions or preliminary estimators. The collection of estimators or algorithms is given and can be parametric, nonparametric or semi-parametric nature. Given a set of estimators, the MS-aggregation consists in constructing a new estimator which is approximately as good as the best estimator in the set, while the objective of C-aggregation is to construct a new estimator which is approximately as good as the best convex combination of the elements in the set, for more details see [4, 16, 43, 87, 88, 104]. Besides, to construct an aggregate of σ^2, we first introduce two independent learning samples: $\mathcal{D}_n = \{(X_i, Y_i), i = 1, \dots, n\}$ and $\mathcal{D}_N = \{(X_i, Y_i), i = n+1, \dots, n+N\}$ which consist of respectively, n and N i.i.d. copies of (X, Y).

2.2.1 Model selection aggregation

In this first paragraph, we detail how we perform MS-aggregation in order to estimate of the conditional variance function σ^2 by MS. It consists of two steps: one step for the estimation of the regression function f^* and a second one devoted to the estimation of σ^2. More precisely, in the first one we build

M_1 estimators of the regression function $\hat{f}_1, \ldots, \hat{f}_{M_1}$ based on $\mathcal{D}_n$ with $2 \leq M_1 < \infty$. Then, we use the second dataset $\mathcal{D}_N$ to estimate f^* by MS: we select the optimal index, denoted $\hat{s}$ as follows

$$\hat{s} \in \underset{s \in [M_1]}{\operatorname{argmin}} \hat{\mathcal{R}}_N(\hat{f}_s), \quad \text{where} \quad \hat{\mathcal{R}}_N(\hat{f}_s) = \frac{1}{N} \sum_{i=n+1}^{n+N} |Y_i - \hat{f}_s(X_i)|^2 \ , \tag{2.1}$$

and the aggregate of the regression function, denoted by $\hat{f}_{\text{MS}}$, is then given by

$$\hat{f}_{\text{MS}} := \hat{f}_{\hat{s}}. \tag{2.2}$$

In the second step, given the estimator $\hat{f}_{\text{MS}}$ built on $\mathcal{D}_N$, we construct using back the sample $\mathcal{D}_n$ M_2 estimators of the variance function σ^2, denoted $\hat{\sigma}^2_{\hat{s},1}, \ldots, \hat{\sigma}^2_{\hat{s},M_2}$, by residual-based method with $2 \leq M_2 < \infty$. Finally, based on $\mathcal{D}_N$ again, we select the optimal single, denoted $\hat{m}$, as follows

$$\hat{m} \in \underset{m \in [M_2]}{\operatorname{argmin}} \hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) \ \text{ where } \ \hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) = \frac{1}{N} \sum_{i=n+1}^{n+N} |\hat{Z}_i - \hat{\sigma}^2_{\hat{s},m}(X_i)|^2$$

with $\hat{Z}_i = \left(Y_i - \hat{f}_{\text{MS}}(X_i)\right)^2$. Therefore, the aggregate of the variance function, called MS-estimator and denoted $\hat{\sigma}^2_{\text{MS}}$, is defined as

$$\hat{\sigma}^2_{\text{MS}} := \hat{\sigma}^2_{\hat{s},\hat{m}}. \tag{2.3}$$

2.2.2 Convex aggregation

Convex aggregation procedures for nonparametric regression are discussed in [3, 16, 87]. We describe here an algorithm for aggregating estimates of the conditional variance function σ^2 by C-aggregation. As for MS-aggregation, the construction of the aggregate of σ^2 needs two independent datasets $\mathcal{D}_n$ and $\mathcal{D}_N$. The aggregation still proceeds in two steps: one for estimating f^* and the second for the estimation of σ^2. Each step is again decomposed in two parts. Firstly, we consider M_1 estimators of the regression function f^*, $\{\hat{f}_1, \ldots, \hat{f}_{M_1}\}$, based on $\mathcal{D}_n$, and for any $\lambda \in \Lambda^{M_1}$ we define the linear combinations $\hat{f}_\lambda$

$$\hat{f}_\lambda = \sum_{j=1}^{M_1} \lambda_j \hat{f}_j.$$

Then, aggregates of the regression function based on the sample $\mathcal{D}_N$ have the form

$$\hat{f}_{\text{C}} := \hat{f}_{\hat{\lambda}} = \sum_{j=1}^{M_1} \hat{\lambda}_j \hat{f}_j \ , \tag{2.4}$$

where the estimator $\hat{\lambda}$ is defined by

$$\hat{\lambda} \in \underset{\lambda \in \Lambda^{M_1}}{\operatorname{argmin}} \hat{\mathcal{R}}_N(\hat{f}_\lambda).$$

Once $\hat{f}_{\text{C}}$ is obtained, we focus on the estimation of σ^2. Based on the sample $\mathcal{D}_n$, we build M_2 estimators for the conditional variance function by the residual-based method, denoted $\hat{\sigma}^2_{\hat{\lambda},1}, \ldots, \hat{\sigma}^2_{\hat{\lambda},M_2}$, and for any $\beta \in \Lambda^{M_2}$ we define $\hat{\sigma}^2_{\hat{\lambda},\beta}$ as follows

$$\hat{\sigma}^2_{\hat{\lambda},\beta} = \sum_{j=1}^{M_2} \beta_j \hat{\sigma}^2_{\hat{\lambda},j} \ .$$

Finally, based on $\mathcal{D}_N$, the aggregate estimate for σ^2 is given by

$$\hat{\sigma}^2_{\text{C}} := \hat{\sigma}^2_{\hat{\lambda},\hat{\beta}} \ , \tag{2.5}$$

where the estimator $\hat{\beta}$ is defined by

$$\hat{\beta} \in \underset{\beta \in \Lambda^{M_2}}{\operatorname{argmin}} \hat{R}_N(\hat{\sigma}^2_{\hat{\lambda},\beta}), \quad \text{where} \quad \hat{R}_N(\hat{\sigma}^2_{\hat{\lambda},\beta}) = \frac{1}{N} \sum_{i=n+1}^{n+N} |\hat{Z}_i - \hat{\sigma}^2_{\hat{\lambda},\beta}(X_i)|^2$$

with $\hat{Z}_i = (Y_i - \hat{f}_{\text{C}}(X_i))^2$. We called $\hat{\sigma}^2_{\text{C}}$ the C-estimator.

2.3 Main results

This section is devoted to studying the L^2-error of MS-estimator and C-estimator. Firstly, we introduce general conditions required on the model in Section 2.3.1. Secondly, we show the consistency of our methods in Sections 2.3.2 and 2.3.3.

2.3.1 Assumptions

The following assumptions are the bedrock of our theoretical analysis:

Assumption 1. *The functions f^* and σ^2 are bounded.*

Assumption 2. *Y is bounded or Y satisfies the gaussian model*

$$Y = f^*(X) + \sigma(X)\xi, \tag{2.6}$$

where ξ is independent of X and distributed according to a standard normal distribution.

These assumptions are relatively weak and play a key role in our approach. They allow to use the Hoeffding's inequality in the case of boundness of Y or ξ. In particular, it is important to emphasize that Assumptions 1 and 2 guarantee that the variable $Y - f^*(X)$ is sub-Gaussian (see Lemma 9 in the case where Y is bounded).

2.3.2 Upper bound for $\hat{\sigma}^2_{\text{MS}}$

We study the L^2-error of the MS-estimator $\hat{\sigma}^2_{\text{MS}}$. Let $\mathcal{R}(\hat{f}_s) = \mathbb{E}\left[|Y - \hat{f}_s(X)|^2\right]$ for all $s \in [M_1]$. We define s^* as follows

$$s^* \in \operatorname*{argmin}_{s\in[M_1]} \mathcal{R}(\hat{f}_s) \ . \tag{2.7}$$

Besides, we need the following assumptions in the case of MS-aggregation:

Assumption 3. *For all $s \in [M_1]$ and all $m \in [M_2]$, $\hat{f}_s$ and $\hat{\sigma}^2_{s,m}$ are bounded a.s $\mathcal{D}_n$. More precisely, there exist two positive constants K_1 and K_2 such that for all $n \in \mathbb{N}^*$*

$$\max_{s\in[M_1]} \|\hat{f}_s\|_\infty \leq K_1, \ \text{ and } \ \max_{(s,m)\in[M_1]\times[M_2]} \|\hat{\sigma}^2_{s,m}\|_\infty \leq K_2.$$

Assumption 4 (Separability hypothesis)**.** *There exists $\delta_0 > 0$ such that*

$$\delta^*\left(\mathcal{D}_n\right) = \min_{s\neq s^*}\{|\mathcal{R}(\hat{f}_s) - \mathcal{R}(\hat{f}_{s^*})|\} > \delta_0 \ .$$

Both assumptions are used to control the L^2-error of the MS-estimator $\hat{\sigma}^2_{\text{MS}}$. Assumption 3 describes the boundedness of the estimators. It is in particular satisfied if the functions in the dictionaries are bounded. In our construction, the constants K_1 and K_2 do not need to be known. In practice, the response variable Y in the sample is finite and then it ensures that the candidates in the dictionaries are bounded. Assumption 4 is used for MS and helps us to ensure that the minimum of the risk $\mathcal{R}$ is well defined. It cannot be verified in practice since it depends on the distribution $\mathbb{P}$. Let $\mathbb{E}$ be the expectation which is taken with respect to both X and the samples $\mathcal{D}_n$ and $\mathcal{D}_N$. We establish the following result

Theorem 9. *Let $\hat{f}_{MS}$ and $\hat{\sigma}^2_{MS}$ be two MS-estimators of f^* and σ^2 defined in Eq. (2.2) and (2.3), respectively. Then, under Assumptions 1- 4, there exist two absolute constants $C_1 > 0$ and $C_2 > 0$ such that*

$$\mathbb{E}\left[|\hat{\sigma}^2_{MS}(X) - \sigma^2(X)|^2\right] \leq \mathbb{E}\left[\min_{m\in[M_2]} \mathbb{E}_X\left[|\hat{\sigma}^2_{s^*,m}(X) - \sigma^2(X)|^2\right]\right] + C_1\left\{\min_{s\in[M_1]} \mathbb{E}\left[\|\hat{f}_s - f^*\|^2_N\right]\right\}^{1/2p} + C_2\left(\frac{\log(M_1)}{N}\right)^{1/4p},$$

with $p = 1$ if Y is bounded and $p = 2$ otherwise.

The proof of this result is postponed to the Appendix. Let's give a sketch of the proof. The L^2-error is the excess risk of $\hat{\sigma}^2_{\texttt{MS}}$ where $\mathbb{E}\left[|\hat{\sigma}^2_{\texttt{MS}}(X) - \sigma^2(X)|^2\right] := \mathbb{E}\left[R(\hat{\sigma}^2_{\texttt{MS}}) - R(\sigma^2)\right]$ with $R(\sigma^2) = \mathbb{E}\left[|Z - \sigma^2(X)|^2\right]$. We introduce the minimizer $\bar{\sigma}^2_{\texttt{MS}} := \hat{\sigma}^2_{\hat{s},\bar{m}}$ where $\bar{m} \in \operatorname{argmin}_{m\in[M_2]} R(\hat{\sigma}^2_{\hat{s},m})$. We consider the decomposition $\mathbb{E}\left[|\hat{\sigma}^2_{\texttt{MS}}(X) - \sigma^2(X)|^2\right] = \mathbb{E}\left[R(\hat{\sigma}^2_{\texttt{MS}}) - R(\bar{\sigma}^2_{\texttt{MS}})\right] + \mathbb{E}\left[R(\bar{\sigma}^2_{\texttt{MS}}) - R(\sigma^2)\right]$. We control the two terms in the right side separately. The first one is the estimation error (variance term). To control it, we need to introduce $\tilde{\sigma}^2_{\texttt{MS}} := \hat{\sigma}^2_{\hat{s},\tilde{m}}$ where $\tilde{m} \in \operatorname{argmin}_{m\in[M_2]} R_N(\hat{\sigma}^2_{\hat{s},m})$, with $R_N(\hat{\sigma}^2_{\hat{s},m}) = \frac{1}{N}\sum_{i=n+1}^{n+N} |Z_i - \hat{\sigma}^2_{\hat{s},m}(X_i)|^2$. The upper bound of the variance depends on the L^2-error of the aggregate $\hat{f}_{\texttt{MS}}$ with respect to the empirical norm. The second one is the approximation error. Its upper-bound is linked to $\mathbb{P}(\hat{s} \neq s^*)$.

Theorem 9 gives the upper-bound for L^2-error of $\hat{\sigma}^2_{\texttt{MS}}$. This bound consists of two parts: the first part is the bias of MS-estimator $\hat{\sigma}^2_{\texttt{MS}}$ and depends on the deterministic selector s^*; the second part is composed of the two remaining terms and corresponds to the estimation error (variance). The first term is the bias term of $\hat{f}_{\texttt{MS}}$ expressed in terms of the empirical norm $\|\cdot\|^2_N$, and the second one characterises the price to pay for MS-aggregation and is of order $(\log(M_1)/N)^{1/4p}$ where $p = 1$ if Y is bounded and $p = 2$ otherwise. Note that this rate is slower than in the case of the estimation of the regression function f^*. This slow rate is due to the double aggregation that we need to perform for the estimation of the conditional variance function.

2.3.3 Upper bound for $\hat{\sigma}^2_{\texttt{C}}$

In this part, we focus in studying the L^2-error of $\hat{\sigma}^2_{\texttt{C}}$. The construction of $\hat{\sigma}^2_{\texttt{C}}$ needs the following estimators $\{\hat{f}_i\}_{i=1}^{M_1}$ and $\{\hat{\sigma}_{\hat{\lambda},i}\}_{i=1}^{M_2}$. We require the following conditions

Assumption 5. *For all $i \in [M_1]$, all $\lambda \in \Lambda^{M_1}$ and all $j \in [M_2]$, $\hat{f}_i$ and $\hat{\sigma}^2_{\lambda,j}$ are bounded a.s. $\mathcal{D}_n$.*

Assumption 6. *Suppose that there exists a constant $K \geq 0$ such that for every $j \in [M_2]$*

$$\mathbb{E}\left[|\hat{\sigma}^2_{\lambda_1,j}(X) - \hat{\sigma}^2_{\lambda_2,j}(X)|\right] \leq K\|\lambda_1 - \lambda_2\|_{1,M_1}, \quad \forall \lambda_1, \lambda_2 \in \Lambda^{M_1} \ \ a.s\ \mathcal{D}_n.$$

Assumption 5 describes the boundedness of the candidates as Assumption 3. Assumption 6 is a strong condition. However, it holds, for instance, for estimators of the form $\hat{\sigma}^2_{\lambda,j}(X) = \sum_i \omega_i(X)(Y_i - \hat{f}_\lambda(X_i))^2$ where $\omega_i(X)$ are weight functions, that are nonnegative and sum to one. The next theorem is the main result of this section, it displays the upper-bound of L^2-error for $\hat{\sigma}^2_{\texttt{C}}$.

Theorem 10. *Let $\hat{f}_C$ and $\hat{\sigma}^2_C$ be two C-estimators of f^* and σ^2 defined in Eq. (2.4) and (2.5), respectively. Then, under Assumptions 1, 2, 5, and 6, there exist two absolute constants $C_1 > 0$ and $C_2 > 0$ such that*

$$\mathbb{E}\left[|\hat{\sigma}^2_C(X) - \sigma^2(X)|^2\right] \leq \mathbb{E}\left[\inf_{\beta\in\Lambda^{M_2}} \mathbb{E}_X\left[|\hat{\sigma}^2_{\hat{\lambda},\beta}(X) - \sigma^2(X)|^2\right]\right] + C_1\left\{\inf_{\lambda\in\Lambda^{M_1}} \mathbb{E}\left[\|\hat{f}_\lambda - f^*\|^2_N\right]\right\}^{1/2p} + C_2\left(\frac{\log(M_1)}{N}\right)^{1/4p},$$

with $p = 1$ if Y is bounded and $p = 2$ otherwise.

As for Theorem 9, the upper-bound for the L^2-error of C-estimator $\hat{\sigma}^2_{\texttt{C}}$ is composed of three terms. The first one is the bias term of $\hat{\sigma}^2_{\texttt{C}}$ which depends on the random selector $\hat{\lambda}$, the second and third ones is a bound of the variance term that rely on the bias term of $\hat{f}_{\texttt{C}}$ with respect to the empirical norm $\|\cdot\|^2_N$ and on the price to pay for convex aggregation which is of order $(\log(M_1)/N)^{1/4p}$ where $p = 1$ if Y is bounded and $p = 2$ otherwise.

We notice that both procedures MS and C have the same rate. Indeed, the variance term of $\hat{\sigma}^2_{\texttt{MS}}$ and $\hat{\sigma}^2_{\texttt{C}}$ is based on the upper bound for $\hat{f}_{\texttt{MS}}$ and $\hat{f}_{\texttt{C}}$. Moreover, the aggregates $\hat{f}_{\texttt{MS}}$ and $\hat{f}_{\texttt{C}}$ have the same rate which is of order $(\log(M_1)/N)^{1/2}$ with respect to the empirical norm $\|\cdot\|^2_N$, see Proposition 2 and Proposition 3 in the Appendix. Let us now compare with the rates of $\hat{f}_{\texttt{MS}}$ and $\hat{f}_{\texttt{C}}$ with respect to

L^2-error and L^2-risk. For the Gaussian and bounded regression model, the rate of the variance term of $\hat{f}_{\text{MS}}$ and $\hat{f}_{\text{C}}$ is of order $\frac{\log(M_1)}{N}$ and $\frac{M_1}{N}$ if $M_1 \leq \sqrt{N}$, respectively, $\sqrt{\frac{1}{N}\log\left(\frac{M_1}{\sqrt{N}}+1\right)}$ if $M_1 \geq \sqrt{N}$ in both cases [16, 47, 48, 87]. We can deduce that our rates are very slow because our procedures need to estimate at the same time the unknown regression function f^* and the variance function σ^2 by aggregation procedures.

2.4 Numerical results

This section is devoted to the numerical analysis of our procedures. In Section 2.4.1, we describe four heteroscedastic models in the gaussian case and two models when Y is bounded. Second, we evaluate the performances of MS-estimator and C-estimator for different examples in Section 2.4.2. Once we have calibrated our estimate of the variance function σ^2, we exploit it to consider the quantile regression in Section 2.4.4.

2.4.1 Data

Our numerical study relies on synthetic data:
Heteroscedastic models: we propose four examples of heteroscedastic models (2.6):

- Model 1: let $a \in \{1/4, 1\}$ and $X = (X_1, X_2, X_3)$ have a uniform distribution on $[0,1]^3$. Let
  1. $f^*(X) = 0.1\cos(X_1) + \exp(-X_3^2)$;
  2. $\sigma^2(X) = a\left(0.1 + \exp(-7(X_1-0.2)^2) + \exp(-10(X_2-0.5)^2 + \exp(-50(X_3-0.9)^2\right)$.
- Model 2: let $X = (X_1, \ldots, X_{10})$ have a uniform distribution on $[0,1]^{10}$. We define
  1. $f^*(X) = 0.1 + \exp(-X_1^2) + 0.2\sin(X_2 + X_3 + X_4 + 0.1X_5^2)$;
  2. $\sigma^2(X) = \frac{1}{2}(0.5+\sqrt{X_1(1-X_2)}+0.8X_3X_4+X_5X_6X_7^2+0.9\exp(-500(X_8+X_9+X_{10}-0.5)^2))^2$.
- Model 3: sparse model
  1. $f^*(X) = X\beta, \quad \beta \in \mathbb{R}^p$;
  2. $\sigma^2(X) = \frac{1}{2}\left(0.3 + \sqrt{X_1(1-X_1)}\sin\left(\frac{2.1\pi}{X_2+0.05}\right) + 0.5X_3 + X_4\right)^2$.

Bounded Y: we consider the following regression model when Y is bounded

$$Y = f^*(X) + \sigma(X)\varrho$$

where ϱ have a uniform distribution on $[-\sqrt{3}, \sqrt{3}]$. We give the following examples of models :

- Model 4: let X have a uniform distribution on $[0,1]^2$ and
  1. $f^*(X) = X_1 + \exp(-X_2^2)$;
  2. $\sigma^2(X) = 0.01 + X_1\exp\left(-(X_2-0.9)^2\right)$.
- Model 5: let $X = (X_1, X_2, X_3)$ have a uniform distribution on $[0,1]^3$ and
  1. $f^*(X) = X_1 + X_2 + 0.5\cos(X_3)$;
  2. $\sigma^2(X) = \left(0.3 + \sqrt{X_1(1-X_1)}\sin\left(\frac{(2.1)\pi}{X_2+0.05}\right) + X_3\right)^2$.

We describe the previous models. We display in Figures 2.1 and 2.2 the histograms of the variance function for every model. Model 1 is a multivariate model in which the regression and variance functions are regular functions. In the case $a = 1/4$, the problem of estimation of σ^2 is hard since it takes a large proportion of values smaller than 1, while the case $a = 1$ is simpler because about 76.3% of the values of σ^2 are larger than 1 and 0.04% larger than 3. Moreover, Model 2 is also a multivariate

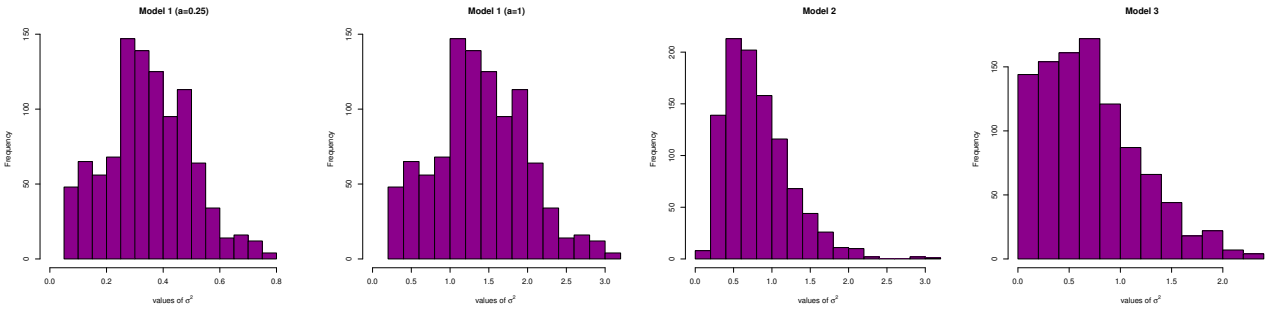


Figure 2.1 – Histogram of values of σ^2 in Gaussian models.

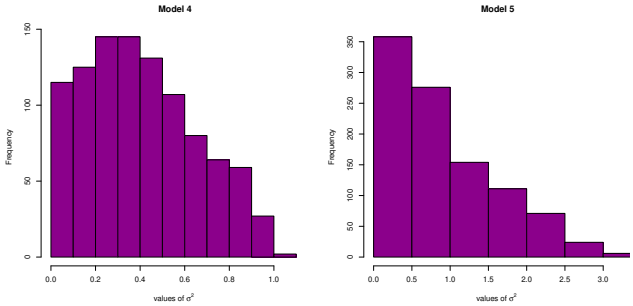


Figure 2.2 – Histogram of values of σ^2 in regression models for bounded responses.

model where we introduce higher order terms in the variance function. In this sense, the estimation of the variance function is hard since in addition, there are only 28% of values of σ^2 greater than 1. In Model 3 we consider a sparse model for the regression function where X is an $N \times p$ matrix (p is the number of predictors) with independent uniform entries, $\beta \in \mathbb{R}^p$ is a vector of weights, and $\xi \in \mathbb{R}^N$ is a standard Gaussian noise vector and is independent of the feature X. We fix $p = 50$. The vector β is chosen to be s-sparse where $s < p$, that means β has only first s coordinates different from 0; $\beta_i = \mathbb{1}_{\{i \leq s\}}$. Here, we choose $s = 14$. In addition, the variance function in this model is less difficult. Indeed, σ^2 takes only 24.8% values greater than 1. Finally, the last two examples are two models when Y is bounded. Model 4 is bivariate regression model where the estimation of σ^2 is difficult (about 99.8% of the values are less than 1). Lastly, considering Model 5, the values of σ^2 are between 0 and 3.12. There are 36.6% of values that are larger than 1. From this perspective the estimation of the variance function is less complicated. However, the presence of higher order terms makes the problem harder.

2.4.2 Benefit of aggregation

In this section, we improve the classical methods based on residual-based approach by considering aggregation. In the same time we compare `MS` and `C` aggregation.

Machines and simulation scheme

The construction of the aggregates $\hat{\sigma}^2_{\text{MS}}$ and $\hat{\sigma}^2_{\text{C}}$ is described in Sections 2.2.1 and 2.2.2. We recall that we focus on the residual-based method to compute the candidates of the variance function σ^2. One of the advantages of using the aggregation approach is that the collection of candidates is chosen by the practitioner and can be arbitrary. We build three dictionaries $\mathcal{F} = \{\hat{f}_s\}_{s=1}^{12}$, $\mathcal{G}_1 = \left\{\hat{\sigma}^2_{\hat{s},12}\right\}_{m=1}^{12}$ and $\mathcal{G}_2 = \{\hat{\sigma}^2_{\hat{\lambda},j}\}_{j=1}^{12}$ that contain 12 machines each: the random forest with different number of trees (ntree=$50, 150, 500$), the kNN with different values of k ($k = 7, 13, 22$), the Lasso with different values of tuning parameter ($\lambda = 0.5, 2$), the Ridge with different values of tuning parameter ($\lambda = 0.9, 3$),

regression tree and the Elastic Net regression with a penalty term $\lambda = 1$ and a parameter $\alpha = 0.6$ that compromises between the ℓ_1 and the ℓ_2 terms in the penalty. The first dictionary is exploited to compute the aggregates $\hat{f}_{\text{MS}}$ and $\hat{f}_{\text{C}}$ while the last two are used to calculate respectively, $\hat{\sigma}^2_{\text{MS}}$ and $\hat{\sigma}^2_{\text{C}}$ with those 12 machines. For the 12 algorithms, we use the following R packages:

- Regression tree (R package `tree`, [72]);
- k-nearest neighbours regression (R package `FNN`, [53]);
- RandomForest regression (R package `randomForest`, [54]);
- Lasso regression (R package `glmnet`, [32]);
- Ridge regression (R package `glmnet`);
- Elastic Net regression (R package `glmnet`).

Other parameters are set by default. In addition to that, we use `Optim` function in `R` which is based on method `BFGS` to compute $\hat{\lambda}$ and $\hat{\beta}$. Now, we evaluate the performances of $\hat{\sigma}^2_{\text{MS}}$ and $\hat{\sigma}^2_{\text{C}}$ on previous models. Besides, we provide estimation of the L^2-error for $\hat{\sigma}^2_{\text{MS}}$ and $\hat{\sigma}^2_{\text{C}}$ and repeat independently $L = 100$ times the following steps:

1. simulate three datasets $\mathcal{D}_n$, $\mathcal{D}_N$ and $\mathcal{D}_T$ with $n \in \{100, 1000\}$, $N \in \{100, 1000\}$ and $T = 1000$;
2. based on $\mathcal{D}_n$, we compute the dictionary $\mathcal{F}$, and then based on $\mathcal{D}_N$, we compute the aggregates $\hat{f}_{\text{MS}}$ (that is $\hat{s}$) and $\hat{f}_{\text{C}}$ (that is $\hat{\lambda}$) of the regression function f^* provided in Eqs (2.2) and (2.4);
3. based on $\mathcal{D}_n$ and $\hat{f}_{\text{MS}}$ (resp. $\hat{f}_{\text{C}}$), we compute the collection $\mathcal{G}_1$ (resp. $\mathcal{G}_2$) and we calculate $\hat{\sigma}^2_{\text{MS}}$ and $\hat{\sigma}^2_{\text{C}}$ on $\mathcal{D}_N$;
4. based on $D_n \cup D_N$: firstly, we compute the collection $\mathcal{F}$[1]; secondly, for each estimate $\hat{f}_s$ in $\mathcal{F}$ we calculate the estimators $\{\hat{\sigma}^2_{s,m}\}_{1\le m\le 12}$ of σ^2 corresponding to the 12 procedures in $\mathcal{F}$;
5. finally, over $\mathcal{D}_T$, we compute the empirical L^2-error of the aggregates $\hat{\sigma}^2_{\text{MS}}$ and $\hat{\sigma}^2_{\text{C}}$. On the other hand, we compute the collection of estimators of the variance function $\{\sigma^2_{s,m}\}_{1\le s,m\le 12}$ and we choose the best among them in terms of empirical L^2-error (always on the datatest $\mathcal{D}_T$). That means, we take the smallest of empirical L^2-error as follow: for all $(s, m) \in [12] \times [12]$

$$\min \frac{1}{T} \sum_{i=1}^{T} \left(\hat{\sigma}^2_{s,m}(X_i) - \sigma^2(X_i)\right)^2 \quad , \tag{2.8}$$

and denote the best method. Finally, we compare it with our aggregation methods.

From these experiments, we compute the means and standard deviations of both empirical L^2-errors $\widehat{\text{Err}}$ for $\hat{\sigma}^2_{\text{MS}}$, $\hat{\sigma}^2_{\text{C}}$ and the best method and we display the boxplot of the empirical L^2-error.

Results

We present our results in Figures 2.3-2.8 and Tables 2.1 and 2.2. We make several observations. First, the convex aggregation method is better than the model selection aggregation method in all models. The best can only serve as a benchmark to see the performance of a real estimator. For this reason, it is perfectly natural that "best" has better performances than our aggregation procedures on the test sample $\mathcal{D}_T$ because the selection of the best couple (s, m) (see Eq. (2.8)) depends precisely on $\mathcal{D}_T$. Second, we notice that when n and N are enough, the `MS`-estimator $\hat{\sigma}^2_{\text{MS}}$ and the `C`-estimator $\hat{\sigma}^2_{\text{C}}$ have a similar performance, that is close to the performance of the best method. These results reflect our theory: the consistency of `MS`-estimator and the `C`-estimator. Third, we observe that the empirical

[1]Note that this set of estimators differ from the dictionary computed in step 2. since it is computed in the whole data $D_n \cup D_N$. We abuse in the notation to avoid extra notation that are irrelevant for the understanding.

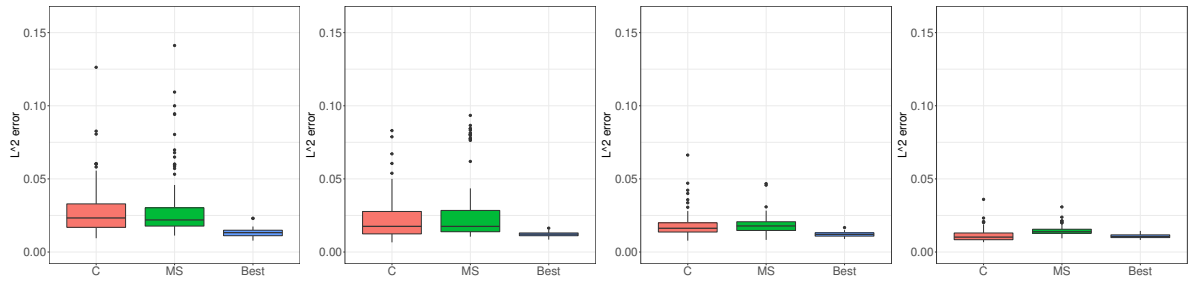
L^2-error of $\hat{\sigma}^2_{\text{MS}}$ and $\hat{\sigma}^2_{\text{C}}$ decreases faster in the simpler models (with respect to the estimation of the variance function) when n and N increase (see the evolution of the boxplots in Figures 2.4 and 2.8 as compared to Figures 2.3 and 2.6. In addition, our numerical results highlight an interesting fact: when we split data, it is advantageous to put more data in the second dataset $\mathcal{D}_N$ used in the aggregation step. Indeed, it seems as illustrated in Table 2.2 that the methods have better performance for large samples $\mathcal{D}_N$ is all cases. As an example, the mean error in Model 1 with $a = 1$ for C-aggregation is 0.33 when $n = 1000$ and $N = 100$ and 0.26 when $n = 100$ and $N = 1000$.

	$n = N = 100$			$n = N = 1000$		
	C	MS	Best	C	MS	Best
Model	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$
Model 1 ($a = 0.25$)	0.028 (0.018)	0.031 (0.023)	0.013 (0.003)	0.011 (0.004)	0.014 (0.003)	0.011 (0.001)
Model 1 ($a = 1$)	0.407 (0.214)	0.428 (0.279)	0.200 (0.44)	0.155 (0.044)	0.200 (0.040)	0.164 (0.013)
Model 2	0.247 (0.133)	0.272 (0.180)	0.110 (0.025)	0.106 (0.046)	0.100 (0.093)	0.070 (0.010)
Model 3	0.287 (0.092)	0.302(0.125)	0.218 (0.019)	0.194 (0.021)	0.198 (0.044)	0.164 (0.011)
Model 4	0.032 (0.027)	0.034 (0.036)	0.010 (0.005)	0.010 (0.005)	0.011 (0.003)	0.009 (0.001)
Model 5	0.382 (0.116)	0.405 (0.168)	0.264 (0.032)	0.209 (0.040)	0.223 (0.024)	0.178 (0.016)

Table 2.1 – Average and standard deviation of the empirical L^2-error of the three estimators with $n = N$.

	$n = 1000$ $N = 100$			$n = 100$, $N = 1000$		
	C	MS	Best	C	MS	Best
Model	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$
Model 1 ($a = 0.25$)	0.023 (0.015)	0.028 (0.023)	0.012 (0.002)	0.018 (0.008)	0.019 (0.006)	0.012 (0.002)
Model 1 ($a = 1$)	0.335 (0.265)	0.381 (0.343)	0.170 (0.014)	0.262 (0.090)	0.278 (0.081)	0.169 (0.018)
Model 2	0.193 (0.132)	0.227 (0.189)	0.074 (0.013)	0.159 (0.055)	0.148 (0.054)	0.073 (0.010)
Model 3	0.252 (0.082)	0.278 (0.149)	0.180 (0.015)	0.259 (0.029)	0.270 (0.035)	0.179 (0.015)
Model 4	0.021 (0.014)	0.026 (0.027)	0.009 (0.002)	0.019 (0.012)	0.017 (0.015)	0.009 (0.002)
Model 5	0.295 (0.144)	0.336 (0.209)	0.195 (0.019)	0.313 (0.079)	0.317 (0.095)	0.194 (0.015)

Table 2.2 – Average and standard deviation of the empirical L^2-error of the three estimators with $n \neq N$.

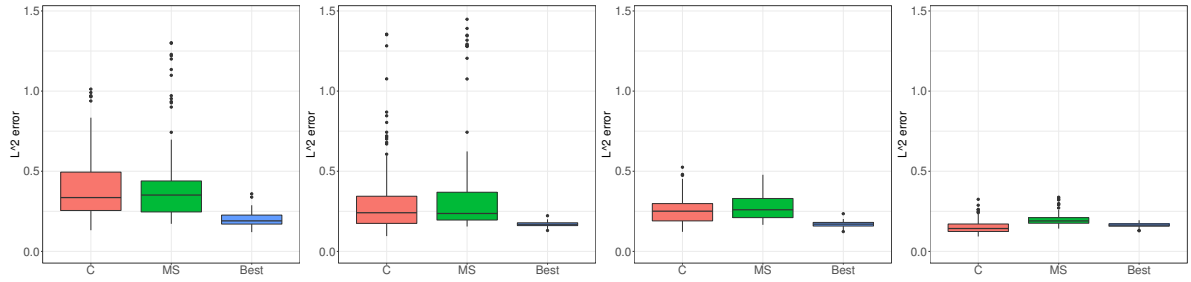


(a) $n = N = 100$ (b) $n = 1000, N = 100$ (c) $n = 100, N = 1000$ (d) $n = N = 1000$

Figure 2.3 – Boxplot of the empirical L^2-error of the estimators in Model 1 ($a = 0.25$)

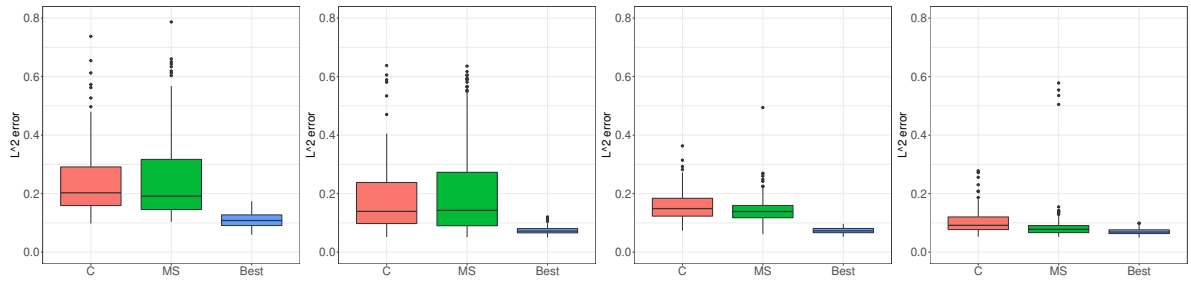
2.4.3 Real datasets

In this part, we consider two real datasets which are available on the UCI database. The first dataset is *Concrete Compressive Strength*. The concrete compressive strength is a highly nonlinear function of age and ingredients (Cement, Water, Blast Furnace Slag, ...). It contains 1030 observations of 8 numerical features. The output takes its values in $[2.330, 82.598]$. The second dataset is *Airfoil Self-Noise*. It is obtained from a series of aerodynamic and acoustic tests of two and three-dimensional airfoil



(a) $n = N = 100$ (b) $n = 1000, N = 100$ (c) $n = 100, N = 1000$ (d) $n = N = 1000$

Figure 2.4 – Boxplot of the empirical L^2-error of the estimators in Model 1 ($a = 1$)



(a) $n = N = 100$ (b) $n = 1000, N = 100$ (c) $n = 100, N = 1000$ (d) $n = N = 1000$

Figure 2.5 – Boxplot of the empirical L^2-error of the estimators in Model 2

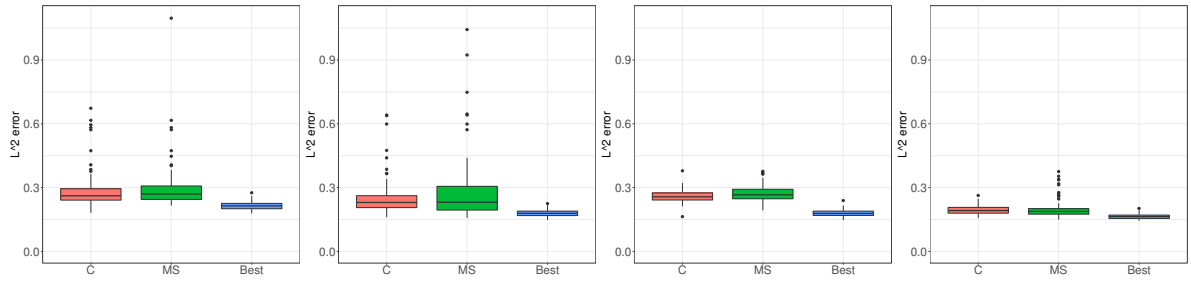
blade sections conducted in an anechoic wind tunnel. It contains 1503 observations of 5 numerical features. The output takes its values in $[103, 140]$. We display the histogram of an estimate of the variance function σ^2 produced by the random forest algorithm for both datasets in Figure 2.9. The estimated values of $\hat{\sigma}^2$ are large in two real datasets: 30% and 41% of the values are larger than 10 in Concrete Compressive Strength and Airfoil Self-Noise respectively. Now, we use the same steps in Section 2.4.2 to illustrate the performance of our methods with the following modifications: we reduce the dictionaries $\mathcal{F}$, $\mathcal{G}_1$ and $\mathcal{G}_2$ into 4 candidates: Lasso, kNN, random forest and support vector machines[2] methods where the parameters of the first two algorithms are chosen by cross-validation and the last two are chosen by default from `glmnet`, `FNN`, `randomForest` and `e1071` packages. We set $k \in \{5, 10, 13, 15, 17, 22, 35, 50, 75, 85, 100, 125\}$ for kNN. In step 4, based on $D_n \cup D_N$, we firstly compute the collection $\mathcal{F}$; secondly, for each estimate $\hat{f}_s$ in $\mathcal{F}$ we compute all possible true estimators $\{\hat{\sigma}^2_{s,m}\}_{1 \leq m \leq 4}$ of σ^2 corresponding to the 4 procedures in $\mathcal{F}$. In the last step, we compute the empirical L^2-risk of

- `MS`-method: $\frac{1}{T}\sum_{i=1}^{T}\left((Y_i - \hat{f}_{\text{MS}}(X_i))^2 - \hat{\sigma}^2_{\text{MS}}(X_i)\right)^2$;
- `C`-method: $\frac{1}{T}\sum_{i=1}^{T}\left((Y_i - \hat{f}_{\text{C}}(X_i))^2 - \hat{\sigma}^2_{\text{C}}(X_i)\right)^2$;
- `best`-method (best empirical L^2-risk): based on Step 4, we take the minimum of

$$\frac{1}{T}\sum_{i=1}^{T}\left((Y_i - \hat{f}_s(X_i))^2 - \hat{\sigma}^2_{s,m}(X_i)\right)^2 \quad \text{for all} \quad (s,m) \in [4] \times [4].$$

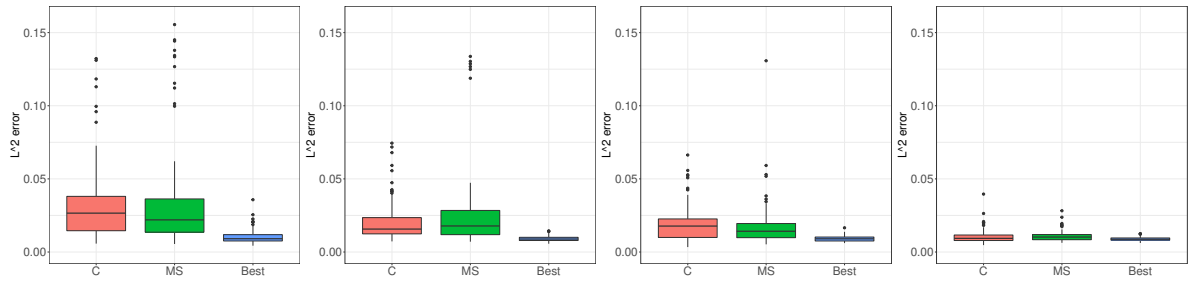
From this estimates, we compute the mean and the standard deviation of the empirical L^2-risk. The associated boxplots are given in Figure 2.10. Here, we fix $T = 200$. We take $n \in \{150, 415, 680\}$, $N \in \{680, 415, 150\}$ for the first real dataset and $n \in \{150, 652, 1153\}$, $N \in \{1153, 651, 150\}$ for the

[2] We simplify it by `svm` and use the R package with default parameters: `e1071`.



(a) $n = N = 100$ (b) $n = 1000, N = 100$ (c) $n = 100, N = 1000$ (d) $n = N = 1000$

Figure 2.6 – Boxplot of the L^2-error of the estimators in sparse model when $p = 50$, and $s = 14$.



(a) $n = N = 100$ (b) $n = 1000, N = 100$ (c) $n = 100, N = 1000$ (d) $n = N = 1000$

Figure 2.7 – Boxplot of the L^2-error of the estimators in Model 4.

second dataset. We observe that both of our aggregation methods achieve the same performance. Note that the aggregation methods are outperformed by the best method when n is large. This is mainly due to the fact that the method called " best" is computed without splitting the data (as explained before). Finally, we notice that it is advantageous to put a lot of data in the first dataset.

2.4.4 Application of variance function: Quantile regression

In this section, we illustrate the performance of our aggregation methods for quantile regression. Quantile regression allows a comprehensive analysis of the relationships between a response Y and input variables X. It is interesting in the entire conditional distribution of the dependent variable, and not only on its mean. For more details see for instance [44, 80, 83]. We recall that in our work we observe $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ such that

$$Y = f^*(X) + \sigma(X)\xi,$$

where ξ is the known noise with mean zero and unit variance. Let $\tau \in (0, 1)$. The conditional τ-quantile of Y given $X = x$, denoted by q_τ, is given by

$$q_\tau(Y|X) = \inf\{y : F_{Y|X}(y) \geq \tau\},$$

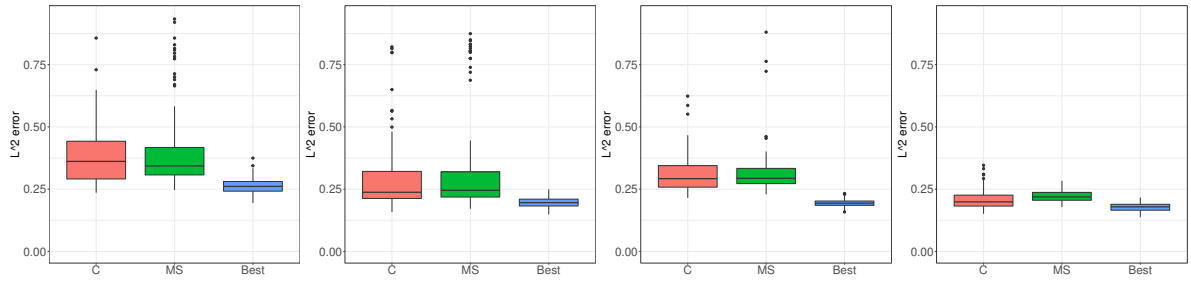
where $F_{Y|X}$ is the cumulative distribution function of $Y|X$. In particular, the quantity $q_\tau(Y|X)$ has the following form

$$q_\tau(Y|X) = f^*(X) + \sigma(X)F_\xi^{-1}(\tau),$$

where F_ξ is the cumulative distribution function of ξ and is known. The plug-in approach is a possible procedure to estimate q_τ. Given an estimator $\hat{f}$ of f^* and an estimator $\hat{\sigma}^2$ of σ^2, the plug-in estimator of q_τ is

$$\hat{q}_\tau(Y|X) = \hat{f}(X) + \hat{\sigma}(X)F_\xi^{-1}(\tau).$$

It is clear that a good estimate of the conditional τ-quantile is related to a good estimate of the regression function f^* and the variance function σ^2. We evaluate the performance of $\hat{q}_\tau(Y|X)$ according to



(a) $n = N = 100$ (b) $n = 1000, N = 100$ (c) $n = 100, N = 1000$ (d) $n = N = 1000$

Figure 2.8 – Boxplot of the L^2-error of the estimators in Model 5.

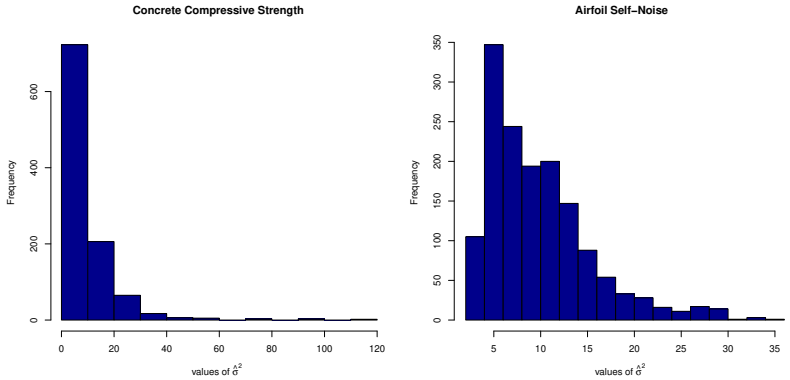


Figure 2.9 – Histogram of the values of the estimates of the variance function.

four different estimations of the regression function and the variance function: random forests, kNN, tree regression, svm and our two aggregation approaches, in the following gaussian model which was introduced in [83]

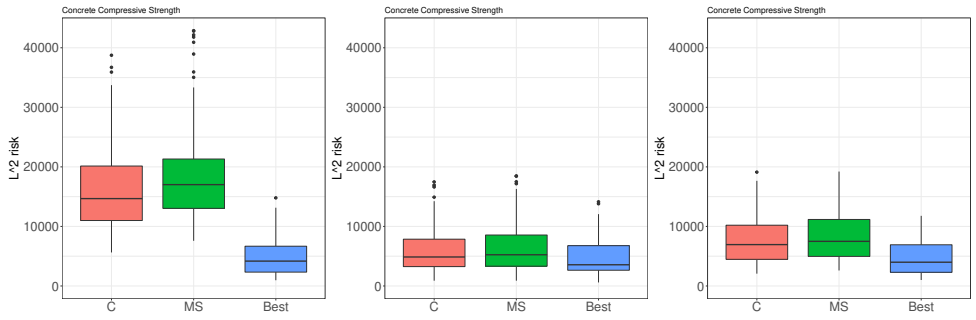
$$Y = \mathrm{sinc}(X) + 0.1 \exp(1 - X)\xi,$$

where X is drawn uniformly from $[-1, 1]$, sinc is the normalized sinc function, and $\xi \sim \mathcal{N}(0, 1)$. We fix $maxnodes = 25$ for the `rf` algorithm, and $n = N = T = 1000$. The performance of $\hat{q}_\tau(Y|X)$ is measured by the empirical quadratic error. The performances obtained from 100 independent runs, computed using the same methods mentioned in the previous section, are provided in Table 2.3, Figure 2.11 and Figure 2.12 for three different quantiles $\tau \in \{0.1, 0.5, 0.9\}$. For $\tau \in \{0.1, 0.9\}$ this is an estimate of the first and last deciles and for $\tau = 0.5$ an estimate of the median.

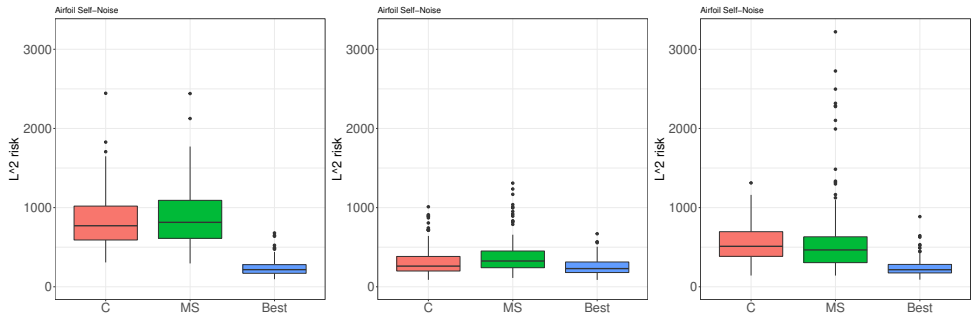
Table 2.3 – Performances of the six plug-in regression quantiles. We compute the means and standard deviations (between parentheses) of the L^2-error of $\hat{q}_\tau(Y|X)$.

	C	MS	knn	rf	svm	Tree
τ	$\widehat{\mathrm{Err}}$	$\widehat{\mathrm{Err}}$	$\widehat{\mathrm{Err}}$	$\widehat{\mathrm{Err}}$	$\widehat{\mathrm{Err}}$	$\widehat{\mathrm{Err}}$
0.1	0.0037 (0.0028)	0.0043 (0.0050)	0.0027 (0.0019)	0.0091 (0.0033)	0.0217 (0.0056)	0.0071 (0.0024)
0.5	0.0013 (0.0011)	0.0013 (0.0013)	0.0013 (0.0008)	0.0042 (0.0018)	0.0006 (0.0005)	0.0027 (0.0012)
0.9	0.0039 (0.0028)	0.0043 (0.0047)	0.0026 (0.0016)	0.0089 (0.0029)	0.0216 (0.0051)	0.0123 (0.0033)

Firstly, we can see here again that our both aggregation based approaches have the same performances. Secondly, the `svm` method which is built on the union of two samples $\mathcal{D}_n$ and $\mathcal{D}_N$ is the best method for the median problem estimation ($\tau = 0.5$) and `knn` method is the best procedure for

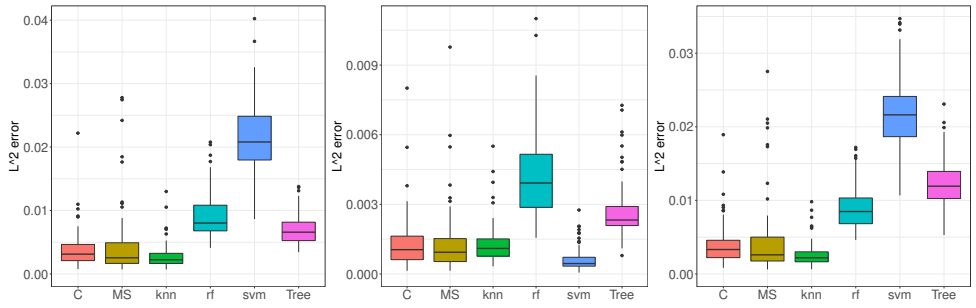


(a) $n = 150$, $N = 680$ (b) $n = 680$, $N = 150$ (c) $n = N = 415$



(d) $n = 150$, $N = 1153$ (e) $n = 1153$, $N = 150$ (f) $n = N = 652$

Figure 2.10 – Boxplots of the L^2-risk of our aggregation methods and the best method.



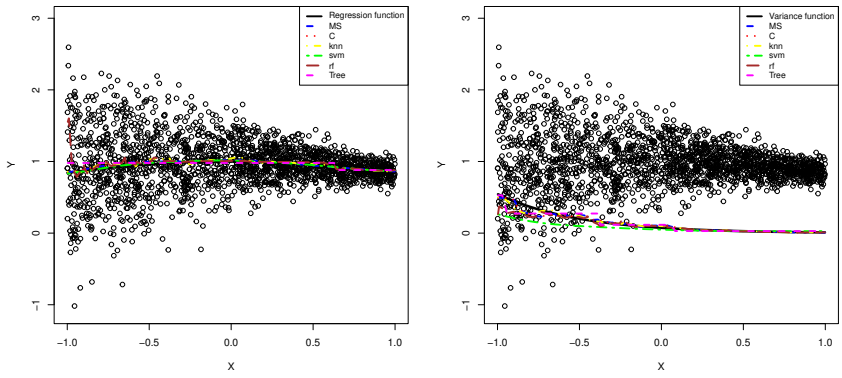
(a) $\tau = 0.1$ (b) $\tau = 0.5$ (c) $\tau = 0.9$

Figure 2.11 – Boxplots of the L^2-error of six methods with $\tau \in \{0.1, 0.5, 0.9\}$.

the decile function. Finally, we can deduce that a good estimation of the regression and variance functions ensures a good estimation of the conditional τ-quantile.

2.5 Conclusion

In the regression setting, we estimated the variance function by the model selection and convex aggregation when the set of initial estimators are constructed by the residual based-method. We called the estimators of the two procedures the `MS`-estimator and `C`-estimator, respectively. We established the consistency of our estimators under mild assumptions and provided rate of convergence for these methods in L_2-norm that are of order $O((\log(M_1)/N)^{1/8})$ when Y is satisfied the gaussian model; and $O((\log(M_1)/N)^{1/4})$ when Y is bounded. In future work, we are interested in studying on the one hand the optimality of our aggregation methods and on the other hand the performance of aggregation methods on high-dimensional features.



(a) Estimators of the regression function (b) Estimators of the variance function

Figure 2.12 – Curves of six estimators of the regression function and the variance function.

2.6 Appendix

This section gathers the proof of our results.

2.6.1 Proof of Theorem 9

Note that the quantity $\mathbb{E}\left[|\hat{\sigma}^2_{\text{MS}}(X) - \sigma^2(X)|^2\right]$ is the excess risk of the estimator $\hat{\sigma}^2_{\text{MS}}$ and defines as follows:

$$\mathbb{E}\left[|\hat{\sigma}^2_{\text{MS}}(X) - \sigma^2(X)|^2\right] := \mathbb{E}\left[R(\hat{\sigma}^2_{\text{MS}}) - R(\sigma^2)\right] \quad ,$$

where $R(\sigma^2) = \mathbb{E}\left[|Z - \sigma^2(X)|^2\right]$ is the true risk of the variance function. Besides, we introduce a minimizer of the risk R, denoted by $\bar{\sigma}^2_{\text{MS}}$ and given

$$\bar{\sigma}^2_{\text{MS}} := \hat{\sigma}^2_{\hat{s},\bar{m}} \ , \text{ where } \quad \bar{m} \in \underset{m\in[M_2]}{\operatorname{argmin}}\, R(\hat{\sigma}^2_{\hat{s},m}). \tag{2.9}$$

We consider the following decomposition

$$R(\hat{\sigma}^2_{\text{MS}}) - R(\sigma^2) = \underbrace{R(\hat{\sigma}^2_{\text{MS}}) - R(\bar{\sigma}^2_{\text{MS}})}_{\text{estimation error}} + \underbrace{R(\bar{\sigma}^2_{\text{MS}}) - R(\sigma^2)}_{\text{approximation error}}. \tag{2.10}$$

Each of these errors is obviously positive. The random term $R(\hat{\sigma}^2_{\text{MS}}) - R(\bar{\sigma}^2_{\text{MS}})$ is called the estimation error (or the variance). It measures how close $\hat{\sigma}^2_{\text{MS}}$ is to the best possible rule in $[M_2]$ in terms of the risk R. The deterministic term $R(\bar{\sigma}^2_{\text{MS}}) - R(\sigma^2)$ is called the approximation error (or the bias). We start with the following lemma

Lemma 1. *Let $\bar{\sigma}^2_{MS}$ be an aggregate defined in Equation (2.9). Then,*

$$\mathbb{E}\left[R(\bar{\sigma}^2_{MS}) - R(\sigma^2)\right] = \mathbb{E}\left[\min_{m\in[M_2]} \mathbb{E}_X\left[|\hat{\sigma}^2_{\hat{s},m}(X) - \sigma^2(X)|^2\right]\right].$$

This result explicitly determines the approximation error.

Proof of Lemma 1. For all $m \in [M_2]$, the excess risk of $\hat{\sigma}^2_{\hat{s},m}$ is given as follows

$$\mathbb{E}\left[|Z - \hat{\sigma}^2_{\hat{s},m}(X)|^2\right] - \mathbb{E}\left[|Z - \sigma^2(X)|^2\right] = \mathbb{E}_X\left[|\hat{\sigma}^2_{\hat{s},m}(X) - \sigma^2(X)|^2\right]. \tag{2.11}$$

We apply min in Equation (2.11) and we get

$$\mathbb{E}\left[R(\bar{\sigma}^2_{\text{MS}}) - R(\sigma^2)\right] \quad = \quad \mathbb{E}\left[\min_{m\in[M_2]} \mathbb{E}_X\left[|\hat{\sigma}^2_{\hat{s},m}(X) - \sigma^2(X)|^2\right]\right].$$

□

Proof of Theorem 9. We thank the decomposition in Eq. (2.10), we have

$$\mathbb{E}\left[|\hat{\sigma}^2_{\text{MS}}(X) - \sigma^2(X)|^2\right] = \mathbb{E}\left[R(\hat{\sigma}^2_{\text{MS}}) - R(\bar{\sigma}^2_{\text{MS}})\right] + \mathbb{E}\left[R(\bar{\sigma}^2_{\text{MS}}) - R(\sigma^2)\right]. \tag{2.12}$$

Step 1. Study of the term $\mathbb{E}\left[R(\bar{\sigma}^2_{\text{MS}}) - R(\sigma^2)\right]$. We begin with the following Lemma

Lemma 2. *Let $\hat{s}$ and s^* be two estimators defined in (2.1) and (2.7), respectively. Then, under Assumptions 1, 2, 3 and 4, there exists an absolute constant C such that*

$$\mathbb{P}\left(\hat{s} \neq s^*\right) \leq C\left(\frac{\log(M_1)}{N}\right)^{1/2}.$$

Proof. Under Assumption 4, we have firstly $\delta^*\left(\mathcal{D}_n\right) = \min_{s\neq s^*}\left\{|\mathcal{R}(\hat{f}_{s^*}) - \mathcal{R}(\hat{f}_s)|\right\} > \delta_0 > 0$. Recall that $\mathcal{R}(\hat{f}_{s^*}) \leq \mathcal{R}(\hat{f}_{\hat{s}})$. On the event $\{\hat{s} \neq s^*\}$, we have two cases

- $\hat{\mathcal{R}}_N(\hat{f}_{\hat{s}}) < \mathcal{R}(\hat{f}_{s^*})$, and then

$$\delta^*\left(\mathcal{D}_n\right) \leq |\mathcal{R}(\hat{f}_{\hat{s}}) - \mathcal{R}(\hat{f}_{s^*})| \leq |\hat{\mathcal{R}}_N(\hat{f}_{s^*}) - \mathcal{R}(\hat{f}_{s^*})| \leq \max_{s\in[M_1]}|\hat{\mathcal{R}}_N(\hat{f}_s) - \mathcal{R}(\hat{f}_s)|.$$

- $\hat{\mathcal{R}}_N(\hat{f}_{\hat{s}}) \geq \mathcal{R}(\hat{f}_{s^*})$, and then

$$\delta^*\left(\mathcal{D}_n\right) \leq |\hat{\mathcal{R}}_N(\hat{f}_{\hat{s}}) - \mathcal{R}(\hat{f}_{\hat{s}})| + |\hat{\mathcal{R}}_N(\hat{f}_{s^*}) - \mathcal{R}(\hat{f}_{s^*})| \leq 2\max_{s\in[M_1]}|\hat{\mathcal{R}}_N(\hat{f}_s) - \mathcal{R}(\hat{f}_s)|.$$

Therefore,

$$\mathbb{P}\left(\hat{s} \neq s^*\right) \leq \mathbb{P}\left(\max_{s\in[M_1]}|\hat{\mathcal{R}}_N(\hat{f}_s) - \mathcal{R}(\hat{f}_s)| \geq \delta_0/2\right)$$

We control this term using Bernstein's inequality. We check that the conditions for Bernstein's inequality are satisfied. For all $s \in [M_1]$, set $V_i(s) = |Y_i - \hat{f}_s(X_i)|^2 = |f^*(X_i) - \hat{f}_s(X_i) + \sigma(X_i)\xi_i|^2$ for all $i = n+1, \ldots, n+N$. First, Assumptions 1 and 3 ensure that there exist a positive constants L_1 and L_2 such that $|f^*(X) - \hat{f}_s(X)| \leq L_1$ and $|\sigma^2(X)| \leq L_2$. Second, note that since the variables $V_i(s)$ are i.i.d. and by the elementary inequality $(x+y)^4 \leq 2^3(x^4+y^4)$ for all $x, y \in \mathbb{R}$, by Lemma 4, and by the elementary inequality $x^4 + y^4 \leq (x+y)^4$ for all $x, y \geq 0$ we have

$$\sum_{i=n+1}^{n+N}\mathbb{E}\left[V_i^2(s)\right] \leq 2^3\sum_{i=n+1}^{n+N}\mathbb{E}\left[|f^*(X) - \hat{f}_s(X)|^4 + \sigma^4(X_i)\xi_i^4\right] \leq 2^7 N(L_1 + \sqrt{L_2})^4 := v_N \ ,$$

and for $k \geq 3$ we follow the elementary inequality $(x+y)^{2k} \leq 2^{2k-1}(x^{2k}+y^{2k})$ for all $x, y \in \mathbb{R}$, Lemma 4, and the following elementary inequality $x^{2k} + y^{2k} \leq (x+y)^{2k}$ for all $x, y \geq 0$

$$\begin{aligned}
\sum_{i=n+1}^{n+N}\mathbb{E}\left[(V_i^k(s) \vee 0)\right] &= \sum_{i=n+1}^{n+N}\mathbb{E}\left[|f^*(X_i) - \hat{f}_s(X_i) + \sigma(X_i)\xi_i|^{2k}\right] \\
&\leq 2^{2k-1}\sum_{i=n+1}^{n+N}\mathbb{E}\left[|f^*(X_i) - \hat{f}_s(X_i)|^{2k} + |\sigma^2(X_i)|^k|\xi_i|^{2k}\right] \\
&\leq \frac{1}{2}2^{2k}N\left(L_1^{2k} + 2^{k+1}(\sqrt{L_2})^{2k}(k)!\right) \\
&\leq \frac{1}{2}2^{3k+1}N\left(L_1 + \sqrt{L_2}\right)^{2k}k! \\
&\leq \frac{1}{2}v_N c^{k-2}k! \ .
\end{aligned}$$

where $c := 8\left(L_1 + \sqrt{L_2}\right)^2$. Using the Bernstein's inequality (Lemma 7), we get for all $s \in [M_1]$

$$\mathbb{P}\left(|\hat{\mathcal{R}}_N(\hat{f}_s) - \mathcal{R}(\hat{f}_s)| \geq \frac{\delta_0}{2}\right) \leq 2\exp\left(-\frac{N\delta_0^2}{2^{10}(L_1+\sqrt{L_2})^4 + 4c\delta_0}\right)$$

By union bound on $s \in [M_1]$, we obtain

$$\mathbb{P}\left(\hat{s} \neq s^*\right) \leq 2\exp\left(\log(M_1) - \frac{N\delta_0^2}{2^{10}(L_1+\sqrt{L_2})^4 + 4c\delta_0}\right) \leq C\left(\frac{\log(M_1)}{N}\right)^{1/2} ,$$

where C is a positive constant which depends on L_1, L_2 and δ_0. □

By Lemmas 1 and 2, and under Assumptions 1 and 3 we get

$$\begin{aligned}\mathbb{E}\left[R(\bar{\sigma}^2_{\text{MS}}) - R(\sigma^2)\right] &= \mathbb{E}\left[\min_{m\in[M_2]} \mathbb{E}_X\left[|\hat{\sigma}^2_{\hat{s},m}(X) - \sigma^2(X)|^2\left\{\mathbb{1}_{\{\hat{s}=s^*\}} + \mathbb{1}_{\{\hat{s}\neq s^*\}}\right\}\right]\right] \\ &\leq \mathbb{E}\left[\min_{m\in[M_2]} \mathbb{E}_X\left[|\hat{\sigma}^2_{s^*,m}(X) - \sigma^2(X)|^2\right]\right] + C\left(\frac{\log(M_1)}{N}\right)^{1/2}\end{aligned}$$

where C is a constant which depends on K_2 , σ^2 and the constant in Lemma 2.

Step 2. Study of the term $\mathbb{E}\left[R(\hat{\sigma}^2_{\text{MS}}) - R(\bar{\sigma}^2_{\text{MS}})\right]$. To treat the estimation error, we introduce an aggregate $\tilde{\sigma}^2_{\text{MS}}$ which is based on minimization of the empirical risk of R

$$\tilde{\sigma}^2_{\text{MS}} := \hat{\sigma}^2_{\hat{s},\tilde{m}} \ , \text{ where } \quad \tilde{m} \in \operatorname*{argmin}_{m\in[M_2]} R_N(\hat{\sigma}^2_{\hat{s},m}) \ ,$$

with $R_N(\hat{\sigma}^2_{\hat{s},m}) = \frac{1}{N}\sum_{i=n+1}^{n+N} |Z_i - \hat{\sigma}^2_{\hat{s},m}(X_i)|^2$. Moreover, we consider the decomposition

$$\mathbb{E}\left[R(\hat{\sigma}^2_{\text{MS}}) - R(\bar{\sigma}^2_{\text{MS}})\right] = \mathbb{E}\left[R(\hat{\sigma}^2_{\text{MS}}) - R(\tilde{\sigma}^2_{\text{MS}})\right] + \mathbb{E}\left[R(\tilde{\sigma}^2_{\text{MS}}) - R(\bar{\sigma}^2_{\text{MS}})\right].$$

Step 2.1. Study of the term $\mathbb{E}\left[R(\tilde{\sigma}^2_{\text{MS}}) - R(\bar{\sigma}^2_{\text{MS}})\right]$. We decompose the term $\mathbb{E}\left[R(\tilde{\sigma}^2_{\text{MS}}) - R(\bar{\sigma}^2_{\text{MS}})\right]$ into two positive terms

$$\mathbb{E}\left[R(\tilde{\sigma}^2_{\text{MS}}) - R(\bar{\sigma}^2_{\text{MS}})\right] = \mathbb{E}\left[R(\tilde{\sigma}^2_{\text{MS}}) - R_N(\tilde{\sigma}^2_{\text{MS}})\right] + \mathbb{E}\left[R_N(\tilde{\sigma}^2_{\text{MS}}) - R(\bar{\sigma}^2_{\text{MS}})\right]. \tag{2.13}$$

We use the fact that $R_N(\tilde{\sigma}^2_{\text{MS}}) \leq R_N(\bar{\sigma}^2_{\text{MS}})$ in Eq. (2.13), and we get the uniform bound

$$\mathbb{E}\left[R(\tilde{\sigma}^2_{\text{MS}}) - R(\bar{\sigma}^2_{\text{MS}})\right] \leq 2\mathbb{E}\left[\max_{(s,m)\in[M_1]\times[M_2]} |R_N(\hat{\sigma}^2_{s,m}) - R(\hat{\sigma}^2_{s,m})|\right].$$

Then using Assumption 2, for some $(s,m) \in [M_1]\times[M_2]$, set $T_i(s,m) = |Z_i - \hat{\sigma}^2_{s,m}(X_i)|^2 = |\sigma^2(X_i)\xi_i^2 - \hat{\sigma}^2_{s,m}(X_i)|^2$ for all $i = n+1,\dots,n+N$. First, note that since the variables $T_i(s,m)$ are i.i.d. , conditionally on $\mathcal{D}_n$ we have

$$\begin{aligned}|R_N(\hat{\sigma}^2_{s,m}) - R(\hat{\sigma}^2_{s,m})| &= \left|\frac{1}{N}\sum_{i=n+1}^{n+N}(T_i(s,m) - \mathbb{E}[T_i(s,m)])\right| \\ &\leq \left|\frac{1}{N}\sum_{i=n+1}^{n+N}(T_i(s,m) - \mathbb{E}[T_i(s,m)])\mathbb{1}_{\{|\xi_i|\leq L\}}\right| \\ &\quad + \left|\frac{1}{N}\sum_{i=n+1}^{n+N}(T_i(s,m) - \mathbb{E}[T_i(s,m)])\mathbb{1}_{\{|\xi_i|> L\}}\right|\end{aligned}$$

for any $L > 0$. Therefore, conditionally on $\mathcal{D}_n$

$$\begin{aligned}\mathbb{E}\left[\max_{(s,m)\in[M_1]\times[M_2]} |R_N(\hat{\sigma}^2_{s,m}) - R(\hat{\sigma}^2_{s,m})|\right] &\leq \mathbb{E}\left[\max_{(s,m)\in[M_1]\times[M_2]}\left|\frac{1}{N}\sum_{i=n+1}^{n+N}(T_i(s,m) - \mathbb{E}[T_i(s,m)])\mathbb{1}_{\{|\xi_i|\leq L\}}\right|\right] \\ &\quad + \mathbb{E}\left[\max_{(s,m)\in[M_1]\times[M_2]}\left|\frac{1}{N}\sum_{i=n+1}^{n+N}(T_i(s,m) - \mathbb{E}[T_i(s,m)])\mathbb{1}_{\{|\xi_i|> L\}}\right|\right].\end{aligned} \tag{2.14}$$

Step 2.1.1. We control the first term on the r.h.s. of Eq. (2.14). On the event $\{|\xi| \leq L\}$ and under Assumptions 1 and 3, we get $|T_i(s,m)| \leq c_1 L^4 + 2K_2^2$ for all $i = n+1, \ldots, n+N$ for some $c_1 > 0$ that depends on σ^2. Conditionally on $\mathcal{D}_n$, we apply Hoeffding's inequality, for all $(s,m) \in [M_1] \times [M_2]$, and all $t \geq 0$

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=n+1}^{n+N}(T_i(s,m) - \mathbb{E}[T_i(s,m)])\mathbb{1}_{\{|\xi_i|\leq L\}}\right| \geq t\right) \leq 2\exp\left(-\frac{Nt^2}{2(c_1L^4 + 2K_2^2)^2}\right) ,$$

Conditionally on $\mathcal{D}_n$, by a union bound on $(s,m) \in [M_1] \times [M_2]$, we deduce that for all $t \geq 0$

$$\mathbb{P}\left(\max_{(s,m)\in[M_1]\times[M_2]}\left|\frac{1}{N}\sum_{i=n+1}^{n+N}(T_i(s,m) - \mathbb{E}[T_i(s,m)])\mathbb{1}_{\{|\xi_i|\leq L\}}\right| \geq t\right) \leq 2\exp\left(\log(M_1M_2) - \frac{Nt^2}{2(c_1L^4 + 2K_2^2)^2}\right).$$

We apply Lemma 6. Then, there exists a positive constant $\mathbf{c}$ such that

$$\mathbb{E}\left[\max_{(s,m)\in[M_1]\times[M_2]}\left|\frac{1}{N}\sum_{i=n+1}^{n+N}(T_i(s,m) - \mathbb{E}[T_i(s,m)])\mathbb{1}_{\{|\xi_i|\leq L\}}\right|\right] \leq \mathbf{c}\left(c_2L^4 + c_3\right)\left(\frac{\log(M_1M_2)}{N}\right)^{1/2},$$

where c_2 is a positive constant that depends on c_1 and c_3 depends on K_2.

Step 2.1.2. We control the second term on the r.h.s. of Eq. (2.14). By union bound on $(s,m) \in [M_1] \times [M_2]$, by Cauchy–Schwarz inequality, under Assumptions 1, 2 and 3, and Lemma 3 we obtain

$$\begin{aligned}
&\mathbb{E}\left[\max_{(s,m)\in[M_1]\times[M_2]}\left|\frac{1}{N}\sum_{i=n+1}^{n+N}(T_i(s,m) - \mathbb{E}[T_i(s,m)])\mathbb{1}_{\{|\xi_i|>L\}}\right|\right] \\
\leq\ &\sum_{s=1}^{M_1}\sum_{m=1}^{M_2}\frac{1}{N}\sum_{i=n+1}^{n+N}\mathbb{E}[|T_i(s,m) - \mathbb{E}[T_i(s,m)]|\mathbb{1}_{\{\xi_i>L\}}] \\
\leq\ &\sum_{s=1}^{M_1}\sum_{m=1}^{M_2}\frac{1}{N}\sum_{i=n+1}^{n+N}\sqrt{\mathbb{E}[|T_i(s,m) - \mathbb{E}[T_i(s,m)]|^2]\mathbb{P}(|\xi_i| > L)} \\
\leq\ &cM_1M_2\sqrt{\mathbb{P}(|\xi_1| > L)} \\
\leq\ &cM_1M_2\frac{\exp(-L^2/4)}{L^{1/2}} ,
\end{aligned}$$

where c is a positive constant which depends on ξ, σ^2 and K_2.
Combining the results of the **Step 2.1.1** and **Step 2.1.2** in Eq. (2.14)

$$\mathbb{E}\left[\max_{(s,m)\in[M_1]\times[M_2]}|R_N(\hat{\sigma}^2_{s,m}) - R(\hat{\sigma}^2_{s,m})|\right] \leq \mathbf{c}(c_2L^4 + c_3)\left(\frac{\log(M_1M_2)}{N}\right)^{1/2} + cM_1M_2\frac{\exp(-L^2/4)}{L^{1/2}}.$$

Choosing $L = 2\sqrt{\log(N)}$ and we get

$$\mathbb{E}\left[\max_{(s,m)\in[M_1]\times[M_2]}|R_N(\hat{\sigma}^2_{s,m}) - R(\hat{\sigma}^2_{s,m})|\right] \leq C\left(\frac{\log(N)^4\log(M_1M_2)}{N}\right)^{1/2} ,$$

where C is a positive constant that depends on c_2 and $\mathbf{c}$, and **Step 2.1.2** is finished.

We combine the results of the **Step 2.1.1** and **Step 2.1.2** and we get the following bound

$$\mathbb{E}\left[R(\tilde{\sigma}^2_{\text{MS}}) - R(\bar{\sigma}^2_{\text{MS}})\right] \leq 2C\left(\frac{\log(N)^4\log(M_1M_2)}{N}\right)^{1/2}.$$

Remark 1. *It is clear that when Y is bounded, there exists an absolute constant $C > 0$ such that*

$$\mathbb{E}\left[R(\tilde{\sigma}^2_{MS}) - R(\bar{\sigma}^2_{MS})\right] \leq C\left(\frac{\log(M_1M_2)}{N}\right)^{1/2}.$$

Step 2.2. Study of the term $\mathbb{E}\left[R(\hat{\sigma}^2_{\text{MS}}) - R(\tilde{\sigma}^2_{\text{MS}})\right]$. We start with the following decomposition

$$\mathbb{E}\left[R(\hat{\sigma}^2_{\text{MS}}) - R(\tilde{\sigma}^2_{\text{MS}})\right] = \mathbb{E}\left[R(\hat{\sigma}^2_{\text{MS}}) - R_N(\hat{\sigma}^2_{\text{MS}})\right] + \mathbb{E}\left[R_N(\hat{\sigma}^2_{\text{MS}}) - R_N(\tilde{\sigma}^2_{\text{MS}})\right] + \mathbb{E}\left[R_N(\tilde{\sigma}^2_{\text{MS}}) - R(\tilde{\sigma}^2_{\text{MS}})\right]. \tag{2.15}$$

We use the same arguments in **Step 2.1.** to control the first term and the last term on the r.h.s. of Eq. (2.15), and we get the following bound

$$\begin{aligned}\mathbb{E}\left[R(\hat{\sigma}^2_{\text{MS}}) - R_N(\hat{\sigma}^2_{\text{MS}})\right] + \mathbb{E}\left[R_N(\tilde{\sigma}^2_{\text{MS}}) - R(\tilde{\sigma}^2_{\text{MS}})\right] &\leq 2\mathbb{E}\left[\max_{(s,m)\in[M_1]\times[M_2]} |R_N(\hat{\sigma}^2_{s,m}) - R(\hat{\sigma}^2_{s,m})|\right] \\ &\leq C\left(\frac{\log(N)^4\log(M_1M_2)}{N}\right)^{1/2}.\end{aligned}$$

Remark 2. *If Y is bounded, there exists an absolute constant $C > 0$ such that*

$$\mathbb{E}\left[R(\hat{\sigma}^2_{MS}) - R_N(\hat{\sigma}^2_{MS})\right] + \mathbb{E}\left[R_N(\tilde{\sigma}^2_{MS}) - R(\tilde{\sigma}^2_{MS})\right] \leq C\left(\frac{\log(M_1M_2)}{N}\right)^{1/2}.$$

We now study the second term on the r.h.s. of Eq. (2.15). For that, we need the following decomposition

$$\mathbb{E}\left[R_N(\hat{\sigma}^2_{\text{MS}}) - R_N(\tilde{\sigma}^2_{\text{MS}})\right] = \mathbb{E}\left[R_N(\hat{\sigma}^2_{\text{MS}}) - \hat{R}_N(\hat{\sigma}^2_{\text{MS}})\right] + \mathbb{E}\left[\hat{R}_N(\hat{\sigma}^2_{\text{MS}}) - R_N(\tilde{\sigma}^2_{\text{MS}})\right]. \tag{2.16}$$

Using $\hat{R}_N(\hat{\sigma}^2_{\text{MS}}) \leq \hat{R}_N(\tilde{\sigma}^2_{\text{MS}})$ in Eq. (2.16), we obtain the following inequality

$$\mathbb{E}\left[R_N(\hat{\sigma}^2_{\text{MS}}) - R_N(\tilde{\sigma}^2_{\text{MS}})\right] \leq 2\mathbb{E}\left[\max_{m\in[M_2]} |\hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) - R_N(\hat{\sigma}^2_{\hat{s},m})|\right].$$

We control the term $\mathbb{E}\left[\max_{m\in[M_2]} |\hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) - R_N(\hat{\sigma}^2_{\hat{s},m})|\right]$. By definition of $\hat{R}_N$ and R_N, and under Assumption 3, we get for all $m \in [M_2]$

$$|\hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) - R_N(\hat{\sigma}^2_{\hat{s},m})| \leq \frac{1}{N}\sum_{i=n+1}^{n+N} |\hat{Z}_i - Z_i|^2 + \frac{2}{N}\sum_{i=n+1}^{n+N} |\hat{Z}_i - Z_i|(|Z_i| + K_2),$$

where K_2 is the bound of $\hat{\sigma}^2_{\hat{s},m}$. The upper-bound of $|\hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) - R_N(\hat{\sigma}^2_{\hat{s},m})|$ does not depend on m, therefore

$$\mathbb{E}\left[\max_{m\in[M_2]} |\hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) - R_N(\hat{\sigma}^2_{\hat{s},m})|\right] \leq \mathbb{E}\left[\frac{1}{N}\sum_{i=n+1}^{n+N} |\hat{Z}_i - Z_i|^2\right] + 2\mathbb{E}\left[\frac{1}{N}\sum_{i=n+1}^{n+N} |\hat{Z}_i - Z_i|(|Z_i| + K_2)\right]. \tag{2.17}$$

Note that, by Assumptions 1 and 3 we obtain for all $i = n+1, \ldots, n+N$

$$|f^*(X_i) - \hat{f}_{\text{MS}}(X_i)| \leq \|f^*\|_\infty + \max_{s\in[M_1]} \|\hat{f}_s\|_\infty \leq \|f^*\|_\infty + K_1 \leq L_1 < \infty.$$

Since $x^2 - y^2 = (x-y)(x+y)$, $(x+y)^2 \leq 2(x^2+y^2)$, we obtain the following inequality for all $i = n+1, \ldots, n+N$

$$\begin{aligned}|\hat{Z}_i - Z_i|^2 &= |(Y_i - \hat{f}_{\text{MS}}(X_i))^2 - (Y_i - f^*(X_i))^2|^2 \\ &= |(f^*(X_i) - \hat{f}_{\text{MS}}(X_i))(2(Y_i - f^*(X_i)) + (f^*(X_i) - \hat{f}_{\text{MS}}(X_i))|^2 \\ &\leq |f^*(X_i) - \hat{f}_{\text{MS}}(X_i)|^2|(8|Y_i - f^*(X_i)|^2 + 2L_1^2|,\end{aligned} \tag{2.18}$$

Control of $\mathbb{E}\left[\frac{1}{N}\sum_{i=n+1}^{n+N} |\hat{Z}_i - Z_i|^2\right]$. First, since Assumptions 1-2 are satisfied, we have that for all

$i = n+1, \ldots, n+N$, $\mathbb{E}\left[|Y_i - f^*(X_i)|^4\right] \leq k_1 < \infty$. Second, by inequality (2.18), Cauchy-Schwarz inequality, Jensen's inequality, and under Assumptions 1, and 2, one gets

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{N}\sum_{i=n+1}^{n+N}|\hat{Z}_i - Z_i|^2\right] &\leq 2L_1^2\mathbb{E}\left[\|\hat{f}_{\text{MS}} - f^*\|_N^2\right] + \frac{8}{N}\sum_{i=n+1}^{n+N}\mathbb{E}\left[|Y_i - f^*(X_i)|^2|f^*(X_i) - \hat{f}_{\text{MS}}(X_i)|^2\right] \\
&\leq 2L_1^2\mathbb{E}\left[\|\hat{f}_{\text{MS}} - f^*\|_N^2\right] + \frac{8}{N}\sum_{i=n+1}^{n+N}\sqrt{\mathbb{E}\left[|Y_i - f^*(X_i)|^4\right]}\sqrt{\mathbb{E}\left[|f^*(X_i) - \hat{f}_{\text{MS}}(X_i)|^4\right]} \\
&\leq 2L_1^2\mathbb{E}\left[\|\hat{f}_{\text{MS}} - f^*\|_N^2\right] + \frac{8\sqrt{k_1}L_1}{N}\sum_{i=n+1}^{n+N}\sqrt{\mathbb{E}\left[|f^*(X_i) - \hat{f}_{\text{MS}}(X_i)|^2\right]} \\
&\leq 2L_1^2\mathbb{E}\left[\|\hat{f}_{\text{MS}} - f^*\|_N^2\right] + 8\sqrt{k_1}L_1\sqrt{\mathbb{E}\left[\frac{1}{N}\sum_{i=n+1}^{n+N}|f^*(X_i) - \hat{f}_{\text{MS}}(X_i)|^2\right]} \\
&\leq C_1\sqrt{\mathbb{E}\left[\|\hat{f}_{\text{MS}} - f^*\|_N^2\right]}\ ,
\end{aligned}$$

where C_1 is a positive constant that depends on k_1 and L_1.

Control of $\mathbb{E}\left[\frac{1}{N}\sum_{i=n+1}^{n+N}|\hat{Z}_i - Z_i|(|Z_i| + K_2)\right]$**.** First, since Assumptions 1-2 are satisfied, we have that for all $i = n+1, \ldots, n+N$, $\mathbb{E}\left[(|Y_i - f^*(X_i)|^2 + K_2)^2\right] \leq k_2 < \infty$. Second, by Cauchy-Schwarz inequality and Jensen's inequality, one gets

$$\begin{aligned}
\frac{1}{N}\sum_{i=n+1}^{n+N}\mathbb{E}\left[|\hat{Z}_i - Z_i|(|Z_i| + K_2)\right] &\leq \frac{1}{N}\sum_{i=n+1}^{n+N}\sqrt{\mathbb{E}\left[|\hat{Z}_i - Z_i|^2\right]}\sqrt{\mathbb{E}\left[(|Y_i - f^*(X_i)|^2 + K_1)^2\right]} \\
&\leq \frac{\sqrt{k_2}}{N}\sum_{i=n+1}^{n+N}\sqrt{\mathbb{E}\left[|\hat{Z}_i - Z_i|^2\right]} \\
&\leq \sqrt{k_2}\sqrt{\mathbb{E}\left[\frac{1}{N}\sum_{i=n+1}^{n+N}|\hat{Z}_i - Z_i|^2\right]} \\
&\leq C_2\mathbb{E}\left[\|\hat{f}_{\text{MS}} - f^*\|_N^2\right]^{1/4}\ ,
\end{aligned}$$

where C_2 is a positive constant that depends on C_1 and k_2. Thus, there exists an absolute constant C such that

$$\mathbb{E}\left[\max_{m\in[M_2]}|\hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) - R_N(\hat{\sigma}^2_{\hat{s},m})|\right] \leq C\mathbb{E}\left[\|\hat{f}_{\text{MS}} - f^*\|_N^2\right]^{1/4}.$$

We need the following proposition:

Proposition 2. *Let $\hat{f}_{MS}$ the aggregate defined in Eq. (2.2). Then, under Assumptions 2 and 3 there exists an absolute constant C such that*

$$\mathbb{E}\left[\|\hat{f}_{MS} - f^*\|_N^2\right] \leq \min_{s\in[M_1]}\mathbb{E}\left[\|\hat{f}_s - f^*\|_N^2\right] + C\left(\frac{\log(M_1)}{N}\right)^{1/2}.$$

This result studies the upper-bound of empirical norm risk of the aggregate $\hat{f}_{\text{MS}}$ and the proof of it exists in [88]. Besides, the Proposition 2 and the elementary inequality $(x+y)^{1/4} \leq x^{1/4} + y^{1/4}$ for all $x, y \geq 0$ give us the following inequality

$$\mathbb{E}\left[R(\hat{\sigma}^2_{\text{MS}}) - R(\tilde{\sigma}^2_{\text{MS}})\right] \leq C'\left\{\min_{s\in[M_1]}\mathbb{E}\left[\|\hat{f}_s - f^*\|_N^2\right]\right\}^{1/4} + C"\left(\frac{\log(M_1)}{N}\right)^{1/8}\ ,$$

where C' is a constant which depends on C_2 and $C"$ is a constant which depends on C_2 and the constant in Proposition 2.

Merging the results of the **Step 1** and **Step 2** in Eq. (2.12) and we get the result.

Remark 3. *In the case where Y is bounded and from Eq.* (2.18)*, we observe that there exists a constant C_3 such that*

$$|\hat{Z}_i - Z_i|^2 \leq C_3 |f^*(X_i) - \hat{f}_{MS}(X_i)|^2 \ . \tag{2.19}$$

By Jensen's inequality twice an inequality (2.17) *and from Eq.*(2.19)*, one gets there exists an absolute constant C_4 such that*

$$\mathbb{E}\left[\max_{m\in[M_2]} |\hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) - R_N(\hat{\sigma}^2_{\hat{s},m})|\right] \leq C_4 \mathbb{E}\left[\|\hat{f}_{MS} - f^*\|_N^2\right]^{1/2}. \tag{2.20}$$

Finally, we apply Proposition 2 in Eq.(2.20) *to get the result.*

□

2.6.2 Proof of Proposition 2

From the definition of MS-estimator $\hat{f}_{\text{MS}}$, we get by a simple algebra that, for any $s \in [M_1]$

$$\|\hat{f}_{\text{MS}} - f^*\|_N^2 \leq \|\hat{f}_s - f^*\|_N^2 + 2 < \hat{f}_{\text{MS}} - \hat{f}_s, Y - f^* >,$$

where $< \hat{f}_{\text{MS}} - \hat{f}_s, Y - f^* > := \frac{1}{N}\sum_{i=n+1}^{n+N} \left((\hat{f}_{\text{MS}}(X_i) - \hat{f}_s(X_i))(Y_i - f^*(X_i))\right)$. Therefore, one gets for any $s \in [M_1]$

$$\mathbb{E}\left[\|\hat{f}_{\text{MS}} - f^*\|_N^2\right] \leq \mathbb{E}\left[\|\hat{f}_s - f^*\|_N^2\right] + 2\mathbb{E}\left[< \hat{f}_{\text{MS}} - \hat{f}_s, Y - f^* >\right]. \tag{2.21}$$

We control the second term in the r.h.s. of Eq (2.21). Firstly, we notice that

$$\mathbb{E}\left[< \hat{f}_{\text{MS}} - \hat{f}_s, Y - f^* >\right] \leq \mathbb{E}\left[\max_{1\leq j\leq M_1} < \hat{f}_j - \hat{f}_s, Y - f^* >\right].$$

Secondly, since $Y - f^*$ is ρ-subgaussian where ρ is a positive constant that depends on $Y - f^*$, then the variables $< \hat{f}_j - \hat{f}_s, Y - f^* >$ is $\bar{\rho}$-subgaussian where $\bar{\rho}^2 = \frac{\rho^2 \|\hat{f}_j - \hat{f}_s\|_N^2}{N}$. Moreover, under Assumption 3, it is clear that $\max_{1\leq j\leq M_1} \|\hat{f}_j - \hat{f}_s\|_N^2 \leq B$ where B is a constant which depends on K_1. Therefore, we use Lemma 5 and we get

$$\mathbb{E}\left[\max_{1\leq j\leq M_1} < \hat{f}_j - \hat{f}_s, Y - f^* >\right] \leq \rho\sqrt{B}\sqrt{\frac{2\log(M_1)}{N}}.$$

Thus,

$$\mathbb{E}\left[\|\hat{f}_{\text{MS}} - f^*\|_N^2\right] \leq \min_{s\in[M_1]} \mathbb{E}\left[\|\hat{f}_s - f^*\|_N^2\right] + 2\rho\sqrt{B}\sqrt{\frac{2\log(M_1)}{N}}.$$

2.6.3 Proof of Theorem 10

We introduce the following aggregates

$$\tilde{\sigma}^2_{\text{C}} := \hat{\sigma}^2_{\hat{\lambda},\tilde{\beta}} \ , \text{ where } \quad \tilde{\beta} \in \operatorname*{argmin}_{\beta\in\Lambda^{M_2}} R_N(\hat{\sigma}^2_{\hat{\lambda},\beta}) \ ,$$

and

$$\bar{\sigma}^2_{\text{C}} := \hat{\sigma}^2_{\hat{\lambda},\bar{\beta}} \ , \text{ where } \quad \bar{\beta} \in \operatorname*{argmin}_{\beta\in\Lambda^{M_2}} R(\hat{\sigma}^2_{\hat{\lambda},\beta}).$$

Consider the following decomposition

$$\mathbb{E}\left[|\hat{\sigma}^2_{\text{C}}(X) - \sigma^2(X)|^2\right] = \mathbb{E}\left[R(\hat{\sigma}^2_{\text{C}}) - R(\tilde{\sigma}^2_{\text{C}})\right] + \mathbb{E}\left[R(\tilde{\sigma}^2_{\text{C}}) - R(\bar{\sigma}^2_{\text{C}})\right] + \mathbb{E}\left[R(\bar{\sigma}^2_{\text{C}}) - R(\sigma^2)\right]. \tag{2.22}$$

Step 1. Study of the term $\mathbb{E}\left[R(\bar{\sigma}_{\mathrm{C}}^2) - R(\sigma^2)\right]$. We use the same proof of Lemma 1, and we get

$$\mathbb{E}\left[R(\bar{\sigma}_{\mathrm{C}}^2) - R(\sigma^2)\right] \leq \mathbb{E}\left[\inf_{\beta \in \Lambda^{M_2}} \mathbb{E}_X\left[|\hat{\sigma}^2_{\hat{\lambda},\beta}(X) - \sigma^2(X)\right]\right].$$

Step 2. Study of the term $\mathbb{E}\left[R(\tilde{\sigma}_{\mathrm{C}}^2) - R(\bar{\sigma}_{\mathrm{C}}^2)\right]$. We use the fact that $R_N(\tilde{\sigma}_{\mathrm{C}}^2) \leq R_N(\bar{\sigma}_{\mathrm{C}}^2)$, and we get the uniform bound

$$\mathbb{E}\left[R(\tilde{\sigma}_{\mathrm{C}}^2) - R(\bar{\sigma}_{\mathrm{C}}^2)\right] \leq 2\mathbb{E}\left[\sup_{(\lambda,\beta)\in\Lambda^{M_1}\times\Lambda^{M_2}} |R_N(\hat{\sigma}^2_{\lambda,\beta}) - R(\hat{\sigma}^2_{\lambda,\beta})|\right].$$

Since Λ^{M_2} (resp. Λ^{M_1}) is compact, we have $\Lambda^{M_2} \subset \bar{B}(0,1)$ (the closed unit ball) (resp. $\Lambda^{M_1} \subset \bar{B}(0,1)$), and there exists an ϵ_2-net $\Lambda^{M_2}_{\epsilon_2}$ of Λ^{M_2} (resp. an ϵ_1-net $\Lambda^{M_1}_{\epsilon_1}$ of Λ^{M_1}) *w.r.t.* $\|\cdot\|_{1,M_2}$ (resp. $\|\cdot\|_{1,M_1}$) such that $|\Lambda^{M_2}_{\epsilon_2}| \leq (3/\epsilon_2)^{M_2}$ (resp. $|\Lambda^{M_1}_{\epsilon_1}| \leq (3/\epsilon_1)^{M_1}$). In particular, for all $\beta \in \Lambda^{M_2}$ (resp. $\lambda \in \Lambda^{M_1}$) there exists $\beta^{\epsilon_2} \in \Lambda^{M_2}_{\epsilon_2}$ (resp. $\lambda^{\epsilon_1} \in \Lambda^{M_1}_{\epsilon_1}$) such that $\|\beta - \beta^{\epsilon_2}\|_{1,M_2} \leq \epsilon_2$ (resp. $\|\lambda - \lambda^{\epsilon_1}\|_{1,M_1} \leq \epsilon_1$). From triangle inequality, one gets

$$\begin{aligned}|R_N(\hat{\sigma}^2_{\lambda,\beta}) - R(\hat{\sigma}^2_{\lambda,\beta})| \leq |R_N(\hat{\sigma}^2_{\lambda,\beta}) - R_N(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}})| + |R_N(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}) - R_N(\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}})| + |R_N(\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}) - R(\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}})| \\ + |R(\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}) - R(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}})| + |R(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}) - R(\hat{\sigma}^2_{\lambda,\beta})|.\end{aligned}$$

1. **Control of** $|R(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}) - R(\hat{\sigma}^2_{\lambda,\beta})|$**.** By Jensen's inequality, under assumptions 1- 2- 5 and $\mathbb{E}[\xi^2] = 1$ we obtain

$$\begin{aligned}|R(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}) - R(\hat{\sigma}^2_{\lambda,\beta})| &\leq \mathbb{E}\left[\left||Z - \hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}(X)|^2 - |Z - \hat{\sigma}^2_{\lambda,\beta}(X)|^2\right|\right] \\ &= \mathbb{E}\left[\left|\left(\sum_{j=1}^{M_2}\left(\beta_j - \beta_j^{\epsilon_2}\right)\hat{\sigma}^2_{\lambda,j}(X)\right)\left(2Z - \hat{\sigma}^2_{\lambda,\beta}(X) - \hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}(X)\right)\right|\right] \\ &\leq C_1\epsilon_2\ ,\end{aligned}$$

   where C_1 is a constant that depends on the upper bounds of σ^2 and $\hat{\sigma}^2_{\lambda,j}$.

2. **Control of** $|R_N(\hat{\sigma}^2_{\lambda,\beta}) - R_N(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}})|$**.** Since Assumptions 1, 2, and 3 are satisfied, we obtain

$$\begin{aligned}|R_N(\hat{\sigma}^2_{\lambda,\beta}) - R_N(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}})| &\leq \frac{1}{N}\sum_{i=n+1}^{n+N}\left(\sum_{j=1}^{M_2}|\beta_j - \beta_j^{\epsilon_2}||\hat{\sigma}^2_{\lambda,j}(X_i)|\right)|2\sigma^2(X_i)\xi_i^2 - \hat{\sigma}^2_{\lambda,\beta}(X_i) - \hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}(X_i)| \\ &\leq k\epsilon_2\left(\frac{C_2}{N}\sum_{i=n+1}^{n+N}\xi_i^2 + C_3\right),\end{aligned}$$

   where k is the bound of $\hat{\sigma}^2_{\lambda,j}$, C_2 is the constant which depends on σ^2 and C_3 is the constant which depends on the upper bounds $\hat{\sigma}^2_{\lambda,\beta}$ and $\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}$.

3. **Control of** $|R(\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}) - R(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}})|$**.** Under Assumptions 1, 2, 5, and 6, we get

$$\begin{aligned}|R(\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}) - R(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}})| &\leq \mathbb{E}\left[\sum_{j=1}^{M_2}\beta_j^{\epsilon_2}|\hat{\sigma}^2_{\lambda,j}(X) - \hat{\sigma}^2_{\lambda^{\epsilon_1},j}(X)||2\sigma^2(X)\xi^2 - \hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}(X) - \hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}(X)|\right] \\ &\leq \sum_{j=1}^{M_2}\beta_j^{\epsilon_2}\Big(2\mathbb{E}\left[\mathbb{E}\left[|\hat{\sigma}^2_{\lambda,j}(X) - \hat{\sigma}^2_{\lambda^{\epsilon_1},j}(X)|\sigma^2(X)\xi^2|\mathcal{D}_n, X\right]\right] \\ &+ \mathbb{E}\left[|\hat{\sigma}^2_{\lambda,j}(X) - \hat{\sigma}^2_{\lambda^{\epsilon_1},j}(X)||\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}(X) - \hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}(X)|\right]\Big) \\ &\leq \sum_{j=1}^{M_2}\beta_j^{\epsilon_2}\Big(2\mathbb{E}\left[|\hat{\sigma}^2_{\lambda,j}(X) - \hat{\sigma}^2_{\lambda^{\epsilon_1},j}(X)|\sigma^2(X)\mathbb{E}\left[\xi^2\right]\right] \\ &+ \mathbb{E}\left[|\hat{\sigma}^2_{\lambda,j}(X) - \hat{\sigma}^2_{\lambda^{\epsilon_1},j}(X)||\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}(X) - \hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}(X)|\right]\Big) \\ &\leq C_4\epsilon_1,\end{aligned}$$

where C_4 is constant which depends on K and the upper bounds of σ^2, $\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}$ and $\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}$.

4. **Control of** $|R_N(\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}) - R_N(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}})|$**.** We use the same way as 3. and we obtain

$$|R_N(\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}) - R_N(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}})| \leq \epsilon_1 \left(\frac{C_2}{N} \sum_{i=n+1}^{n+N} \xi_i^2 + C_5 \right),$$

where C_5 is constant which depends on the upper bounds of $\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}$ and $\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}$.

Therefore, we deduce that

$$\mathbb{E}\left[\sup_{(\lambda,\beta)\in\Lambda^{M_1}\times\Lambda^{M_2}} |R_N(\hat{\sigma}^2_{\lambda,\beta}) - R(\hat{\sigma}^2_{\lambda,\beta})| \right] \leq C_{k,C_2,C_3,C_4,C_5}(\epsilon_1+\epsilon_2) + \mathbb{E}\left[\sup_{(\lambda,\beta)\in\Lambda^{M_1}_{\epsilon_1}\times\Lambda^{M_2}_{\epsilon_2}} |R_N(\hat{\sigma}^2_{\lambda,\beta}) - R(\hat{\sigma}^2_{\lambda,\beta})| \right].$$

For some $(\lambda,\beta) \in \Lambda^{M_1}_{\epsilon_1} \times \Lambda^{M_2}_{\epsilon_2}$, set $T_i(\lambda,\beta) = |Z_i - \hat{\sigma}^2_{\lambda,\beta}(X_i)|^2 = |\sigma^2(X_i)\xi_i^2 - \hat{\sigma}^2_{\lambda,\beta}(X_i)|^2$ for all $i = n+1,\dots,n+N$. Let $L>0$. Since the variables $T_i(\lambda,\beta)$ are i.i.d. , we have

$$\begin{aligned}\mathbb{E}\left[\sup_{(\lambda,\beta)\in\Lambda^{M_1}_{\epsilon_1}\times\Lambda^{M_2}_{\epsilon_2}} |R_N(\hat{\sigma}^2_{\lambda,\beta}) - R(\hat{\sigma}^2_{\lambda,\beta})| \right] &\leq \mathbb{E}\left[\sup_{(\lambda,\beta)\in\Lambda^{M_1}_{\epsilon_1}\times\Lambda^{M_2}_{\epsilon_2}} \left| \frac{1}{N}\sum_{i=n+1}^{n+N} (T_i(\lambda,\beta) - \mathbb{E}[T_i(\lambda,\beta)])\mathbb{1}_{\{|\xi_i|\leq L\}} \right| \right] \\ &\quad + \mathbb{E}\left[\sup_{(\lambda,\beta)\in\Lambda^{M_1}_{\epsilon_1}\times\Lambda^{M_2}_{\epsilon_2}} \left| \frac{1}{N}\sum_{i=n+1}^{n+N} (T_i(\lambda,\beta) - \mathbb{E}[T_i(\lambda,\beta)])\mathbb{1}_{\{|\xi_i|> L\}} \right| \right]. \end{aligned} \tag{2.23}$$

Step 2.1. We control the first term on the r.h.s. of Eq. (2.23). On the event $\{|\xi| \leq L\}$ and under assumptions 1, 2 and 5, we get $|T_i(\lambda,\beta)| \leq c_1 L^4 + \bar{c}_1$ for all $i = n+1,\dots,n+N$ where c_1 is a positive constant which depends on the upper bound of σ^2 and $\bar{c}_1$ depends on the upper bound of $\hat{\sigma}^2_{\lambda,\beta}$. Conditionally on $\mathcal{D}_n$, we apply Hoeffding's inequality, for all $(\lambda,\beta) \in \Lambda^{M_1}_{\epsilon_1} \times \Lambda^{M_2}_{\epsilon_2}$, and all $t \geq 0$

$$\mathbb{P}\left(\left| \frac{1}{N}\sum_{i=n+1}^{n+N} (T_i(\lambda,\beta) - \mathbb{E}[T_i(\lambda,\beta)])\mathbb{1}_{\{|\xi_i|\leq L\}} \right| \geq t \right) \leq 2\exp\left(-\frac{-Nt^2}{2(c_1L^4+\bar{c}_1)^2} \right),$$

By a union bound on $(\lambda,\beta) \in \Lambda^{M_1}_{\epsilon_1} \times \Lambda^{M_2}_{\epsilon_2}$ and choosing $\epsilon_1 = \epsilon_2 = \frac{3}{N}$, we deduce that for all $t \geq 0$

$$\begin{aligned}&\mathbb{P}\left(\sup_{(\lambda,\beta)\in\Lambda^{M_1}_{\epsilon_1}\times\Lambda^{M_2}_{\epsilon_2}} \left| \frac{1}{N}\sum_{i=n+1}^{n+N} (T_i(\lambda,\beta) - \mathbb{E}[T_i(\lambda,\beta)])\mathbb{1}_{\{|\xi_i|\leq L\}} \right| \geq t \right) \\ &\qquad \leq 2\exp\left((M_1+M_2)\log(N) - \frac{-Nt^2}{2(c_1L^4+\bar{c}_1)^2} \right).\end{aligned}$$

We apply Lemma 6. Then, there exists a positive constant $\mathbf{c}$ such that

$$\mathbb{E}\left[\sup_{(\lambda,\beta)\in\Lambda^{M_1}_{\epsilon_1}\times\Lambda^{M_2}_{\epsilon_2}} \left| \frac{1}{N}\sum_{i=n+1}^{n+N} (T_i(\lambda,\beta) - \mathbb{E}[T_i(\lambda,\beta)])\mathbb{1}_{\{|\xi_i|\leq L\}} \right| \right] \leq \mathbf{c}(c_2L^4+\bar{c}_2)\left(\frac{(M_1+M_2)\log(N)}{N} \right)^{1/2},$$

where c_2 is constant which depends on c_1 and $\bar{c}_2$ on $\bar{c}_1$.

Step 2.2. We control the second term on the r.h.s. of Eq. (2.23). Thanks to the boundness of σ^2 and $\hat{\sigma}^2_{\lambda,\beta}$ and $\mathbb{E}[\xi^4] = 3$, we get $\mathbb{E}[T_i(\lambda,\beta)] \leq c_3$ and $T_i(\lambda,\beta) \leq c_4\xi_i^4 + c_5$ for all $i = n+1,\dots,n+N$ where c_4 and c_5 are constants which depend on the upper bounds of σ^2 and $\hat{\sigma}^2_{\lambda,\beta}$, respectively. By

Cauchy–Schwarz inequality and Lemma 3 we obtain

$$\mathbb{E}\left[\sup_{(\lambda,\beta)\in\Lambda_{\epsilon_1}^{M_1}\times\Lambda_{\epsilon_2}^{M_2}}\left|\frac{1}{N}\sum_{i=n+1}^{n+N}(T_i(s,m)-\mathbb{E}[T_i(s,m)])\mathbb{1}_{\{|\xi_i|>L\}}\right|\right]$$
$$\leq \frac{c_4}{N}\sum_{i=n+1}^{n+N}\mathbb{E}[\xi_i^4\mathbb{1}_{\{|\xi_i|>L\}}]+(c_3+c_5)\mathbb{P}(|\xi_1|>L)$$
$$\leq \bar{c}_4\sqrt{\mathbb{P}(|\xi_1|>L)}+(c_3+c_5)\mathbb{P}(|\xi_1|>L)$$
$$\leq \frac{\bar{c}_4\exp(-L^2/4)}{\sqrt{L}}+\frac{(c_3+c_5)\exp(-L^2/2)}{L}\ ,$$

where $\bar{c}_4$ is a positive constant that depends on c_4 and ξ.
Merging the results of the **Step 2.1** and **Step 2.2** in Eq.(2.23), and we obtain

$$\mathbb{E}\left[\sup_{(\lambda,\beta)\in\Lambda_{\epsilon_1}^{M_1}\times\Lambda_{\epsilon_2}^{M_2}}|R_N(\hat{\sigma}^2_{\lambda,\beta})-R(\hat{\sigma}^2_{\lambda,\beta})|\right]\leq \mathbf{c}(c_2L^4+\bar{c}_2)\left(\frac{(M_1+M_2)\log(N)}{N}\right)^{1/2}+\frac{\bar{c}_4\exp(-L^2/4)}{\sqrt{L}}$$
$$+\frac{(c_3+c_5)\exp(-L^2/2)}{L}.$$

Putting $L=\sqrt{2\log(N)}$, and we get

$$\mathbb{E}\left[\sup_{(\lambda,\beta)\in\Lambda_{\epsilon_1}^{M_1}\times\Lambda_{\epsilon_2}^{M_2}}|R_N(\hat{\sigma}^2_{\lambda,\beta})-R(\hat{\sigma}^2_{\lambda,\beta})|\right]\leq c_6\left(\frac{(M_1+M_2)\log^5(N)}{N}\right)^{1/2}\ ,$$

where c_6 is constant which depends on c_2. Thus,

$$\mathbb{E}\left[R(\tilde{\sigma}^2_{\mathrm{C}})-R(\bar{\sigma}^2_{\mathrm{C}})\right]\leq C\left(\frac{(M_1+M_2)\log^5(N)}{N}\right)^{1/2}\ ,$$

where C is constant which depends on c_6 and $\mathbf{c}$.

Remark 4. *When Y is bounded, it is clear that there exists an absolute constant $C>0$*

$$\mathbb{E}\left[R(\tilde{\sigma}^2_{C})-R(\bar{\sigma}^2_{C})\right]\leq C\left(\frac{(M_1+M_2)\log(N)}{N}\right)^{1/2}\ .$$

Step 3. Study of the term $\mathbb{E}\left[R(\hat{\sigma}^2_{\mathrm{C}})-R(\tilde{\sigma}^2_{\mathrm{C}})\right]$. We use the same arguments of proof of Theorem 9 (**Step 2.2**), and we get that there exists two positive constants C_1 and C_2 such that

$$\mathbb{E}\left[R(\hat{\sigma}^2_{\mathrm{C}})-R(\tilde{\sigma}^2_{\mathrm{C}})\right]\leq C_1\left\{\mathbb{E}\left[\|\hat{f}_{\mathrm{C}}-f^*\|_N^2\right]\right\}^{1/p}+C_2\alpha_N\ , \tag{2.24}$$

where $p=2$ if Y is bounded, $p=4$ otherwise, and

$$\alpha_N=\begin{cases}\left(\frac{(M_1+M_2)\log(N)}{N}\right)^{1/2} & \text{if } Y \text{ is bounded;}\\ \left(\frac{(M_1+M_2)\log^5(N)}{N}\right)^{1/2} & \text{otherwise.}\end{cases}$$

In the sequel, we give the following proposition

Proposition 3. *Let $\hat{f}_C$ be the aggregate defined in Eq. (2.4). Then, under Assumptions 2 and 5 there exists an absolute constant $C>0$*

$$\mathbb{E}\left[\|\hat{f}_C-f^*\|_N^2\right]\leq\min_{\lambda\in\Lambda^{M_1}}\mathbb{E}\left[\|\hat{f}_\lambda-f^*\|_N^2\right]+C\sqrt{\frac{\log(M_1)}{N}}\ .$$

The proof of this proposition is similar of the proof of Proposition 2. Thus, we apply Proposition 3 in inequality (2.24) and we get

$$\mathbb{E}\left[R(\hat{\sigma}^2_{\mathrm{C}}) - R(\tilde{\sigma}^2_{\mathrm{C}})\right] \leq C_1 \left\{ \min_{\lambda \in \Lambda^{M_2}} \mathbb{E}\left[\|\hat{f}_\lambda - f^*\|^2_N\right] \right\}^{1/p} + \bar{C}_1 \phi^{\mathrm{C}}_N(M_1) \ ,$$

where $\bar{C}_1$ is a constant that depends on C_1 and the constant in Proposition 3, where $p = 2$ if Y is bounded, $p = 4$ otherwise, and

$$\phi^{\mathrm{C}}_N(M_1) = \begin{cases} \left(\frac{\log(M_1)}{N}\right)^{1/4} & \text{if } Y \text{ is bounded;} \\ \left(\frac{\log(M_1)}{N}\right)^{1/8} & \text{otherwise.} \end{cases}$$

Combining **Step** 1, **Step** 2 and **Step** 3 in Eq (2.22) yields the result.

2.6.4 Technical lemmas

In this section, we gather several technical results which are used to derive the proof of results of this chapter.

Lemma 3. *Let X be the standard gaussian distribution, then for any $x > 0$, it holds*

$$\mathbb{P}(X > x) \leq \frac{\exp(-x^2/2)}{\sqrt{2\pi}x} \ , \ \textit{and} \ \ \mathbb{P}(|X| > x) \leq \sqrt{\frac{2}{\pi}} \frac{\exp(-x^2/2)}{x} \ .$$

Proof. Since $X \sim \mathcal{N}(0,1)$, one gets

$$\mathbb{P}\left(X > x\right) = \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} \exp(-u^2/2) du \leq \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} \frac{u}{x} \exp(-u^2/2) du = \frac{\exp(-x^2/2)}{\sqrt{2\pi}x}.$$

The second inequality follows from symmetry and the last one using the union bound

$$\mathbb{P}(|X| > x) \leq 2\mathbb{P}\left(X > x\right).$$

□

Lemma 4. *Let $X \sim \mathcal{N}(0,1)$ and $k \geq 1$, then*

$$\mathbb{E}\left[|X|^{2k}\right] \leq 2^{k+1} k!.$$

Proof.

$$\begin{aligned} \mathbb{E}\left[|X|^{2k}\right] = \int_0^{+\infty} \mathbb{P}\left(|X|^{2k} > t\right) dt \quad &= \quad \int_0^{+\infty} \mathbb{P}\left(|X| > t^{\frac{1}{2k}}\right) dt \\ &\leq \quad 2\int_0^{+\infty} \exp\left(-t^{\frac{1}{k}}/2\right) dt \\ &\overset{u=t^{\frac{1}{k}}/2}{=} \quad 2^{k+1} k \int_0^{+\infty} u^{k-1} \exp(-u) du = 2^{k+1} k!. \end{aligned}$$

□

Lemma 5. *Let $X_1, \ldots, X_M$ be zero mean ν-subgaussian random variables,* i.e., $\mathbb{E}\left[\exp(rX_i)\right] \leq exp\left(\frac{r^2\nu^2}{2}\right)$ *for all $r > 0$. Then*

$$\mathbb{E}\left[\max_{1 \leq i \leq M} X_i\right] \leq \nu\sqrt{2\log(M)}.$$

Proof. By Jensen's inequality, for any $r > 0$

$$\begin{aligned}\mathbb{E}\left[\max_{1\le i\le N} X_i\right] = \frac{1}{r}\mathbb{E}\left[\log\left(\exp\left(r\max_{1\le i\le M} X_i\right)\right)\right] &\le \frac{1}{r}\log\left(\mathbb{E}\left[\exp\left(r\max_{1\le i\le M} X_i\right)\right]\right)\\ &= \frac{1}{r}\log\left(\mathbb{E}\left[\max_{1\le i\le M}\exp\left(rX_i\right)\right]\right)\\ &\le \frac{1}{r}\log\left(\sum_{i=1}^{M}\mathbb{E}\left[\exp\left(rX_i\right)\right]\right)\\ &\le \frac{1}{r}\log\left(\sum_{i=1}^{M}\mathbb{E}\left[\exp\left(\frac{r^2\nu^2}{2}\right)\right]\right) = \frac{\log(M)}{r} + \frac{\nu^2 r}{2}\ ,\end{aligned}$$

taking $r = \sqrt{\frac{2\log(M)}{\nu^2}}$ and we get the result. □

Lemma 6. *Let $N \in \mathbb{N}^*$, $a \ge 1$, b and c be two non negative real numbers. Consider Z a positive random variable such that*

$$\mathbb{P}\left(Z \ge t\right) \le \min(1, \exp(a - bNt^2)\ . \tag{2.25}$$

Then, there exists a constant $C > 0$ not depending of N such that

$$\mathbb{E}[Z] \le C\left(\frac{a}{bN}\right)^{1/2}.$$

Proof. By condition (2.25), we have

$$\mathbb{E}[Z] \le \int_0^{+\infty}\min(1, \exp(a - bNt^2)dt \le \left(\frac{a}{bN}\right)^{1/2} + \int_{\left(\frac{a}{bN}\right)^{1/2}}^{+\infty}\exp(a - bNt^2)dt. \tag{2.26}$$

The following elementary inequality $(x-y)^2 \le x^2 - y^2$ for all $x, y \ge 0$ yields to

$$\begin{aligned}\int_{\left(\frac{a}{bN}\right)^{1/2}}^{+\infty}\exp(a - bNt^2)dt &\le \int_{\left(\frac{a}{bN}\right)^{1/2}}^{+\infty}\exp\left(-bN\left(t - \left(\frac{a}{bN}\right)^{1/2}\right)^2\right)dt\\ &= \int_0^{+\infty}\exp\left(-bNu^2\right)du \le C\left(\frac{1}{bN}\right)^{1/2}.\end{aligned} \tag{2.27}$$

Combining Equation (2.27) in Equation (2.26) to yield the result. □

Lemma 7 (Bernstein's inequality)**.** *Let $T_1, \ldots, T_n$ be independent real valued random variables. Assume that there exists some positive numbers v and c such that*

$$\sum_{i=1}^{n}\mathbb{E}[T_i^2] \le v\ ,$$

and for all integers $k \ge 3$

$$\sum_{i=1}^{n}\mathbb{E}[(T_i \vee 0)^k] \le \frac{k!}{2}vc^{k-2}\ .$$

Let $S = \sum_{i=1}^{n}\left(T_i - \mathbb{E}[T_i]\right)$, then for every any positive x we have

$$\mathbb{P}\left(|S| \ge x\right) \le 2\exp\left(-\frac{x^2}{2(v + cx)}\right)\ .$$

Lemma 8 (Hoeffding's inequality)**.** *Let $N \in \mathbb{N}^*$ and $a > 0$ be a real number. Let $X_1, \ldots, X_N$ be independent random variables having values in $[-a, a]$, then for all $t > 0$*

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}(X_i - \mathbb{E}[X_i])\right| > t\right) \le 2\exp\left(-\frac{Nt^2}{2a^2}\right).$$

Lemma 9 (Hoeffding's Lemma)**.** *Let $X \in [a, b]$ be a bounded random variable with $\mathbb{E}[X] = 0$. Then, for all $\lambda \in \mathbb{R}$*

$$\mathbb{E}\left[\exp(\lambda X)\right] \leq \exp\left(\frac{\lambda^2 (b-a)^2}{8}\right).$$

Chapter 3

Regression with reject option and application to kNN

Abstract: We investigate the problem of regression where one is allowed to abstain from predicting. We refer to this framework as *regression with reject option* as an extension of classification with reject option. In this context, we focus on the case where the rejection rate is fixed and derive the optimal rule which relies on thresholding the conditional variance function. We provide a semi-supervised estimation procedure of the optimal rule involving two datasets: a first *labeled* dataset is used to estimate both regression function and conditional variance function while a second *unlabeled* dataset is exploited to calibrate the desired rejection rate. The resulting predictor with reject option is shown to be almost as good as the optimal predictor with reject option both in terms of risk and rejection rate. We additionally apply our methodology with kNN algorithm and establish rates of convergence for the resulting kNN predictor under mild conditions. Finally, a numerical study is performed to illustrate the benefit of using the proposed procedure.

Contents

3.1 Introduction

Confident prediction is a fundamental problem in statistical learning for which numerous efficient algorithms have been designed, *e.g.*, neural-networks, kernel methods, or k-Nearest-Neighbors (kNN) to name a few. However, even state-of-art methods may fail in some situations, leading to bad decision-making. Obvious damageable incidences of an erroneous decision may occur in several fields such as medical diagnosis, where a wrong estimation can be fatal. In this work, we provide a novel statistical procedure designed to handle these cases. In the specific context of regression, we build a prediction algorithm that allows to abstain from predicting when the doubt is too important. As a generalization of the classification with reject option setting [18, 19, 23, 42, 49, 64, 94], this framework is naturally referred to as *regression with reject option*. In the spirit of [23], we opt here for a strategy where the predictor can abstain up to a fraction $\varepsilon \in (0, 1)$ of the data. The merit of this approach is that it allows human action on the proportion of the data where the prediction is too difficult while standard machine learning algorithms can be exploited to perform the predictions on the other fraction of the data. The difficulty to address a prediction is then automatically evaluated by the procedure. From this perspective, this strategy may improve the efficiency of the human intervention.

In this chapter, we investigate the regression problem with reject option when the rejection (or abstention) rate is controlled. Specifically, we provide a statistically principled and computationally efficient algorithm tailored to this problem. We first formally define the regression with reject option framework, and explicitly exhibit the optimal predictor with bounded rejection rate in Section 3.2. This optimal rule relies on a thresholding of the conditional variance function. This result is the bedrock of our work and suggests the use of a plug-in approach. We propose in Section 3.3 a two-step procedure which first estimates both the regression function and the conditional variance function on a first *labeled* dataset and then calibrates the threshold responsible for abstention using a second *unlabeled* dataset. Under mild assumptions, we show that our procedure performs as well as the optimal predictor both in terms of risk and rejection rate. We emphasize that our procedure can be exploited with any off-the-shell estimator. As an example we apply in Section 3.4 our methodology with the kNN algorithm for which we derive rates of convergence. Finally, we perform numerical experiments in Section 3.5 which illustrate the benefits of our approach. In particular, it highlights the flexibility of the proposed procedure.

Rejection in regression is extremely rarely considered in the literature, an exception being [100] that views the reject option from a different perspective. There, the authors used the reject option from the side of ε-optimality, and therefore ensures that the prediction is inside a ball with radius ε around the regression function with high probability. Their methodology is intrinsically associated with empirical risk minimization procedures. In contrast, our method is applicable to any estimation procedure. Closer related works to ours appears in classification with reject option literature [5, 18, 19, 23, 42, 49, 64, 94]. In particular, the present work can be viewed as an extension of the classification with reject option setting. Indeed, from a general perspective, the present contribution brings a deeper understanding of the reject option. Importantly, the conditional variance function appears to capture the main feature behind the abstention decision. In [23], the authors also provide rates of convergence for plug-in type approaches in the case of bounded rejection rate. However, their rates of convergence holds only under some margin type assumption [2, 71] and a smoothness assumption on the considered estimator. On the contrary, we do not require these assumptions to get valid rates of convergence.

3.2 Regression with reject option

In this section we introduce the regression with reject option setup and derive a general form of the optimal rule in this context. We additionally highlight the case of fixed rejection rate as our main framework. First of all, before we proceed, let us introduce some preliminary notation. Let (X, Y) be a random couple taking its values in $\mathbb{R}^d \times \mathbb{R}$: here X denotes a feature vector and Y is the corresponding output. We denote by $\mathbb{P}$ the joint distribution of (X, Y) and by $\mathbb{P}_X$ the marginal

distribution of the feature X. Let $x \in \mathbb{R}^d$, we introduce the regression function $f^*(x) = \mathbb{E}[Y|X = x]$ as well as the conditional variance function $\sigma^2(x) = \mathbb{E}\left[(Y - f^*(X))^2|X = x\right]$. We will give due attention to these two functions in our analysis. In addition, we denote by $\|\cdot\|$ the Euclidean on $\mathbb{R}^d$. Finally, $|\cdot|$ stands for the cardinality when dealing with a finite set.

3.2.1 Predictor with reject option

Let f be some measurable real-valued function which must be viewed as a prediction function. A *predictor with reject option* Γ_f associated to f is defined as being any function that maps $\mathbb{R}^d$ onto $\mathcal{P}(\mathbb{R})$ such for all $x \in \mathbb{R}^d$, the output $\Gamma_f(x) \in \{\emptyset, \{f(x)\}\}$. We denote by Υ_f the set of all predictors with reject option that relies on f. Hence, in this framework, there are only two options for a particular $x \in \mathbb{R}^d$: whether the predictor with reject option outputs the empty set, meaning that no prediction is produced for x; or the output $\Gamma_f(x)$ is of size 1 and the prediction coincides with the value $f(x)$. The framework of regression with reject option naturally brings into play two important characteristics of a given predictor Γ_f. The first one is the rejection rate that we denote by $r(\Gamma_f) = \mathbb{P}(|\Gamma_f(X)| = 0)$ and the second one is the L_2 error when prediction is performed

$$\mathrm{Err}(\Gamma_f) = \mathbb{E}\left[(Y - f(X))^2 \mid |\Gamma_f(X)| = 1\right] \quad .$$

The ultimate goal in regression with reject option is to build a predictor Γ_f with a small rejection rate that achieves a small conditional L_2 error as well. A natural way to make this happen is to embed these quantities into a measure of performance. To this end, let consider the following risk

$$\mathcal{R}_\lambda(\Gamma_f) = \mathbb{E}\left[(Y - f(X))^2 \mathbb{1}_{\{|\Gamma_f(X)|=1\}}\right] + \lambda\, r(\Gamma_f) \quad ,$$

where $\lambda \geq 0$ is a tuning parameter which is responsible for compromising error and rejection rate: larger λ's result in predictors Γ_f with smaller rejection rates, but with larger errors. Hence, λ can be interpreted as the price to pay for using the reject option. Note that the above risk $\mathcal{R}_\lambda$ has already been considered by [42] in the classification framework.

Minimizing the risk $\mathcal{R}_\lambda$, we derive an explicit expression of an optimal predictor with reject option.

Proposition 4. *Let $\lambda \geq 0$, and consider*

$$\Gamma^*_\lambda \in \arg\min \mathcal{R}_\lambda(\Gamma_f) \quad ,$$

where the minimum is taken over all predictors with rejection option $\Gamma_f \in \Upsilon_f$ and all measurable functions f. Then we have that

1. *The optimal predictor with rejected option Γ^*_λ can be written as*

$$\Gamma^*_\lambda(X) = \begin{cases} \{f^*(X)\} & \text{if } \sigma^2(X) \leq \lambda \\ \emptyset & \text{otherwise} \quad . \end{cases} \tag{3.1}$$

2. *For any $\lambda < \lambda'$, the following holds*

$$\mathrm{Err}(\Gamma^*_\lambda) \leq \mathrm{Err}(\Gamma^*_{\lambda'}) \quad \text{and} \quad r(\Gamma^*_\lambda) \geq r(\Gamma^*_{\lambda'}) \quad .$$

Interestingly, this result shows that the oracle predictor relies on thresholding the conditional variance function σ^2. We believe that this is an important remark that provides an essential characteristic of the reject option in regression but also in classification. Indeed, it has been shown that the optimal classifier with reject option for classification is obtained by thresholding the function f^* (see for instance [42]). However, in the binary case where $Y \in \{0, 1\}$, one has $\sigma^2(x) = f^*(x)(1 - f^*(x))$, and then thresholding σ^2 and f^* are equivalent.

The second point of the proposition shows that the error and the rejection rate of the optimal predictor are working in two opposite directions *w.r.t.* λ and then a compromise is required. We illustrate this aspect with the `airfoil` dataset, and the kNN predictor (see Section 3.5) in the contiguous Figure 3.1. The two curves correspond to the evaluation of the error $\mathrm{Err}(\hat{\Gamma}_\lambda)$ (blue-solid line) and the rejection rate $r(\hat{\Gamma}_\lambda)$ (red-dashed line) as a function of λ. In general any choice of the parameter λ is difficult to interpret. Indeed, one of the major drawbacks of this approach is that any fixed λ (or even an "optimal" value of this parameter) does not allow to control neither of the two parts of the risk function. Especially, the rejection rate can be arbitrary large.

For this reason, we investigate in Section 3.2.2 the setting where the rejection rate is fixed. We understand this rejection rate as a budget one has beforehand.

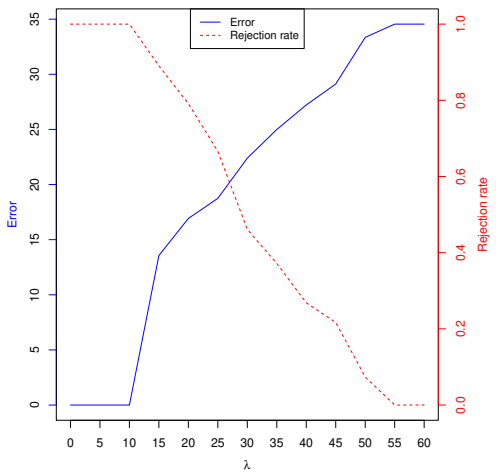


Figure 3.1 – $\widehat{\mathrm{Err}}\left(\hat{\Gamma}_\lambda\right)$ and $\hat{r}\left(\hat{\Gamma}_\lambda\right)$ vs. λ.

3.2.2 Optimal predictor with fixed rejection rate

In this section, we introduce the framework where the rejection rate is fixed or at least bounded. That is to say, for a given predictor with reject option Γ_f and a given rejection rate $\varepsilon \in (0,1)$, we ask that Γ_f satisfies following constraint $r(\Gamma_f) \leq \varepsilon$. This kind of constraint has also been considered by [23] in the classification setting. Our objective becomes to solve the constrained problem[1]:

$$\Gamma_\varepsilon^* \in \arg\min\{\mathrm{Err}(\Gamma_f) \ : \ r(\Gamma_f) \leq \varepsilon\} \ . \tag{3.2}$$

In the same vein as Proposition 4, we aim at writing an explicit expression of Γ_ε^*, referred in what follows to as ε-predictor. However, this expression is not well identified in the general case. Therefore, we make the following mild assumption on the distribution of $\sigma^2(X)$, which translates the fact that the function σ^2 is not constant on any set with non-zero measure *w.r.t.* $\mathbb{P}_X$.

Assumption 7. *The cumulative distribution function F_{σ^2} of $\sigma^2(X)$ is continuous.*

Let us denote by $F_{\sigma^2}^{-1}$ the generalized inverse of the cumulative distribution F_{σ^2} defined for all $u \in (0,1)$ as $F_{\sigma^2}^{-1}(u) = \inf\{t \in \mathbb{R} \ : \ F_{\sigma^2}(t) \geq u\}$. Under Assumption 7 and from Proposition 4, we derive an explicit expression of the ε-predictor Γ_ε^* given by (3.2).

Proposition 5. *Let $\varepsilon \in (0,1)$, and let $\lambda_\varepsilon = F_{\sigma^2}^{-1}(1-\varepsilon)$. Under Assumption 7, we have $\Gamma_\varepsilon^* = \Gamma_{\lambda_\varepsilon}^*$.*

As an immediate consequence of the above proposition and properties on quantile functions is that

$$r(\Gamma_\varepsilon^*) = \mathbb{P}(|\Gamma_\varepsilon^*(X)| = 0) = \mathbb{P}\left(\sigma^2(X) \geq \lambda_\varepsilon\right) = \mathbb{P}\left(F_{\sigma^2}(\sigma^2(X)) \geq 1-\varepsilon\right) = \varepsilon \ ,$$

and then the ε-predictor has rejection rate exactly ε. The continuity Assumption 7 is a sufficient condition to ensure that this property holds true. Besides, from this assumption, the ε-predictor can be expressed as follows

$$\Gamma_\varepsilon^*(x) = \begin{cases} \{f^*(x)\} & \text{if } \ F_{\sigma^2}(\sigma^2(x)) \leq 1-\varepsilon \\ \emptyset & \text{otherwise} \ . \end{cases} \tag{3.3}$$

Finally, as suggested by Proposition 4 and 5, the performance of a given predictor with reject option Γ_f is measured through the risk $\mathcal{R}_\lambda$ when $\lambda = \lambda_\varepsilon$. Then, its excess risk is given by

$$\mathcal{E}_{\lambda_\varepsilon}(\Gamma_f) = \mathcal{R}_{\lambda_\varepsilon}(\Gamma_f) - \mathcal{R}_{\lambda_\varepsilon}(\Gamma_\varepsilon^*) \ ,$$

for which the following result provides a closed formula.

[1] By abuse of notation, we refer to Γ_λ^* as the solution of the penalized problem and to Γ_ε^* as the solution of the constrained problem.

Proposition 6. *Let $\varepsilon \in (0,1)$. For any predictor Γ_f, we have*

$$\mathcal{E}_{\lambda_\varepsilon}(\Gamma_f) = \mathbb{E}_X\left[(f^*(X) - f(X))^2 \mathbb{1}_{\{|\Gamma_f(X)|=1\}}\right] + \mathbb{E}_X\left[|\sigma^2(X) - \lambda_\varepsilon| \mathbb{1}_{\{|\Gamma_f(X)| \neq |\Gamma_\varepsilon^*(X)|\}}\right] \ .$$

The above excess risk consists of two terms that translates two different aspect of the regression with reject option problem. The first one is related to the L_2 risk of the prediction function f and is rather classical in the regression setting. On contrast, the second is related to the reject option problem. It is dictated by the behavior of the conditional variance σ^2 around the threshold λ_ε.

3.3 Plug-in ε-predictor with reject option

We devote this section to the study of a data-driven predictor with reject option based on the *plug-in* principle that mimics the optimal rule derived in Proposition 5.

3.3.1 Estimation strategy

Equation (3.3) indicates that a possible way to estimate Γ_ε^* relies on the plug-in principle. To be more specific, Eq. (3.3) suggests that estimating f^* and σ^2, as well as the cumulative distribution F_{σ^2} would be enough to get an estimator of Γ_ε^*. To build such predictor, we first introduce a learning sample $\mathcal{D}_n = \{(X_i, Y_i), \ \ i = 1, \ldots, n\}$ which consists of n independent copies of (X, Y). This dataset helps us to construct estimators $\hat{f}$ and $\hat{\sigma}^2$ of the regression function f^* and the conditional variance function σ^2 respectively. In this chapter, we focus on estimator $\hat{\sigma}^2$ which relies on the residual-based methods [37]. Based on $\mathcal{D}_n$, the estimator $\hat{\sigma}^2$ is obtained by solving the regression problem of the output variable $(Y - \hat{f}(X))^2$ on the input variable X. Estimating the last quantity F_{σ^2} is rather simple by replacing cumulative distribution function by its empirical version. Since this term only depends on the marginal distribution $\mathbb{P}_X$, we estimate it using a second *unlabeled* dataset $\mathcal{D}_N = \{X_{n+1}, \ldots, X_{n+N}\}$ composed of N independent copies of X. This is an important feature of our methodology since unlabeled data are usually easy to get. The dataset $\mathcal{D}_N$ is assumed to be independent of $\mathcal{D}_n$. We set

$$\hat{F}_{\hat{\sigma}^2}(\cdot) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{\hat{\sigma}^2(X_{n+i}) \leq \cdot\}} \ ,$$

as an estimator for F_{σ^2}. With this notation, the *plug-in ε-predictor* is the predictor with reject option defined for each $x \in \mathbb{R}^d$ as

$$\hat{\Gamma}_\varepsilon(x) = \begin{cases} \left\{\hat{f}(x)\right\} & \text{if } \ \hat{F}_{\hat{\sigma}^2}(\hat{\sigma}^2(x)) \leq 1 - \varepsilon \\ \emptyset & \text{otherwise} \ . \end{cases} \tag{3.4}$$

It is worth noting that the proposed methodology is flexible enough to rely upon any off-the-shelf estimators of the regression function f^* and the conditional variance function σ^2.

3.3.2 Consistency of plug-in ε-predictors

In this part, we investigate the statistical properties of the plug-in ε-predictors with reject option. This analysis requires an additional assumption on the following quantity

$$F_{\hat{\sigma}^2}(\cdot) = \mathbb{P}_X\left(\hat{\sigma}^2(X) \leq \cdot | \mathcal{D}_n\right) \ .$$

Assumption 8. *The cumulative distribution function $F_{\hat{\sigma}^2}$ of $\hat{\sigma}^2(X)$ is continuous.*

This condition is analogous to Assumption 7 but deals with the estimator $\hat{\sigma}^2(X)$ instead of the true conditional variance $\sigma^2(X)$. This difference makes Assumption 8 rather weak as the estimator $\hat{\sigma}^2(X)$ is chosen by the practitioner. Moreover, we can make any estimator satisfy this condition by providing a smoothed version of it. We illustrate this strategy with kNN algorithm in Section 3.4. Next theorem is the main result of this section, it establishes the consistency of the predictor $\hat{\Gamma}_\varepsilon$ to the optimal one.

Theorem 11. *Let $\varepsilon \in (0,1)$. Assume that σ^2 is bounded, $\hat{f}$ is a consistent estimator of f^* w.r.t. the L_2 risk, and $\hat{\sigma}^2$ is a consistent estimator of σ^2 w.r.t. the L_1 risk. Under Assumptions 7- 8, the followings hold*

$$\mathbb{E}\left[\mathcal{E}_{\lambda_\varepsilon}\left(\hat{\Gamma}_\varepsilon\right)\right] \underset{n,N\to+\infty}{\longrightarrow} 0, \quad \textit{and} \quad \mathbb{E}\left[|r(\hat{\Gamma}_\varepsilon) - \varepsilon|\right] \leq CN^{-1/2} \ ,$$

where $C > 0$ is an absolute constant.

This theorem establishes the fact that the plug-in ε-predictor behaves asymptotically as well as the optimal ε-predictor both in terms of risk and rejection rate. The convergence of the rejection rate requires only Assumption 8 which is rather weak and can even be removed following the process detailed in Section 3.4.2. In particular, the theorem shows that the rejection rate of the plug-in ε-predictor is of level ε up to a term of order $O(N^{-1/2})$. This rate is similar to the one obtained in the classification setting [23]. It relies on the difference between the cumulative distribution $F_{\hat{\sigma}^2}$ and its empirical counterpart $\hat{F}_{\hat{\sigma}^2}$ that is controlled using Dvoretzky-Kiefer-Wolfowitz Inequality [61]. Interestingly, this result applies to any consistent estimators of f^* and σ^2.

The estimation of regression function f^* is widely studied and suitable algorithm such as random forests, kernel procedures, or kNN estimators can be used, see [9, 36, 78, 82, 86]. The estimation of the conditional variance function which relies on the residual-based methods has also been extensively studied based on kernel procedures, see for instance [30, 37, 38, 46, 81]. In the next section, we derive rates of convergence in the case where both estimators $\hat{f}$ and $\hat{\sigma}^2$ rely on the kNN algorithm. In particular, we establish rates of convergence for $\hat{\sigma}^2$ in sup norm (see Proposition 9 in the supplementary material).

3.4 Application to kNN algorithm: rates of convergence

The plug-in ε-predictor $\hat{\Gamma}_\varepsilon$ relies on estimators of the regression and the conditional variance functions. In this section, we consider the specific case of kNN based estimations. We refer to the resulting predictor as *kNN predictor with reject option*. Specifically, we establish rates of convergence for this procedure. In addition, since kNN estimator of σ^2 violates Assumption 8, applying our methodology to kNN has the benefit of illustrating the smoothing technique to make this condition be satisfied.

3.4.1 Assumptions

To study the performance of the kNN predictor with reject option in the finite sample regime, we assume that X belongs to a regular compact set $\mathcal{C} \subset \mathbb{R}^d$, see [2]. Besides, we make the following assumptions.

Assumption 9. *The functions f^* and σ^2 are Lipschitz.*

Assumption 10 (Strong density assumption)**.** *We assume that the marginal distribution $\mathbb{P}_X$ admits a density μ w.r.t to the Lebesgue measure such that for all $x \in \mathcal{C}$, we have $0 < \mu_{\min} \leq \mu(x) \leq \mu_{\max}$.*

These two assumptions are rather classical when we deal with rate of convergence and we refer the reader to the baseline books [36, 86]. In particular, we point out that the strong density assumption has been introduced in the context of binary classification for instance in [2]. The last assumption that we require highlights the behavior of σ^2 around the threshold λ_ε.

Assumption 11 (α-exponent assumption)**.** *We say that σ^2 has exponent $\alpha \geq 0$ (at level λ_ε) with respect to $\mathbb{P}_X$ if there exists $c^* > 0$ such that for all $t > 0$*

$$\mathbb{P}_X\left(0 < |\sigma^2(X) - \lambda_\varepsilon| \leq t\right) \leq c^* t^\alpha \ .$$

This assumption has been first introduced in [71] and is also referred as Margin assumption in the binary classification setting, see [60]. For $\alpha > 0$, Assumption 11 ensures that the random variable $\sigma^2(X)$ can not concentrate too much around the threshold λ_ε. It allows to derive faster rates of convergence. Note that, if $\alpha = 0$ there is no assumption.

3.4.2 kNN predictor with reject option

For any $x \in \mathbb{R}^d$, we denote by $(X_{(i,n)}(x), Y_{(i,n)}(x)), i = 1, \dots n$ the reordered data according to the ℓ_2 distance in $\mathbb{R}^d$, meaning that $\|X_{(i,n)}(x) - x\| < \|X_{(j,n)}(x) - x\|$ for all $i < j$ in $\{1, \dots, n\}$. Note that Assumption 10 ensures that ties occur with probability 0 (see [36] for more details). Let $k = k_n$ be an integer. The kNN estimator of f^* and σ^2 are then defined, for all x, as follows

$$\hat{f}(x) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{(i,n)}(x) \quad \text{and} \quad \hat{\sigma}^2(x) = \frac{1}{k_n} \sum_{i=1}^{k_n} \left(Y_{(i,n)}(x) - \hat{f}(X_{(i,n)}(x)) \right)^2 \ .$$

Conditional on $\mathcal{D}_n$, the cumulative distribution function $F_{\hat{\sigma}^2}$ is not continuous and then Assumption 8 does not hold. To avoid this issue, we introduce a random perturbation ζ distributed according to the Uniform distribution on $[0, u]$ that is independent from every other random variable where $u > 0$ is a (small) fixed real number that will be specified later. Then, we define the random variable $\bar{\sigma}^2(X, \zeta) := \hat{\sigma}^2(X) + \zeta$. It is not difficult to see that, conditional on $\mathcal{D}_n$ the cumulative distribution $F_{\bar{\sigma}^2}$ of $\bar{\sigma}^2(X, \zeta)$ is continuous. Furthermore, by the triangle inequality, the consistency of $\hat{\sigma}^2$ implies the consistency of $\bar{\sigma}^2$ provided that u tends to 0. Therefore, we naturally define the kNN predictor with reject option as follows.

Let $(\zeta_1, \dots, \zeta_N)$ be independent copies of ζ and independent of every other random variable. We set

$$\hat{F}_{\bar{\sigma}^2}(.) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{\hat{\sigma}^2(X_{n+i}) + \zeta_i \leq \cdot\}} \ ,$$

and the kNN ε-predictor with reject option is then defined for all x and ζ as

$$\hat{\Gamma}_\varepsilon(x, \zeta) = \begin{cases} \left\{ \hat{f}(x) \right\} & \text{if } \ \hat{F}_{\bar{\sigma}^2}(\bar{\sigma}^2(x, \zeta)) \leq 1 - \varepsilon \\ \emptyset & \text{otherwise} \ . \end{cases}$$

3.4.3 Rates of convergence

In this section, we derive the rates of convergence of the kNN ε-predictor in the following framework. We assume that Y is bounded or that Y satisfies

$$Y = f^*(X) + \sigma(X)\xi \ , \tag{3.5}$$

where ξ is independent of X and distributed according to a standard normal distribution. Note that these assumptions covers a broad class of applications. Under these assumptions, we can state the following result.

Theorem 12. *Grant Assumptions 7, 9, 10, and 11. Let $\varepsilon \in (0, 1)$, if $k_n \propto n^{2/(d+2)}$, and $u \leq n^{-1/(d+2)}$, then the kNN ε-predictor $\hat{\Gamma}_\varepsilon$ satisfies*

$$\mathbb{E}\left[\mathcal{E}_{\lambda_\varepsilon}\left(\hat{\Gamma}_\varepsilon \right) \right] \leq C \left(n^{-2/(d+2)} + \log(n)^{(\alpha+1)} n^{-(\alpha+1)/(d+2)} + N^{-1/2} \right) \ ,$$

where $C > 0$ is a constant which depends on f^*, σ^2, c_0, c^*, α, $\mathcal{C}$, and on the dimension d.

Each part of the above rate describes a given feature of the problem. The first one relies on the estimation error of the regression function. The second one, which depends in part on the parameter α from Assumption 11, is due to the estimation error in sup norm of the conditional variance $\mathbb{E}\left[\left(\sup_{x \in \mathcal{C}} \left| \hat{\sigma}^2(x) - \sigma^2(x) \right| \right) \right] \leq C \log(n) n^{-1/(d+2)}$ stated in Proposition 9 in the supplementary material. Notice that for $\alpha = 1$, the second term is of the same order (up to logarithmic factor) as the term corresponding to the estimation of the regression function. The last term is directly linked to the estimation of the threshold λ_ε. Lastly, for $\alpha > 1$, we observe, provided that the size of the unlabeled sample N is sufficiently large, that this rate is the same as the rate of $\hat{f}$ in L_2 norm which is then the best situation that we can expect for the rejection setting.

3.5 Numerical experiments

In this section, we present numerical experiments to illustrate the performance of the plug-in ε-predictor. The construction process of this predictor is described in Section 3.3.1 and relies on estimators of the regression and the conditional variance functions. The code used for the implementation of the plug-in ε-predictor can be found at `https://github.com/ZaouiAmed/Neurips2020_RejectOption`. For this experimental study, we consider the same algorithm for both estimation tasks and build three plug-in ε-predictors based respectively on support vector machines (`svm`), random forests (`rf`), and kNN (`knn`) algorithms. Besides, to avoid non continuity issues, we add the random perturbation $\zeta \sim \mathcal{U}[0, 10^{-10}]$ to all of the considered methods as described in Section 3.4.2. The performance is evaluated on two benchmark datasets: *QSAR aquatic toxicity* and *Airfoil Self-Noise* coming from the UCI database. We refer to these two datasets as `aquatic` and `airfoil` respectively. For all datasets, we split the data into three parts (50 % train labeled, 20 % train unlabeled, 30 % test). The first part is used to estimate both regression and variance functions, while the second part is used to compute the empirical cumulative distribution function. Finally, for each $\varepsilon \in \{i/10, \quad i = 0, \ldots, 9\}$ and each plug-in ε-predictor, we compute the empirical rejection rate $\hat{r}$ and the empirical error $\widehat{\mathrm{Err}}$ on the test set. This procedure is repeated 100 times and we report the average performance on the test set alongside its standard deviation. We employ the 10-fold cross-validation to select the parameter $k \in \{5, 10, 15, 20, 30, 50, 70, 100, 150\}$ of the kNN algorithm. For random forests and svm procedures, we used respectively the `R` packages `randomForest` and `e1071` with default parameters.

3.5.1 Datasets

The datasets used for the experiments are briefly described below:
QSAR aquatic toxicity has been used to develop quantitative regression QSAR models to predict acute aquatic toxicity towards the fish Pimephales promelas. This dataset is composed of $n = 546$ observations for which 8 numerical features are measured. The output takes its values in $[0.12, 10.05]$.
Airfoil Self-Noise is composed of $n = 1503$ observations for which 5 features are measured. This dataset is obtained from a series of aerodynamic and acoustic tests. The output is the scaled sound pressure level, in decibels. It takes its values in $[103, 140]$.

Since the variance function plays a key role in the construction of the plug-in ε-predictor, we display in Figure 3.2 the histogram of an estimate of $\sigma^2(X)$ produced by the random forest algorithm. More specifically, for each $i = 1, \ldots, n$, we evaluate $\hat{\sigma}^2(X_i)$ by 10-fold cross-validation and build the histogram of $(\hat{\sigma}^2(X_i))_{i=1,\ldots,n}$ thereafter. Left and right panels of Figure 3.2 deal respectively with the `aquatic` and `airfoil` datasets and reflect two different situations where the use of reject option is relevant. The estimated variance in the `airfoil` dataset is typically large (about 40% of the values are larger than 10) and then we may have some doubts in the associated prediction. According to the `aquatic` dataset, main part of the estimated values $\hat{\sigma}^2$ is smaller than 1 and then the use of the reject option may seem less significant. However, in this case, the predictions produced by the plug-in ε-predictors would be very accurate.

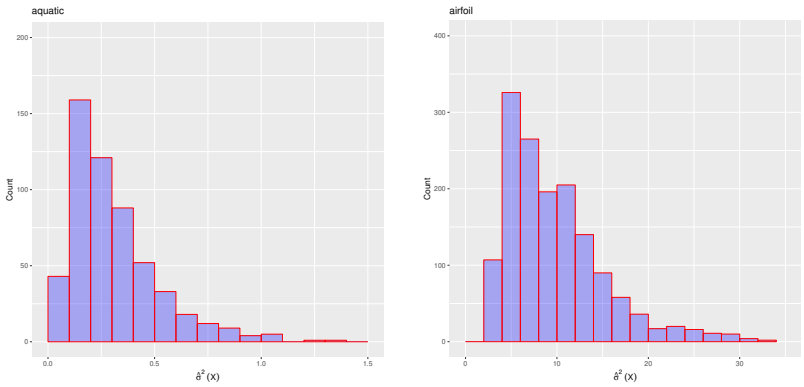


Figure 3.2 – Histogram of the estimates of $\sigma^2(X)$

3.5.2 Results

We present the obtained results in Figure 3.3 and Table 3.1. We make a focus on the values of $\varepsilon \in \{0, 0.2, 0.5, 0.8\}$. As a general picture, the results are reflecting our theory: the empirical errors of the plug-in ε-predictors are decreasing *w.r.t.* ε for both datasets and their empirical rejection rates are very close to their expected values. Indeed, Table 3.1 displays how precise is the estimation of the rejection rate whatever the method used. This is in accordance with our theoretical findings. Moreover, the empirical errors of the plug-in ε-predictors based on the random forests and kNN algorithms are decreasing *w.r.t.* ε for both datasets. As expected, the use of the reject option improves the prediction precision. As an illustration, for `airfoil` dataset and the predictor based on random forests, the error is divided by 2 if we reject 50% of the data. However, we discover that the decrease for the prediction error is not systematic. In the case of plug-in ε-predictor based on the svm algorithm and with the `aquatic` dataset, we observe a strange curve for the error rate (see Figure 3.3-left). We conjecture that this phenomenon is due to a poor estimation of the variance. Indeed, in Figure 3.4, we present the performance of some kind of hybrid plug-in ε-predictors: we still use the svm algorithm to estimate the regression function; the variance function estimation is done based on svm (dashed line), random forests (dotted line), and kNN (dash-dotted line). From Figure 3.4, we observe that the empirical error $\widehat{\mathrm{Err}}$ is now decreasing *w.r.t.* ε for the hybrid predictors based on svm and random forests, and that the performance is quite good.

Table 3.1 – Performances of the three plug-in ε-predictors on the real datasets `aquatic`, and `airfoil`.

	aquatic						airfoil					
	svm		rf		knn		svm		rf		knn	
1-ε	$\widehat{\mathrm{Err}}$	$1-\hat{r}$	$\widehat{\mathrm{Err}}$	$1-\hat{r}$	$\widehat{\mathrm{Err}}$	$1-\hat{r}$	$\widehat{\mathrm{Err}}$	$1-\hat{r}$	$\widehat{\mathrm{Err}}$	$1-\hat{r}$	$\widehat{\mathrm{Err}}$	$1-\hat{r}$
1	1.38 (0.18)	1.00 (0.00)	1.34 (0.18)	1.00 (0.00)	2.29 (0.27)	1.00 (0.00)	11.81 (1.03)	1.00 (0.00)	14.40 (1.04)	1.00 (0.00)	35.40 (2.05)	1.00 (0.00)
0.8	1.08 (0.17)	0.81 (0.05)	1.04 (0.16)	0.80 (0.05)	1.98 (0.26)	0.80 (0.04)	8.27(0.86)	0.80 (0.03)	10.26 (0.95)	0.80 (0.03)	31.13 (1.96)	0.80 (0.03)
0.5	0.91 (0.18)	0.50 (0.06)	0.81 (0.18)	0.50 (0.06)	1.51 (0.30)	0.50 (0.06)	5.15 (0.92)	0.50 (0.04)	7.22 (0.92)	0.50 (0.3)	22.42 (2.13)	0.50 (0.03)
0.2	1.01 (0.32)	0.19 (0.05)	0.55 (0.21)	0.20 (0.05)	0.75 (0.37)	0.19 (0.05)	2.6 (0.64)	0.20 (0.03)	4.00 (0.74)	0.20 (0.03)	17.27 (3.00)	0.19 (0.03)

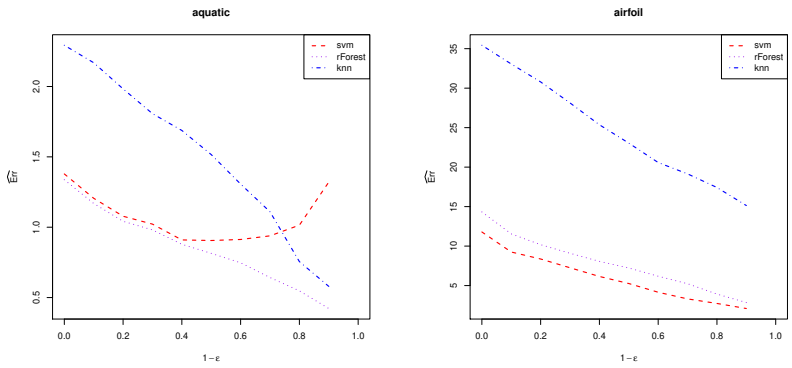


Figure 3.3 – Visual description of the performance of three plug-in ε-predictors on the `aquatic`, and `airfoil` datasets.

3.5.3 Benefit of aggregation in regression with reject option

This section is a part of the article [105]. We illustrate our aggregation methods in the regression with reject option with two real datasets: *Concrete Compressive Strength* and *Airfoil Self-Noise* (see Section 2.4.3).

We may have some doubts in the associated prediction on two real datasets since the estimated variance is large. We evaluate the performance of the procedure on two real datasets considering the same algorithm for both estimation tasks (same approach to estimate the regression and variance functions) and build four plug-in ε-predictors based respectively, on support vector machines (`svm`),

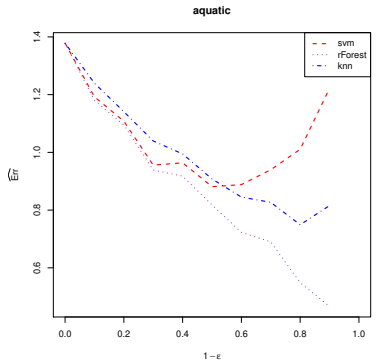


Figure 3.4 – Additional description of the performance of the plug-in procedure on `aquatic` dataset.

random forests (`rf`), and Lasso (`Lasso`) and kNN (`knn`) algorithms. We take $\zeta \sim \mathcal{U}([0, 10^{-7}])$. In particular, we run 100 times the procedure where we split the data each time in three: $\mathcal{D}_{N_1}$ with $N_1 = 780$ for Concrete Compressive Strength and $N_1 = 1253$ for Airfoil Self-Noise, $\mathcal{D}_N$ with $N = 150$ and $\mathcal{D}_T$ with $T = 100$. The dataset $\mathcal{D}_T$ is exploited to calculate the empirical rejection rate $\hat{r}$ and the empirical error $\widehat{\mathrm{Err}}$ for each $\varepsilon \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. From these estimations, we compute the average and standard deviation (between parentheses) of $\hat{r}$ and $\widehat{\mathrm{Err}}$. The results are reported in Table 3.2 with $\varepsilon \in \{0, 0.2, 0.5, 0.8\}$ and in Figure 3.5.

Table 3.2 – Performances of the four plug-in ε-predictors on the real datasets `Concrete compressive strength`, and `airfoil self-Noise`.

`Concrete compressive strength`

	`rf`		`svm`		`knn`		`Lasso`	
ε	$\widehat{\mathrm{Err}}$	$\hat{r}$	$\widehat{\mathrm{Err}}$	$\hat{r}$	$\widehat{\mathrm{Err}}$	$\hat{r}$	$\widehat{\mathrm{Err}}$	$\hat{r}$
0	31.33 (7.82)	0.00 (0.00)	46.83 (8.65)	0.00 (0.00)	87.32 (15.74)	0.00 (0.00)	111.72 (16.13)	0.00 (0.00)
0.2	20.99 (5.72)	0.21 (0.06)	33.98 (7.54)	0.20 (0.06)	65.61 (16.12)	0.20 (0.06)	89.91 (14.63)	0.19 (0.05)
0.5	13.47 (6.08)	0.48 (0.06)	21.73 (7.38)	0.50 (0.06)	46.71 (19.89)	0.50 (0.06)	66.86 (16.13)	0.50 (0.07)
0.8	7.26 (8.29)	0.81 (0.05)	18.25 (11.70)	0.81 (0.05)	28.58 (28.66)	0.79 (0.11)	52.32 (16.99)	0.81 (0.05)

`Airfoil Self-Noise`

	`rf`		`svm`		`knn`		`Lasso`	
ε	$\widehat{\mathrm{Err}}$	$\hat{r}$	$\widehat{\mathrm{Err}}$	$\hat{r}$	$\widehat{\mathrm{Err}}$	$\hat{r}$	$\widehat{\mathrm{Err}}$	$\hat{r}$
0	12.57 (1.99)	0.00 (0.00)	10.36 (2.68)	0.00 (0.00)	33.97 (4.45)	0.00 (0.00)	22.51 (3.62)	0.00 (0.00)
0.2	8.53 (1.48)	0.20 (0.05)	7.42 (2.09)	0.19 (0.05)	29.25 (4.03)	0.20 (0.06)	20.52 (3.17)	0.19 (0.06)
0.5	6.25 (1.41)	0.52 (0.07)	3.89 (1.45)	0.50 (0.06)	22.42 (4.10)	0.49 (0.06)	15.51 (3.37)	0.50 (0.06)
0.8	2.82 (1.00)	0.80 (0.05)	1.93 (0.93)	0.80 (0.05)	14.95 (7.12)	0.80 (0.07)	9.07 (3.19)	0.80 (0.05)

First of all, we recall that for $\varepsilon = 0$, the measure of risk of ε-predictor match with the error of the approach we use to estimate the regression function f^*. We remark that the best plug-in ε-predictor with $\varepsilon = 0$ is the `rf` method for concrete compressive strength and the `svm` method for airfoil self-noise. Their errors are 31.33 and 10.36 respectively. Notice that all the corresponding errors diminishes with ε and rejection rate is close to ε.
We recall that our aim is not to build a regression rule that reduces the error rate. The motivation for introducing the plug-in ε-predictor is only to improve the confidence on prediction. Since the construction of the optimal rule depends on the estimators of f^* and σ^2, poor estimators would lead to bad plug-in ε-predictor. Now, we hope that our aggregation methods improve the accuracy of the

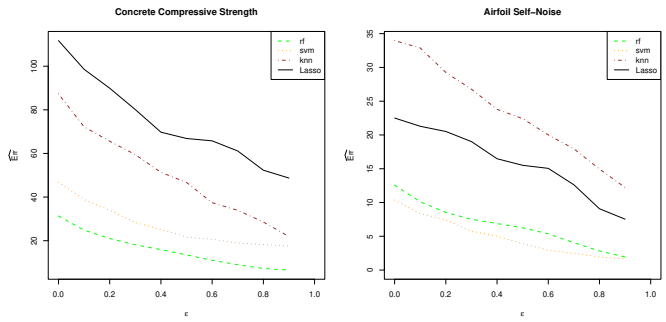


Figure 3.5 – Visual description of the performance of four plug-in ε-predictors.

procedure.
For a comparative study, we evaluate the performance of the *plug-in ε-predictor* considering the same algorithm for both estimation tasks of estimating the regression and the variance functions: we build four plug-in ε-predictors based respectively, on support vector machines (`svm`), random forests (`rf`), and Lasso (`Lasso`), kNN (`knn`) algorithms. We compare these methods to our aggregation procedures `C` and `MS`. The dictionaries $\mathcal{F}$, $\mathcal{G}_1$ and $\mathcal{G}_2$ are described in Section 2.4.3. The labeled $\mathcal{D}_M$ plays the role of $\mathcal{D}_N$ in chapter 2. For each $\varepsilon \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, and each plug-in ε-predictor, we compute the empirical rejection rate $\hat{r}$ and the empirical error $\widehat{\mathrm{Err}}$. We take $\zeta \sim \mathcal{U}([0, 10^{-7}])$. So, we repeat independently 100 times the following steps:

1. simulate four datasets $\mathcal{D}_n$, $\mathcal{D}_M$, $\mathcal{D}_N$ and $\mathcal{D}_T$ with $M = 150$, $N = 150$ and $T = 100$;
2. based on $\mathcal{D}_n$, we compute the estimators in $\mathcal{F}$, and then based on $\mathcal{D}_M$, we compute the aggregates $\hat{f}_{\mathtt{MS}}$ and $\hat{f}_{\mathtt{CM}}$. Then, we compute the `knn`, `Lasso`, `rf` and `svm` estimators of the regression function on $\mathcal{D}_n \cup \mathcal{D}_M$;
3. based on $\mathcal{D}_n$ and $\hat{f}_{\mathtt{MS}}$ (resp. $\hat{f}_{\mathtt{CM}}$), we compute the estimators in $\mathcal{G}_1$ (resp. $\mathcal{G}_2$). Then, based on $\mathcal{D}_M$ we calculate $\hat{\sigma}^2_{\mathtt{MS}}$ and $\hat{\sigma}^2_{\mathtt{CM}}$. From $\mathcal{D}_n \cup \mathcal{D}_M$, we compute the `knn`, `Lasso`, `rf` and `svm` estimators of σ^2;
4. based on $\mathcal{D}_N$, we compute the empirical cumulative distribution function of the randomized estimators $\hat{\hat{\sigma}}^2(X)$;
5. finally, over $\mathcal{D}_T$, we compute the empirical rejection rate $\hat{r}$ and the empirical error $\widehat{\mathrm{Err}}$ for the considered $\hat{\Gamma}_\varepsilon$.

From these estimations, we compute the average and standard deviation (between brackets) of $\hat{r}$ and $\widehat{\mathrm{Err}}$. The results are reported in Table 3.3 with $\varepsilon \in \{0, 0.2, 0.5, 0.8\}$ and in Figure 3.6. Our main observation is that the `C` plug-in ε-predictor has the same performance as `MS` plug-in ε-predictor. According to the rejection rate, we recall that our theory related to this point is distribution free and this is also observed in Table 3.3 since all rejection rate have the approximately prescribed level ε. Importantly, note that both aggregation-based methods require splitting the data (part for estimation and a part for aggregation) while the other plug-in ε-predictors (such as `rf`) do not. However, our plug-in ε-predictors based on aggregation have a similar performance as the best. This result validates the relevance of our strategy.

Table 3.3 – Performances of the six plug-in ε-predictors on the real datasets `Concrete compressive strength`, and `Airfoil Self-Noise`.

`Concrete compressive strength`

	rf		svm		knn		Lasso		C		MS	
ε	$\widehat{\text{Err}}$	$\hat{r}$	$\widehat{\text{Err}}$	$\hat{r}$	$\widehat{\text{Err}}$	$\hat{r}$	$\widehat{\text{Err}}$	$\hat{r}$	$\widehat{\text{Err}}$	$\hat{r}$	$\widehat{\text{Err}}$	$\hat{r}$
0	30.27 (7.16)	0.00 (0.00)	45.03 (9.03)	0.00 (0.00)	85.59 (16.11)	0.00 (0.00)	110.09 (16.01)	0.00 (0.00)	34.59 (7.74)	0.00 (0.00)	35.08 (7.44)	0.00 (0.00)
0.2	20.07 (5.65)	0.20 (0.05)	32.97 (7.03)	0.20 (0.05)	62.40 (13.93)	0.20 (0.06)	86.51 (11.72)	0.20 (0.05)	23.07 (6.70)	0.20 (0.05)	24.76 (6.91)	0.20 (0.05)
0.5	13.25 (5.38)	0.48 (0.07)	22.45 (7.16)	0.49 (0.07)	44.98 (15.11)	0.49 (0.07)	64.95 (11.08)	0.49 (0.07)	13.24 (3.87)	0.49 (0.06)	15.14 (4.18)	0.49 (0.07)
0.8	7.91 (8.26)	0.80 (0.05)	17.92 (13.69)	0.80 (0.04)	30.02 (23.85)	0.78 (0.09)	55.91 (20.09)	0.80 (0.06)	8.50 (5.33)	0.80 (0.05)	10.19 (9.77)	0.80 (0.05)

`Airfoil Self-Noise`

	rf		svm		knn		Lasso		C		MS	
ε	$\widehat{\text{Err}}$	$\hat{r}$	$\widehat{\text{Err}}$	$\hat{r}$	$\widehat{\text{Err}}$	$\hat{r}$	$\widehat{\text{Err}}$	$\hat{r}$	$\widehat{\text{Err}}$	$\hat{r}$	$\widehat{\text{Err}}$	$\hat{r}$
0	12.80 (1.84)	0.00 (0.00)	10.49 (2.18)	0.00 (0.00)	33.78 (4.33)	0.00 (0.00)	23.23 (3.71)	0.00 (0.00)	10.30 (1.98)	0.00 (0.00)	10.77 (2.22)	0.00 (0.00)
0.2	8.67 (1.29)	0.20 (0.05)	7.02 (1.86)	0.21 (0.05)	29.15 (4.60)	0.20 (0.05)	21.16 (3.49)	0.20 (0.06)	6.19 (1.17)	0.21 (0.05)	5.82 (1.41)	0.21 (0.05)
0.5	6.08 (1.13)	0.49 (0.07)	3.84 (0.98)	0.48 (0.07)	21.95 (4.42)	0.49 (0.07)	14.83 (3.39)	0.49 (0.08)	4.09 (0.99)	0.49 (0.07)	3.73 (1.15)	0.49 (0.06)
0.8	2.92 (1.02)	0.80 (0.04)	1.97 (1.06)	0.80 (0.05)	14.25 (6.94)	0.81 (0.08)	8.76 (2.64)	0.80 (0.05)	1.68 (0.64)	0.80 (0.04)	1.67 (0.73)	0.80 (0.05)

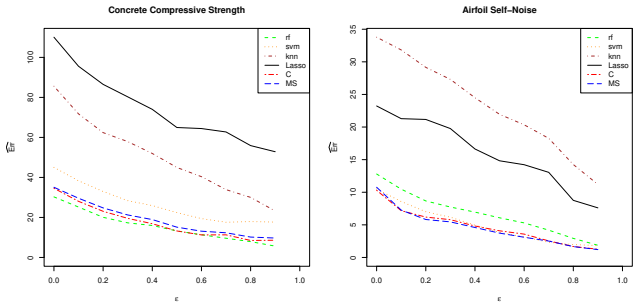


Figure 3.6 – Visual description of the performance of six plug-in ε-predictors.

3.6 Conclusion

We generalized the use of the reject option to the regression setting. We investigated the particular setting where the rejection rate is bounded. In this framework, an optimal rule is derived, it relies on thresholding of the variance function. Based on the *plug-in* principle, we derived a semi-supervised algorithm that can be applied on top of any off-the-shelf estimators of both regression and variance functions. One of the main features of the proposed procedure is that it precisely controls the probability of rejection. We derived general consistency results on rejection rate and on excess risk. We also established rates of convergence for the predictor with reject option when the regression and the variance functions are estimated by kNN algorithm. In future work, we plan to apply our methodology to the high-dimensional setting, taking advantage of sparsity structure of the data.

Approaches based on reject option may be helpful at least from two perspectives. First, when human action is limited by time or any other constraint, reject option is an efficient tool to prioritize the human action. On the other hand, in a world where automatic decisions need to be balanced and considered with caution, abstaining from prediction is one way to prevent from damageable decision-making. In particular, human is more likely able to detect anomalies such as bias in data. In a manner of speaking, the use of the reject option compromises between human and machines! Our numerical and theoretical analyses support this idea, in particular because our estimation of the rejection rate is accurate.

While the rejection rate has to be fixed according to the considered problem, it appears that the main drawback of our approach is that border instances may be automatically treated while they would have deserved a human consideration. From a general perspective, this is a weakness of all methods based on reject option. This inconvenience is even stronger when the conditional variance function is poorly estimated.

3.7 Supplementary material

This supplementary material is organized as follows. Section 3.7.1 provides all proofs of results related to the optimal predictors (that is, Propositions 4, 5 11). In Sections 3.7.2 and 3.7.3 we prove Theorem 11 that establishes the consistency and Theorem 12 that states the rates of convergence of the plug-in ε-predictor $\hat{\Gamma}_\varepsilon$ respectively. We further establish several finite sample guarantees on kNN estimator in Section 3.7.4. To help readability of the chapter, we provide in Section 3.7.5 some technical tools that are used for the proofs.

3.7.1 Proofs for optimal predictors

Proof of Proposition 4. By definition of $\mathcal{R}_\lambda$, we have for any predictor with reject option Γ_f

$$\begin{aligned}
\mathcal{R}_\lambda(\Gamma_f) &= \mathbb{E}\left[(Y-f(X))^2\mathbb{1}_{\{|\Gamma_f(X)|=1\}}\right] + \lambda\mathbb{P}(|\Gamma_f(X)|=0) \\
&= \mathbb{E}\left[(Y-f^*(X)+f^*(X)-f(X))^2\mathbb{1}_{\{|\Gamma_f(X)|=1\}}\right] + \lambda\mathbb{P}(|\Gamma_f(X)|=0) \\
&= \mathbb{E}\left[(Y-f^*(X))^2\mathbb{1}_{\{|\Gamma_f(X)|=1\}}\right] + \mathbb{E}\left[(f^*(X)-f(X))^2\mathbb{1}_{\{|\Gamma_f(X)|=1\}}\right] \\
&\quad +2\mathbb{E}\left[(Y-f^*(X))(f^*(X)-f(X))\mathbb{1}_{\{|\Gamma_f(X)|=1\}}\right] + \lambda\mathbb{P}(|\Gamma_f(X)|=0) \ .
\end{aligned}$$

Since

$$\mathbb{E}\left[(Y-f^*(X))(f^*(X)-f(X))\mathbb{1}_{\{|\Gamma_f(X)|=1\}}\right] = 0 \ ,$$

and

$$\mathbb{E}\left[(Y-f^*(X))^2\mathbb{1}_{\{|\Gamma_f(X)|=1\}}\right] = \mathbb{E}\left[\sigma^2(X)\mathbb{1}_{\{|\Gamma_f(X)|=1\}}\right] \ ,$$

we deduce,

$$\begin{aligned}
\mathcal{R}_\lambda(\Gamma_f) &= \mathbb{E}\left[(f^*(X)-f(X))^2\mathbb{1}_{\{|\Gamma_f(X)|=1\}}\right] + \mathbb{E}\left[\sigma^2(X)\mathbb{1}_{\{|\Gamma_f(X)|=1\}} + \lambda(1-\mathbb{1}_{\{|\Gamma_f(X)|=1\}})\right] \\
&= \mathbb{E}\left[\left\{(f^*(X)-f(X))^2 + (\sigma^2(X)-\lambda)\right\}\mathbb{1}_{\{|\Gamma_f(X)|=1\}}\right] + \lambda \ . \qquad (3.6)
\end{aligned}$$

Clearly, on the event $\{|\Gamma_f(X)|=1\}$, the mapping $f \mapsto (f^*(X)-f(X))^2 + \left(\sigma^2(X)-\lambda\right)$ achieves its minimum at $f=f^*$. Then, it remains to consider the minimization of

$$\Gamma \mapsto \mathbb{E}\left[\left\{(\sigma^2(X)-\lambda)\right\}\mathbb{1}_{\{|\Gamma(X)|=1\}}\right] + \lambda \ ,$$

on the set Υ_{f^*}, which leads to $\{|\Gamma(X)|=1\} = \{\sigma^2(X)\le\lambda\}$. Putting all together, we get

$$\{|\Gamma^*_\lambda(X)|=1\} = \{\sigma^2(X)\le\lambda\} \ \text{ and on this event } \ \Gamma^*_\lambda(X) = \{f^*(X)\} \ ,$$

and point 1. of Proposition 4 is proven. For the second point, we observe that for $0<\lambda<\lambda'$,

$$\{|\Gamma^*_\lambda(X)|=1\} = \{\sigma^2(X)\le\lambda\} \subset \{\sigma^2(X)\le\lambda'\} = \{|\Gamma^*_{\lambda'}(X)|=1\} \ .$$

From this inclusion, we deduce $r(\Gamma^*_{\lambda'}) \le r(\Gamma^*_\lambda)$. Furthermore, using the relation $\{|\Gamma^*_\lambda(X)|=1\} = \{\sigma^2(X)\le\lambda\}$ and if we denote by $a_\lambda = \mathbb{P}\left(|\Gamma^*_\lambda(X)|=1\right)$ we have

$$\begin{aligned}
\mathrm{Err}\left(\Gamma^*_\lambda\right) - \mathrm{Err}\left(\Gamma^*_{\lambda'}\right) &= \frac{1}{a_\lambda}\mathbb{E}\left[(Y-f^*(X))^2\mathbb{1}_{\{\sigma^2(X)\le\lambda\}}\right] - \frac{1}{a_{\lambda'}}\mathbb{E}\left[(Y-f^*(X))^2\mathbb{1}_{\{\sigma^2(X)\le\lambda'\}}\right] \\
&= \left(\frac{1}{a_\lambda}-\frac{1}{a_{\lambda'}}\right)\mathbb{E}\left[(Y-f^*(X))^2\mathbb{1}_{\{\sigma^2(X)\le\lambda\}}\right] \\
&\quad - \frac{1}{a_{\lambda'}}\mathbb{E}\left[(Y-f^*(X))^2\mathbb{1}_{\{\lambda<\sigma^2(X)\le\lambda'\}}\right] \ . \qquad (3.7)
\end{aligned}$$

By definition of $\sigma^2(X)$, we can write

$$\begin{aligned}\mathbb{E}\left[(Y-f^*(X))^2\mathbb{1}_{\{\sigma^2(X)\leq\lambda\}}\right] &= \mathbb{E}\left[\mathbb{1}_{\{\sigma^2(X)\leq\lambda\}}\mathbb{E}\left[(Y-f^*(X))^2|X\right]\right]\\ &= \mathbb{E}\left[\mathbb{1}_{\{\sigma^2(X)\leq\lambda\}}\sigma^2(X)\right]\leq\lambda a_\lambda,\end{aligned}$$

and then

$$\left(\frac{1}{a_\lambda}-\frac{1}{a_{\lambda'}}\right)\mathbb{E}\left[(Y-f^*(X))^2\mathbb{1}_{\{\sigma^2(X)\leq\lambda\}}\right]\leq\lambda\left(1-\frac{a_\lambda}{a_{\lambda'}}\right).$$

In the same way, we obtain

$$\frac{1}{a_{\lambda'}}\mathbb{E}\left[(Y-f^*(X))^2\mathbb{1}_{\{\lambda\leq\sigma^2(X)\leq\lambda'\}}\right]\geq\frac{\lambda}{a_{\lambda'}}\left(a_{\lambda'}-a_\lambda\right)=\lambda\left(1-\frac{a_\lambda}{a_{\lambda'}}\right).$$

From Equation (3.7), we then get $\mathrm{Err}\left(\Gamma^*_\lambda\right)\leq\mathrm{Err}\left(\Gamma^*_{\lambda'}\right)$. □

Proof of Proposition 5. First of all, observe that for any $\varepsilon\in(0,1)$, if we set $\lambda_\varepsilon=F^{-1}_{\sigma^2}(1-\varepsilon)$, then the optimal predictor Γ^*_λ given by (4.2) with $\lambda=\lambda_\varepsilon$ is such that,

$$r\left(\Gamma^*_{\lambda_\varepsilon}\right)=\mathbb{P}\left(\sigma^2(X)\geq\lambda_\varepsilon\right)=\mathbb{P}\left(F_{\sigma^2}(\sigma^2(X))\geq1-\varepsilon\right)=\varepsilon\ .$$

We need to prove that any predictor Γ_f such that $r\left(\Gamma_f\right)=\varepsilon'$ with $\varepsilon'\leq\varepsilon$, satisfies $\mathrm{Err}\left(\Gamma_f\right)\geq\mathrm{Err}\left(\Gamma^*_{\lambda_\varepsilon}\right)$. To this end, consider $\Gamma^*_{\lambda_{\varepsilon'}}$ with $\lambda_{\varepsilon'}=F^{-1}_{\sigma^2}(1-\varepsilon')$. On one hand, by optimality of $\Gamma^*_{\lambda_{\varepsilon'}}$ (*cf.* point 1. of Proposition 4), we have

$$\mathrm{Err}\left(\Gamma_f\right)-\mathrm{Err}\left(\Gamma^*_{\lambda_{\varepsilon'}}\right)=\frac{1}{1-\varepsilon'}\left(\mathcal{R}_{\lambda_{\varepsilon'}}\left(\Gamma_f\right)-\mathcal{R}_{\lambda_{\varepsilon'}}\left(\Gamma^*_{\lambda_{\varepsilon'}}\right)\right)\geq0\ .$$

On the other hand, since $\varepsilon'\leq\varepsilon$ implies $\lambda_\varepsilon\leq\lambda_{\varepsilon'}$, point 2. of Proposition 4 reads as

$$\mathrm{Err}\left(\Gamma^*_{\lambda_\varepsilon}\right)\leq\mathrm{Err}\left(\Gamma^*_{\lambda_{\varepsilon'}}\right)\ .$$

Combining these two facts gives the desired result. □

Proof of Proposition 11. First, from Equation (3.6), we have the following decomposition

$$\begin{aligned}\mathcal{R}_{\lambda_\varepsilon}(\Gamma_f) &= \mathbb{E}\left[\left\{(f^*(X)-f(X))^2+\sigma^2(X)-\lambda_\varepsilon\right\}\mathbb{1}_{\left\{|\Gamma_f(X)|=1\right\}}\right]+\lambda_\varepsilon\\ &= \mathbb{E}\left[(f^*(X)-f(X))^2\mathbb{1}_{\left\{|\Gamma_f(X)|=1\right\}}\right]+\mathbb{E}\left[(\sigma^2(X)-\lambda_\varepsilon)\mathbb{1}_{\left\{|\Gamma_f(X)|=1\right\}}\right]+\lambda_\varepsilon\ .\end{aligned}$$

Therefore, we deduce

$$\mathcal{E}\left(\Gamma_f\right)=\mathbb{E}\left[(f^*(X)-f(X))^2\mathbb{1}_{\left\{|\Gamma_f(X)|=1\right\}}\right]+\mathbb{E}\left[\left(\sigma^2(X)-\lambda_\varepsilon\right)\left\{\mathbb{1}_{\left\{|\Gamma_f(X)|=1\right\}}-\mathbb{1}_{\{|\Gamma^*_\varepsilon(X)|=1\}}\right\}\right],$$

and the result follows from the fact that all non zero values of $\mathbb{1}_{\left\{|\Gamma_f(X)|=1\right\}}-\mathbb{1}_{\{|\Gamma^*_\varepsilon(X)|=1\}}$ equal the sign of $\left(\sigma^2(X)-\lambda_\varepsilon\right)$ due to the fact that $\{|\Gamma^*_\varepsilon(X)|=1\}=\left\{\sigma^2(X)-\lambda_\varepsilon\leq0\right\}$. □

3.7.2 Proof of the consistency results: Theorem 11

The consistency of $\hat{\Gamma}_\varepsilon$ consists in the introduction of a pseudo oracle ε-predictor $\tilde{\Gamma}_\varepsilon$ defined for all $x\in\mathbb{R}^d$ by

$$\tilde{\Gamma}_\varepsilon(x)=\begin{cases}\left\{\hat{f}(x)\right\} & \text{if } \hat{\sigma}^2(x)\leq F^{-1}_{\hat{\sigma}^2}(1-\varepsilon)\\ \emptyset & \text{otherwise}\ .\end{cases}\tag{3.8}$$

This predictor differs from $\hat{\Gamma}_\varepsilon$ in that it knows the marginal distribution $\mathbb{P}_X$ and then it has rejection rate exactly ε. Then, we consider the following decomposition

$$\mathbb{E}\left[\mathcal{E}_{\lambda_\varepsilon}\left(\hat{\Gamma}_\varepsilon\right)\right]=\mathbb{E}\left[\mathcal{R}_{\lambda_\varepsilon}(\hat{\Gamma}_\varepsilon)-\mathcal{R}_{\lambda_\varepsilon}(\tilde{\Gamma}_\varepsilon)\right]+\mathbb{E}\left[\mathcal{E}_{\lambda_\varepsilon}\left(\tilde{\Gamma}_\varepsilon\right)\right]\ ,\tag{3.9}$$

and show that both terms in the r.h.s. go to zero.

• **Step 1.** $\mathbb{E}\left[\mathcal{E}_{\lambda_\varepsilon}\left(\tilde{\Gamma}_\varepsilon\right)\right]\to0$. We use Proposition 11 and get the following result.

Proposition 7. *Let $\varepsilon \in (0,1)$. Under Assumptions 7 and 8, the following holds*

$$\mathbb{E}\left[\mathcal{E}_{\lambda_\varepsilon}\left(\tilde{\Gamma}_\varepsilon\right)\right] \leq \mathbb{E}\left[(\hat{f}(X)-f^*(X))^2\right] + \mathbb{E}\left[|\hat{\sigma}^2(X)-\sigma^2(X)|\right] + C\mathbb{E}\left[|F_{\hat{\sigma}^2}(\lambda_\varepsilon) - F_{\sigma^2}(\lambda_\varepsilon)|\right],$$

where $C>0$ is constant which depends on the upper bounds of σ^2 and λ_ε.

Proof of Proposition 7. Let $\varepsilon \in (0,1)$. First, we recall our notation $F_{\hat{\sigma}^2}(\cdot) = \mathbb{P}_X\left(\hat{\sigma}^2(X) \leq \cdot | \mathcal{D}_n\right)$ and $\lambda_\varepsilon = F^{-1}_{\sigma^2}(1-\varepsilon)$. We also introduce $\tilde{\lambda}_\varepsilon = F^{-1}_{\hat{\sigma}^2}(1-\varepsilon)$ for the pseudo-oracle counterpart of λ_ε. A direct application of Proposition 11 yields

$$\mathcal{E}\left(\tilde{\Gamma}_\varepsilon\right) \leq \mathbb{E}_X\left[\left(\hat{f}(X)-f^*(X)\right)^2\right] + \mathbb{E}_X\left[|\sigma^2(X)-\lambda_\varepsilon|\mathbb{1}_{\left\{|\tilde{\Gamma}_\varepsilon(X)|\neq|\Gamma^*_\varepsilon(X)|\right\}}\right]. \tag{3.10}$$

We first observe that if $\sigma^2(X) \leq \lambda_\varepsilon$ and $\hat{\sigma}^2(X) \geq \tilde{\lambda}_\varepsilon$, we have either of the two cases

- $\tilde{\lambda}_\varepsilon \geq \lambda_\varepsilon$ and then $|\sigma^2(X)-\lambda_\varepsilon| \leq |\hat{\sigma}^2(X)-\sigma^2(X)|$;
- $\tilde{\lambda}_\varepsilon \leq \lambda_\varepsilon$ and then either $|\sigma^2(X)-\lambda_\varepsilon| \leq |\hat{\sigma}^2(X)-\sigma^2(X)|$ or $\hat{\sigma}^2(X) \in (\tilde{\lambda}_\varepsilon, \lambda_\varepsilon)$.

Similar reasoning holds in the case where $\sigma^2(X) \geq \lambda_\varepsilon$ and $\hat{\sigma}^2(X) \leq \tilde{\lambda}_\varepsilon$. Therefore

$$\begin{aligned}
&\mathbb{E}\left[|\sigma^2(X)-\lambda_\varepsilon|\mathbb{1}_{\left\{|\tilde{\Gamma}_\varepsilon(X)|\neq|\Gamma^*_\varepsilon(X)|\right\}}|\mathcal{D}_n\right] \\
&\qquad \leq \mathbb{E}\left[|\sigma^2(X)-\lambda_\varepsilon|\mathbb{1}_{\{|\sigma^2(X)-\lambda_\varepsilon|\leq|\hat{\sigma}^2(X)-\sigma^2(X)|\}}|\mathcal{D}_n\right] \\
&\qquad\qquad + \mathbb{1}_{\{\lambda_\varepsilon\leq\tilde{\lambda}_\varepsilon\}}\mathbb{E}\left[|\sigma^2(X)-\lambda_\varepsilon|\mathbb{1}_{\{\lambda_\varepsilon\leq\hat{\sigma}^2(X)\leq\tilde{\lambda}_\varepsilon\}}|\mathcal{D}_n\right] \\
&\qquad\qquad\qquad + \mathbb{1}_{\{\tilde{\lambda}_\varepsilon\leq\lambda_\varepsilon\}}\mathbb{E}\left[|\sigma^2(X)-\lambda_\varepsilon|\mathbb{1}_{\{\tilde{\lambda}_\varepsilon\leq\hat{\sigma}^2(X)\leq\lambda_\varepsilon\}}|\mathcal{D}_n\right] \quad .
\end{aligned}$$

From the above inequality, since σ^2 is bounded, there exists a constant $C>0$ such that

$$\mathbb{E}\left[|\sigma^2(X)-\lambda_\varepsilon|\mathbb{1}_{\left\{|\tilde{\Gamma}_\varepsilon(X)|\neq|\Gamma^*_\varepsilon(X)|\right\}}\right] \leq \mathbb{E}\left[|\hat{\sigma}^2(X)-\sigma^2(X)|\right] + C\mathbb{E}\left[|F_{\hat{\sigma}^2}(\tilde{\lambda}_\varepsilon) - F_{\hat{\sigma}^2}(\lambda_\varepsilon)|\right] \quad .$$

Now, from Assumptions 7 and 8, we have that $F_{\sigma^2}(\lambda_\varepsilon) = 1-\varepsilon = F_{\hat{\sigma}^2}(\tilde{\lambda}_\varepsilon)$. Therefore, we deduce that

$$\mathbb{E}\left[|\sigma^2(X)-\lambda_\varepsilon|\mathbb{1}_{\left\{|\tilde{\Gamma}_\varepsilon(X)|\neq|\Gamma^*_\varepsilon(X)|\right\}}\right] \leq \mathbb{E}\left[|\hat{\sigma}^2(X)-\sigma^2(X)|\right] + C\mathbb{E}\left[|F_{\sigma^2}(\lambda_\varepsilon) - F_{\hat{\sigma}^2}(\lambda_\varepsilon)|\right] \quad .$$

Putting this into Equation (3.10) gives the result in Proposition 7. □

Since $\hat{f}$ and $\hat{\sigma}^2$ are consistent *w.r.t.* the L_2 and L_1 risks respectively, the first two terms in the bound of Proposition 7 converge to zero. It remains to study the convergence of the last term. To this end, we prove that

$$\begin{aligned}
\mathbb{E}\left[|F_{\sigma^2}(\lambda_\varepsilon) - F_{\hat{\sigma}^2}(\lambda_\varepsilon)|\right] &= \mathbb{E}\left[|\mathbb{1}_{\{\sigma^2(X)\leq\lambda_\varepsilon\}} - \mathbb{1}_{\{\hat{\sigma}^2(X)\leq\lambda_\varepsilon\}}|\right] \\
&\leq \mathbb{P}\left(|\sigma^2(X)-\hat{\sigma}^2(X)| \geq |\sigma^2(X)-\lambda_\varepsilon|\right) \quad .
\end{aligned}$$

Hence, for any $\beta>0$, using Markov's Inequality we have

$$\begin{aligned}
\mathbb{E}\left[|F_{\sigma^2}(\lambda_\varepsilon) - F_{\hat{\sigma}^2}(\lambda_\varepsilon)|\right] &\leq \mathbb{P}\left(|\sigma^2(X)-\lambda_\varepsilon| \leq \beta\right) + \mathbb{P}\left(|\sigma^2(X)-\hat{\sigma}^2(X)| \geq \beta\right) \\
&\leq \mathbb{P}\left(|\sigma^2(X)-\lambda_\varepsilon| \leq \beta\right) + \frac{\mathbb{E}\left[|\hat{\sigma}^2(X)-\sigma^2(X)|\right]}{\beta} \quad .
\end{aligned}$$

Combining this last inequality with Proposition 7 and the consistency of $\hat{f}$ and $\hat{\sigma}^2$ *w.r.t.* the L_2 and L_1 risks respectively implies that for all $\beta>0$

$$\limsup_{n,N\to+\infty} \mathbb{E}\left[\mathcal{E}_{\lambda_\varepsilon}\left(\tilde{\Gamma}_\varepsilon\right)\right] \leq C\mathbb{P}\left(|\sigma^2(X)-\lambda_\varepsilon| \leq \beta\right) \quad .$$

Since the above inequality holds for all $\beta > 0$, under Assumption 7, we deduce that

$$\mathbb{E}\left[\mathcal{E}_{\lambda_\varepsilon}\left(\tilde{\Gamma}_\varepsilon\right)\right] \to 0 \ ,$$

and then this step of the proof is complete.

• **Step 2.** $\mathbb{E}\left[\mathcal{R}_{\lambda_\varepsilon}(\hat{\Gamma}_\varepsilon) - \mathcal{R}_{\lambda_\varepsilon}(\tilde{\Gamma}_\varepsilon)\right] \to 0$. Thanks to Equation (3.6), we have that

$$\mathcal{R}_{\lambda_\varepsilon}(\hat{\Gamma}_\varepsilon) - \mathcal{R}_{\lambda_\varepsilon}(\tilde{\Gamma}_\varepsilon) = \mathbb{E}_X\left[\left\{(f^*(X) - \hat{f}(X))^2 + (\sigma^2(X) - \lambda_\varepsilon)\right\}\left(\mathbb{1}_{\{|\hat{\Gamma}_\varepsilon(X)|=1\}} - \mathbb{1}_{\{|\tilde{\Gamma}_\varepsilon(X)|=1\}}\right)\right] \ .$$

Therefore, since σ^2 is bounded, there exists a constant $C > 0$ such that

$$\begin{aligned}\mathbb{E}\left[\left|\mathcal{R}_{\lambda_\varepsilon}(\hat{\Gamma}_\varepsilon) - \mathcal{R}_{\lambda_\varepsilon}(\tilde{\Gamma}_\varepsilon)\right|\right] &\leq 2\mathbb{E}\left[(f^*(X) - \hat{f}(X))^2\right] + C\mathbb{E}\left[\left|\mathbb{1}_{\{|\hat{\Gamma}_\varepsilon(X)|=1\}} - \mathbb{1}_{\{|\tilde{\Gamma}_\varepsilon(X)|=1\}}\right|\right] \\ &\leq 2\mathbb{E}\left[(f^*(X) - \hat{f}(X))^2\right] + CA_\varepsilon \ , \end{aligned} \tag{3.11}$$

where

$$A_\varepsilon = \mathbb{E}\left[\left|\mathbb{1}_{\{|\hat{\Gamma}_\varepsilon(X)|=1\}} - \mathbb{1}_{\{|\tilde{\Gamma}_\varepsilon(X)|=1\}}\right|\right] = \mathbb{E}\left[\left|\mathbb{1}_{\{\hat{F}_{\hat{\sigma}^2}(\hat{\sigma}^2(X))\geq 1-\varepsilon\}} - \mathbb{1}_{\{F_{\hat{\sigma}^2}(\hat{\sigma}^2(X))\geq 1-\varepsilon\}}\right|\right] \ . \tag{3.12}$$

Considering the fact that $\hat{f}$ is consistent *w.r.t.* the L_2 risk, it remains to treat the term A_ε. We have

$$A_\varepsilon \leq \mathbb{P}\left(|\hat{F}_{\hat{\sigma}^2}(\hat{\sigma}^2(X)) - F_{\hat{\sigma}^2}(\hat{\sigma}^2(X))| \geq |F_{\hat{\sigma}^2}(\hat{\sigma}^2(X)) - (1-\varepsilon)|\right) \ ,$$

and then, for all $\beta > 0$, the following holds

$$A_\varepsilon \leq \mathbb{P}\left(|F_{\hat{\sigma}^2}(\hat{\sigma}^2(X)) - (1-\varepsilon)| < \beta\right) + \mathbb{P}\left(|\hat{F}_{\hat{\sigma}^2}(\hat{\sigma}^2(X)) - F_{\hat{\sigma}^2}(\hat{\sigma}^2(X))| \geq \beta\right) \ . \tag{3.13}$$

Under Assumption 8, the random variable $F_{\hat{\sigma}^2}(\hat{\sigma}^2(X))$ is uniformly distributed on $[0,1]$ conditionally on $\mathcal{D}_n$. Therefore, we deduce that

$$\begin{aligned}\mathbb{P}\left(|F_{\hat{\sigma}^2}(\hat{\sigma}^2(X)) - (1-\varepsilon)| < \beta\right) &= \mathbb{E}\left[\mathbb{P}_X\left(|F_{\hat{\sigma}^2}(\hat{\sigma}^2(X)) - (1-\varepsilon)| < \beta\right)|\mathcal{D}_n\right] \\ &= \mathbb{E}\left[2\beta|\mathcal{D}_n\right] = 2\beta \ . \end{aligned} \tag{3.14}$$

According to the second term in the r.h.s. of Equation (3.13). we have that

$$\begin{aligned}\mathbb{P}\left(|\hat{F}_{\hat{\sigma}^2}(\hat{\sigma}^2(X)) - F_{\hat{\sigma}^2}(\hat{\sigma}^2(X))| \geq \beta\right) &\leq \mathbb{P}\left(\sup_{x\in\mathbb{R}}|\hat{F}_{\hat{\sigma}^2}(x) - F_{\hat{\sigma}^2}(x)| \geq \beta\right) \\ &= \mathbb{E}\left[\mathbb{P}_{\mathcal{D}_N}\left(\sup_{x\in\mathbb{R}}|\hat{F}_{\hat{\sigma}^2}(x) - F_{\hat{\sigma}^2}(x)| \geq \beta|\mathcal{D}_n\right)\right] \ , \end{aligned}$$

where $\mathbb{P}_{\mathcal{D}_N}$ is the probability measure *w.r.t.* the dataset $\mathcal{D}_N$. Since, conditionally on $\mathcal{D}_n$, $\hat{F}_{\hat{\sigma}^2}$ is the empirical counterpart of the continuous cumulative distribution function $F_{\hat{\sigma}^2}$, applying the Dvoretzky-Kiefer-Wolfowitz Inequality [61], we deduce that

$$\mathbb{P}\left(|\hat{F}_{\hat{\sigma}^2}(\hat{\sigma}^2(X)) - F_{\hat{\sigma}^2}(\hat{\sigma}^2(X))| \geq \beta\right) \leq 2\exp(-2N\beta^2) \ . \tag{3.15}$$

Putting (3.14) and (3.15) into Eq. (3.13), we have that for all $\beta > 0$

$$A_\varepsilon \leq 2\left(\beta + \exp\left(-2N\beta^2\right)\right) \ . \tag{3.16}$$

Since Equation (3.16) holds for all $\beta > 0$, we have that $A_\varepsilon \to 0$ as $N, n \to +\infty$. Hence, from the above inequality we get the desired result in **Step 2**:

$$\mathbb{E}\left[\left|\mathcal{R}_{\lambda_\varepsilon}(\hat{\Gamma}_\varepsilon) - \mathcal{R}_{\lambda_\varepsilon}(\tilde{\Gamma}_\varepsilon)\right|\right] \to 0 \ .$$

Combining **Step 1** and **Step 2** yields the convergence: $\mathbb{E}\left[\mathcal{E}_{\lambda_\varepsilon}\left(\hat{\Gamma}_\varepsilon\right)\right] \to 0$ as $N, n \to +\infty$.

• **Bound on** $\mathbb{E}\left[r(\hat{\Gamma}_\varepsilon)\right]$**.** To finish the proof of Theorem 11, it remains to control the rejection rate $\mathbb{E}\left[r(\hat{\Gamma}_\varepsilon)\right]$ and show that it satisfies $\mathbb{E}\left[\left|r(\hat{\Gamma}_\varepsilon) - \varepsilon\right|\right] \leq CN^{-1/2}$ for some constant $C > 0$. We observe that

$$\mathbb{E}\left[\left|r(\hat{\Gamma}_\varepsilon) - \varepsilon\right|\right] = \mathbb{E}\left[\left|r(\hat{\Gamma}_\varepsilon) - r(\tilde{\Gamma}_\varepsilon)\right|\right] \leq A_\varepsilon \ ,$$

where A_ε is given by Eq. (3.12). Repeating the same reasoning as in **Step 2** above, we bound A_ε as in Eq. (3.13), and get from Dvoretsky-Kiefer-Wolfowitz Inequality (see Equation (3.15)), that for all $\beta > 0$,

$$\mathbb{P}\left(|\hat{F}_{\hat{\sigma}^2}(\hat{\sigma}^2(X)) - F_{\hat{\sigma}^2}(\hat{\sigma}^2(X))| \geq \beta\right) \leq 2\exp(-2N\beta^2) \ ,$$

and from Equation (3.14),

$$\mathbb{P}\left(|F_{\hat{\sigma}^2}(\hat{\sigma}^2(X)) - (1-\varepsilon)| < \beta\right) = 2\beta \ .$$

These two bounds combined the classical peeling argument of [2] (see Lemma 12 below) imply the desired result:

$$A_\varepsilon \leq CN^{-1/2} \ . \tag{3.17}$$

3.7.3 Proof of rates of convergence: Theorem 12

In this section, we follow the same strategy as in Section 3.7.2 but here we care about rates of convergence. Moreover, we have to pay attention to the randomness we introduced in the predictor because of the use of kNN. As in Section 3.7.2, we introduce some pseudo-oracle predictor. However, this one needs to depend on the randomness we introduced in the definition of $\hat{\Gamma}_\varepsilon(x, \zeta)$. Define the pseudo-oracle ε-predictor $\tilde{\Gamma}_\varepsilon$ for all $x \in \mathbb{R}^d$ and $\zeta \in [0, u]$ as[2]

$$\tilde{\Gamma}_\varepsilon(x, \zeta) = \begin{cases} \left\{\hat{f}(x)\right\} & \text{if } \ \bar{\sigma}^2(x, \zeta) \leq F^{-1}_{\bar{\sigma}^2}(1-\varepsilon) \\ \emptyset & \text{otherwise} \ . \end{cases}$$

To study the excess risk $\mathbb{E}\left[\mathcal{E}_{\lambda_\varepsilon}\left(\hat{\Gamma}_\varepsilon\right)\right]$ of our predictor, we also consider a similar decomposition as in Eq. (3.9) and treat each of the two terms separately.

• **Step 1.** Study of $\mathbb{E}\left[\mathcal{E}_{\lambda_\varepsilon}\left(\tilde{\Gamma}_\varepsilon\right)\right]$. We establish the following result.

Proposition 8. *Assume that Assumptions 10 and 11 are fulfilled for some $\alpha \geq 0$, then the following inequality holds*

$$\mathbb{E}\left[\mathcal{E}_{\lambda_\varepsilon}\left(\tilde{\Gamma}_\varepsilon\right)\right] \leq \mathbb{E}\left[\left(f^*(X) - \hat{f}(X)\right)^2\right] + C\left(\mathbb{E}\left[\left(\sup_{x \in \mathcal{C}} |\hat{\sigma}^2(x) - \sigma^2(x)|\right)^{1+\alpha}\right] + u^{1+\alpha}\right) \ ,$$

where $C > 0$ depends only on c^* and α.

Proof. Let $\varepsilon \in (0, 1)$. We recall that $\lambda_\varepsilon = F^{-1}_{\sigma^2}(1-\varepsilon)$ and $\tilde{\lambda}_\varepsilon = F^{-1}_{\bar{\sigma}^2}(1-\varepsilon)$. Since ζ is distributed according to a Uniform distribution on $[0, u]$, we observe that

$$\left|\bar{\sigma}^2(X, \zeta) - \sigma^2(X)\right| \leq \sup_{x \in \mathcal{C}} \left|\sigma^2(x) - \hat{\sigma}^2(x)\right| + u := \hat{h}_u \ .$$

Hence, according to Theorem 2.12 in [11] (recalled in Lemma 13), we have that conditionally on $\mathcal{D}_n$

$$\left|\tilde{\lambda}_\varepsilon - \lambda_\varepsilon\right| \leq \hat{h}_u \ .$$

[2]The only difference between $\tilde{\Gamma}_\varepsilon(x, \zeta)$ and $\tilde{\Gamma}_\varepsilon(x)$ given in (3.8) is the dependency in ζ that is hidden inside $\bar{\sigma}^2$. To avoid useless additional notation, we write $\tilde{\Gamma}_\varepsilon$ for both pseudo-oracles.

Furthermore, since X and ζ are independent, we can use Proposition 11 and get

$$\mathbb{E}\left[\mathcal{E}_{\lambda_\varepsilon}\left(\tilde{\Gamma}_\varepsilon\right)\right] \leq \mathbb{E}\left[\left(\hat{f}(X) - f^*(X)\right)^2\right] + \mathbb{E}\left[|\sigma^2(X) - \lambda_\varepsilon| \mathbb{1}_{\left\{|\tilde{\Gamma}_\varepsilon(X,\zeta)| \neq |\Gamma^*_\varepsilon(X)|\right\}}\right] \ .$$

On the event $\left\{|\tilde{\Gamma}_\varepsilon(X,\zeta)| \neq |\Gamma^*_\varepsilon(X)\right\}$, we note that

$$|\sigma^2(X) - \lambda_\varepsilon| \leq |\bar{\sigma}^2(X,\zeta) - \sigma^2(X)| + |\tilde{\lambda}_\varepsilon - \lambda_\varepsilon| \ .$$

Therefore, conditional on $\mathcal{D}_n$, we deduce the following

$$\begin{aligned}
\mathbb{E}_{(X,\zeta)}\left[|\sigma^2(X) - \lambda_\varepsilon| \mathbb{1}_{\left\{|\tilde{\Gamma}_\varepsilon(X,\zeta)| \neq |\Gamma^*_\varepsilon(X)|\right\}}\right] &\leq \mathbb{E}_{(X,\zeta)}\left[|\sigma^2(X) - \lambda_\varepsilon| \mathbb{1}_{\left\{|\sigma^2(X)-\lambda_\varepsilon| \leq |\bar{\sigma}^2(X,\zeta)-\sigma^2(X)| + |\tilde{\lambda}_\varepsilon - \lambda_\varepsilon|\right\}}\right] \\
&\leq \mathbb{E}_X\left[|\sigma^2(X) - \lambda_\varepsilon| \mathbb{1}_{\left\{|\sigma^2(X)-\lambda_\varepsilon| \leq 2\hat{h}_u\right\}}\right] \\
&\leq 2\hat{h}_u \mathbb{P}_X\left(|\sigma^2(X) - \lambda_\varepsilon| \leq 2\hat{h}_u\right) \ .
\end{aligned}$$

Finally, applying Assumption 11, we deduce that there exists a constant $C > 0$ such that

$$\mathbb{E}\left[|\sigma^2(X) - \lambda_\varepsilon| \mathbb{1}_{\left\{|\tilde{\Gamma}_\varepsilon(X,\zeta)| \neq |\Gamma^*_\varepsilon(X)|\right\}}\right] \leq C\left(\mathbb{E}\left[\sup_{x\in\mathcal{C}} \left|\sigma^2(x) - \hat{\sigma}^2(x)\right|^{1+\alpha}\right] + u^{1+\alpha}\right) \ ,$$

which ends the proof. □

Based on Proposition 8, the control of $\mathbb{E}\left[\mathcal{E}_{\lambda_\varepsilon}\left(\tilde{\Gamma}_\varepsilon\right)\right]$ requires a bound on $\mathbb{E}\left[\left(f^*(X) - \hat{f}(X)\right)^2\right]$ and on $\mathbb{E}\left[\sup_{x\in\mathcal{C}} \left|\sigma^2(x) - \hat{\sigma}^2(x)\right|^{1+\alpha}\right]$. The first of these two terms relies on estimation of the regression function with kNN algorithm and is rather well studied. In particular, thanks to Proposition 14 we have with the choice $k_n \propto n^{2/(d+2)}$

$$\mathbb{E}\left[\left(\hat{f}(X) - f^*(X)\right)^2\right] \leq Cn^{-2/(d+2)} \ , \tag{3.18}$$

where $C > 0$ is a constant which depends on f^*, c_0, $\mathcal{C}$, and d. Then it remains to bound the second term which is the purpose of Proposition 9 that relies on the rate of convergence of the kNN estimator of the conditional variance $\hat{\sigma}^2$ in supremum norm. This result says that under our assumptions and for the choice $k_n \propto n^{2/(d+2)}$, we have that

$$\mathbb{E}\left[\left(\sup_{x\in\mathcal{C}} \left|\hat{\sigma}^2(x) - \sigma^2(x)\right|\right)^{1+\alpha}\right] \leq C \log(n)^{(\alpha+1)} n^{-(\alpha+1)/(d+2)} \ ,$$

for a constant $C > 0$ that depends on f^*, σ^2, c_0, $\mathcal{C}$, and on the dimension d. Putting this last inequality and Eq. (3.18) into the upper bound on the excess risk of $\tilde{\Gamma}_\varepsilon$ from Proposition 8 we show that when we set $u = u_n \leq n^{-1/(d+2)}$ we can write

$$\mathbb{E}\left[\mathcal{E}_{\lambda_\varepsilon}\left(\tilde{\Gamma}_\varepsilon\right)\right] \leq C\left(n^{-2/(d+2)} + \log(n)^{(\alpha+1)} n^{-(\alpha+1)/(d+2)}\right) \ ,$$

where $C > 0$ is a constant which depends on f^*, σ^2, c_0, c^*, α, $\mathcal{C}$, and on the dimension d. This ends the first step of the proof.

• **Step 2.** Study of $\mathbb{E}\left[\mathcal{R}_{\lambda_\varepsilon}(\hat{\Gamma}_\varepsilon) - \mathcal{R}_{\lambda_\varepsilon}(\tilde{\Gamma}_\varepsilon)\right]$. Since X and ζ are independent, as in **Step 2** of the proof of Theorem 11 (*cf.* Eq. (3.11)), we get

$$\mathbb{E}\left[\left|\mathcal{R}_{\lambda_\varepsilon}(\hat{\Gamma}_\varepsilon) - \mathcal{R}_{\lambda_\varepsilon}(\tilde{\Gamma}_\varepsilon)\right|\right] \leq 2\mathbb{E}\left[\left(f^*(X) - \hat{f}(X)\right)^2\right] + CA_\varepsilon \ ,$$

where A_ε is defined similarly as in Equation (3.12) with a small modification due to the random perturbation we made on $\hat{\sigma}^2$. Similarly we have

$$A_\varepsilon \leq \mathbb{P}\left(|\hat{F}_{\bar{\sigma}^2}(\bar{\sigma}^2(X,\zeta)) - \{F_{\bar{\sigma}^2}(\bar{\sigma}^2(X,\zeta))| \geq |F_{\bar{\sigma}^2}(\bar{\sigma}^2(X,\zeta)) - (1-\varepsilon)|\right) \ .$$

Therefore using the same arguments as in **Step 2** of the proof of Theorem 11 to get (3.17), it is easy to see that there exists $C > 0$ such that $A_\varepsilon \leq CN^{-1/2}$. Then, we deduce

$$\mathbb{E}\left[\left|\mathcal{R}_{\lambda_\varepsilon}(\hat{\Gamma}_\varepsilon) - \mathcal{R}_{\lambda_\varepsilon}(\tilde{\Gamma}_\varepsilon)\right|\right] \leq 2\mathbb{E}\left[\left(f^*(X) - \hat{f}(X)\right)^2\right] + CN^{-1/2}.$$

Finally, an application of Theorem 14 yields

$$\mathbb{E}\left[\left|\mathcal{R}_{\lambda_\varepsilon}(\hat{\Gamma}_\varepsilon) - \mathcal{R}_{\lambda_\varepsilon}(\tilde{\Gamma}_\varepsilon)\right|\right] \leq C\left(n^{-2/(d+2)} + N^{-1/2}\right),$$

where $C > 0$ is a constant which depends on f^*, c_0, $\mathcal{C}$, and d. This ends **Step 2** of the proof.

Lastly, we combine the results in **Step 1** and **Step 2**, together with the decomposition

$$\mathbb{E}\left[\mathcal{E}_{\lambda_\varepsilon}\left(\hat{\Gamma}_\varepsilon\right)\right] = \mathbb{E}\left[\mathcal{R}_{\lambda_\varepsilon}(\hat{\Gamma}_\varepsilon) - \mathcal{R}_{\lambda_\varepsilon}(\tilde{\Gamma}_\varepsilon)\right] + \mathbb{E}\left[\mathcal{E}_{\lambda_\varepsilon}\left(\tilde{\Gamma}_\varepsilon\right)\right] \ ,$$

and get the desired bound on the excess risk.

3.7.4 Rate of convergence for kNN estimator

In this section , we focus on rates of convergence of kNN for the estimation of the regression function f^* and the conditional variance function σ^2. The proofs techniques are largely inspired by those in [9, 36], though we provide some additional steps to build for instance finite sample bounds for the sup norm in the problem of conditional variance estimation.

Regression function estimation

We provide the rate of convergence of the kNN estimator of f^* in the regression model for which we make the following assumptions. We assume that f^* is Lipschitz (Assumption 9) and that Assumption 10 are fulfilled. We recall that from Assumption 10, we have that $\mathbb{P}_X$ is supported on a compact set $\mathcal{C}$. Furthermore, we also assume that $Y - f^*(X)$ satisfies a uniform noise condition: there exists $c_0 > 0$ such that

$$\sup_{x\in\mathcal{C}} \mathbb{E}\left[\exp(\lambda\left(Y - f^*(X)\right)) \mid X = x\right] \leq \exp(c_0^2\lambda^2), \quad \text{for} \quad |\lambda| \leq \frac{1}{c_0} \ . \tag{3.19}$$

This assumption is rather weak and requires that conditional on X is sub-exponential uniformly over $\mathcal{C}$ (see [92]). Using the same notation as in Section 3.3, we recall that the kNN estimator $\hat{f}$ of f is defined as follows

$$\hat{f}(x) = \frac{1}{k_n}\sum_{i=1}^{k_n} Y_{(i,n)}(x) \ .$$

The purpose of the appendix is to provide rates of convergence for the kNN estimator $\hat{f}$ under the above assumption. To this end we require two auxiliary lemmata, which provide a control respectively with high probability and in expectation on the distance between a feature point and its neighbors uniformly over $\mathcal{C}$.

Lemma 10. *Assume Assumptions 9-10 hold. Then there exist $C_1 > 0$, which depends only on $\mathcal{C}$, $\mu_{\min}$, and on d and $C_2 > 0$, which depends on $\mathcal{C}$ and on d, such that for all $t \geq \left(\frac{\log(n)k_n}{C_1 n}\right)^{1/d}$, we have*

$$\mathbb{P}\left(\sup_{x\in\mathcal{C}} \frac{1}{k_n}\sum_{i=1}^{k_n} \left\|X_{(i,n)}(x) - x\right\| \geq t\right) \leq C_2 \exp\left(\log(n) - C_1 t^d n/k_n\right) \ .$$

Proof. For any $a \in \mathbb{R}$, let us denote by $\lfloor a \rfloor$ the largest integer which is smaller or equal to a. Consider some $x \in \mathcal{C}$. Following the same arguments as in proof of Theorem 6.2 in [36], we split the data $X_1, \ldots, X_n$ into $k_n + 1$ folds such that the first k_n folds have the same size $\lfloor \frac{n}{k_n} \rfloor$ and the last fold contains the remaining data if there are. We denote $\tilde{X}_j^x$ the nearest neighbor of x in the jth fold and then obviously

$$\sum_{i=1}^{k_n} \left\| X_{(i,n)}(x) - x \right\| \leq \sum_{j=1}^{k_n} \left\| \tilde{X}_j^x - x \right\| \ .$$

Let $\bar{B}_2(a, r)$ be the closed Euclidean ball in $\mathbb{R}^d$ centered in a with radius $r > 0$. Since $\mathcal{C}$ is compact, we have $\mathcal{C} \subset \bar{B}_2(0, R)$ for some $R > 0$, and therefore there exists an ε-net $\mathcal{C}_\varepsilon$ of $\mathcal{C}$ *w.r.t.* $\|.\|$ such that $|\mathcal{C}_\varepsilon| \leq \left(\frac{3R}{\varepsilon}\right)^d$. In particular, for all $x \in \mathcal{C}$ there exists $x_\varepsilon \in \mathcal{C}_\varepsilon$ such that $\|x - x_\varepsilon\| \leq \varepsilon$. Then, for all $x \in \mathcal{C}$ and all $j \in \{1, \ldots, k_n\}$, there exists $x_\varepsilon \in \mathcal{C}_\varepsilon$ such that

$$\left\| \tilde{X}_j^x - x \right\| \leq \left\| \tilde{X}_j^x - x_\varepsilon \right\| + \varepsilon \ . \tag{3.20}$$

Besides, we observe that

$$\left\| \tilde{X}_j^x - x_\varepsilon \right\| \leq \left\| \tilde{X}_j^{x_\varepsilon} - x_\varepsilon \right\| + 2\varepsilon \ . \tag{3.21}$$

Indeed, if $\left\| \tilde{X}_j^{x_\varepsilon} - x_\varepsilon \right\| + 2\varepsilon < \left\| \tilde{X}_j^x - x_\varepsilon \right\|$ we can write

$$\begin{aligned} \left\| \tilde{X}_j^{x_\varepsilon} - x \right\| + \varepsilon &\leq \left\| \tilde{X}_j^{x_\varepsilon} - x_\varepsilon \right\| + 2\varepsilon \\ &< \left\| \tilde{X}_j^x - x_\varepsilon \right\| \leq \left\| \tilde{X}_j^x - x \right\| + \varepsilon \ , \end{aligned}$$

which contradicts the fact that $\tilde{X}_j^x$ is the nearest neighbor of x in the jth fold. Hence, from Equations (3.20) and (3.21), we deduce that

$$\sup_{x \in \mathcal{C}} \frac{1}{k_n} \sum_{i=1}^{k_n} \left\| X_{(i,n)}(x) - x \right\| \leq 3\varepsilon + \sup_{x \in \mathcal{C}_\varepsilon} \frac{1}{k_n} \sum_{j=1}^{k_n} \left\| \tilde{X}_j^x - x \right\| \ .$$

From the above inequality, we obtain that for $t > 6\varepsilon$,

$$\mathbb{P}\left(\sup_{x \in \mathcal{C}} \frac{1}{k_n} \sum_{i=1}^{k_n} \left\| X_{(i,n)}(x) - x \right\| \geq t \right) \leq \mathbb{P}\left(\sup_{x \in \mathcal{C}_\varepsilon} \frac{1}{k_n} \sum_{j=1}^{k_n} \left\| \tilde{X}_j^x - x \right\| \geq t/2 \right) \ . \tag{3.22}$$

Our goal becomes to bound r.h.s. of the above inequality. Using union bound, we deduce that for all $t > 6\varepsilon$

$$\begin{aligned} \mathbb{P}\left(\sup_{x \in \mathcal{C}_\varepsilon} \frac{1}{k_n} \sum_{j=1}^{k_n} \left\| \tilde{X}_j^x - x \right\| \geq t/2 \right) &\leq \sum_{x \in \mathcal{C}_\varepsilon} \mathbb{P}\left(\frac{1}{k_n} \sum_{j=1}^{k_n} \left\| \tilde{X}_j^x - x \right\| \geq t/2 \right) \\ &\leq \sum_{x \in \mathcal{C}_\varepsilon} \sum_{j=1}^{k_n} \mathbb{P}\left(\left\| \tilde{X}_j^x - x \right\| \geq t/2 \right) \ . \end{aligned} \tag{3.23}$$

For each $x \in \mathcal{C}_\varepsilon$ and $j \in \{1, \ldots, k_n\}$, by definition of $\tilde{X}_j^x$ and since $(X_i)_{i=1,\ldots,n}$ are i.i.d., we have

$$\mathbb{P}\left(\left\| \tilde{X}_j^x - x \right\| \geq t/2 \right) = \mathbb{P}\left(\left\| X_{(1, \lfloor \frac{n}{k_n} \rfloor)} - x \right\| \geq t/2 \right) = \left(\mathbb{P}\left(\|X_1 - x\| \geq t/2 \right) \right)^{\lfloor n/k_n \rfloor} \ . \tag{3.24}$$

On one hand, observe that for $t \geq 4R$, $\left(\mathbb{P}\left(\|X_1 - x\| \geq t/2 \right) = 0$. On the other hand for $t \leq 4R$, using the elementary inequality $\log(1 - a) \leq -a$ for all $a \in [0, 1)$, we have that

$$\left(\mathbb{P}\left(\|X_1 - x\| \geq t/2 \right) \right)^{\lfloor n/k_n \rfloor} \leq \exp\left(- \left\lfloor \frac{n}{k_n} \right\rfloor \mathbb{P}\left(\|X_1 - x\| \leq t/2 \right) \right) \ ,$$

which yields, thanks to Assumption 10, there exists $C > 0$ which depends on $\mu_{\min}$ and d such that

$$\left(\mathbb{P}\left(\|X_1 - x\| \geq t/2\right)\right)^{\lfloor n/k_n \rfloor} \leq \exp\left(-Ct^d \lfloor n/k_n \rfloor\right) \quad .$$

We finally deduce from Equation (3.22), (3.23), and (3.24), that for all $t \geq 6\varepsilon$

$$\mathbb{P}\left(\sup_{x\in\mathcal{C}} \frac{1}{k_n}\sum_{i=1}^{k_n} \left\|X_{(i,n)}(x) - x\right\| \geq t\right) \leq k_n |\mathcal{C}_\varepsilon| \exp\left(-Ct^d \lfloor n/k_n \rfloor\right) \quad .$$

Choosing $\varepsilon = (\frac{k_n}{6^d C n})^{1/d}$, we get that for $t \geq (\frac{k_n}{Cn})^{1/d}$,

$$\mathbb{P}\left(\sup_{x\in\mathcal{C}} \frac{1}{k_n}\sum_{i=1}^{k_n} \left\|X_{(i,n)}(x) - x\right\| \geq t\right) \leq C_2 \exp\left(\log(n) - C_1 t^d n/k_n\right),$$

which yields the expected result. □

The second lemma establishes a control in expectation of the uniform distance.

Lemma 11. *Under Assumption 10, there exist $C > 0$, which depends only on $\mathcal{C}$, $\mu_{\min}$, and on d such that*

$$\mathbb{E}\left[\left(\sup_{x\in\mathcal{C}} \frac{1}{k_n}\sum_{i=1}^{k_n} \left\|X_{(i,n)}(x) - x\right\|\right)^p\right] \leq C\left(\frac{k_n \log(n)}{n}\right)^{p/d} \quad .$$

Proof. Since Assumption 10 holds, we can use Lemma 10. Then there exist two non negative constants C_1 and C_2 such that for all $t \geq \left(\frac{\log(n)k_n}{C_1 n}\right)^{1/d}$, we have

$$\mathbb{P}\left(\sup_{x\in\mathcal{C}} \frac{1}{k_n}\sum_{i=1}^{k_n} \left\|X_{(i,n)}(x) - x\right\| \geq t\right) \leq C_2 \exp\left(\log(n) - C_1 t^d n/k_n\right) \quad .$$

Therefore an application of Lemma 14 implies directly the result. □

Below, we state the main result of this section related to the rate of convergence in sup norm of the kNN estimator of the regression function.

Theorem 13. *Assume Assumption 10 is satisfied. Moreover, let $p \geq 1$ and $k_n \propto n^{2/(d+2)}$. Then*

$$\mathbb{E}\left[\left(\sup_{x\in\mathcal{C}} \left|\hat{f}(x) - f^*(x)\right|\right)^p\right] \leq C \log(n)^p n^{-p/(d+2)},$$

where $C > 0$ is a constant which depends on f^*, c_0, $\mathcal{C}$, $\mu_{\min}$ and d.

Proof. First, we have that

$$\hat{f}(x) - f^*(x) = \frac{1}{k_n}\sum_{i=1}^{k_n} \left(Y_{(i,n)}(x) - f^*(X_{(i,n)}(x))\right) + \frac{1}{k_n}\sum_{i=1}^{k_n} \left(f^*(X_{(i,n)}(x)) - f^*(x)\right) \quad .$$

Therefore, since f^* is L-Lipschitz, we then deduce that

$$\sup_{x\in\mathcal{C}} \left|\hat{f}(x) - f^*(x)\right| \leq \sup_{x\in\mathcal{C}} \left|\frac{1}{k_n}\sum_{i=1}^{k_n} Y_{(i,n)}(x) - f^*(X_{(i,n)}(x))\right| + L \sup_{x\in\mathcal{C}} \frac{1}{k_n}\sum_{i=1}^{k_n} \left\|X_{(i,n)}(x) - x\right\| \quad ,$$

which implies that

$$\begin{aligned}\mathbb{E}\left[\left(\sup_{x\in\mathcal{C}} \left|\hat{f}(x) - f^*(x)\right|\right)^p\right] \leq 2^{p-1}\mathbb{E}\left[\left(\sup_{x\in\mathcal{C}} \left|\frac{1}{k_n}\sum_{i=1}^{k_n} Y_{(i,n)}(x) - f^*(X_{(i,n)}(x))\right|\right)^p\right] \\ + 2^{p-1}L^p\mathbb{E}\left[\left(\sup_{x\in\mathcal{C}} \frac{1}{k_n}\sum_{i=1}^{k_n} \left\|X_{(i,n)}(x) - x\right\|\right)^p\right] \quad . \quad (3.25)\end{aligned}$$

Lemma 11 provides a bound on the second term in the r.h.s. of the above inequality. Then it remains to study the first term in the r.h.s. of Eq. (3.25). Let $x \in \mathcal{C}$, and denote by $\mathcal{N}_{k_n}(x) = \{X_{(1,n)}(x), \dots X_{(k_n,n)}(x)\}$ the set of the k_n-nearest neighbors of x among $\{X_1, \dots, X_n\}$. We denote by $\mathcal{B}$ the set of all closed balls in $\mathbb{R}^d$. We observe that there exists $\rho_x > 0$ such that $\mathcal{N}_{k_n}(x) \subset \{\bar{B}(x, \rho_x) \cap \{X_1, \dots, X_n\}\}$, where $\bar{B}(x, \rho_x)$ is the closed ball centered on x with radius ρ_x. Therefore

$$\{\mathcal{N}_{k_n}(x), \quad x \in \mathcal{C}\} \subset \{\{X_1, \dots, X_n\} \cap B, \quad B \in \mathcal{B}\} \ .$$

Besides, since the VC-dimension of the class of balls in $\mathbb{R}^d$ is upper bounded by $d+2$ (see for instance Corollary 13.2 in [26]), Sauer Lemma implies that

$$|\{\{X_1, \dots, X_n\} \cap B, \quad B \in \mathcal{B}\}| \leq \mathcal{S}(\mathcal{B}, n) \leq (n+1)^{d+2} \ ,$$

where $\mathcal{S}(\mathcal{B}, n)$ denotes the shatter coefficient of $\mathcal{B}$ by n points from $\mathcal{C}$. We then deduce that $|\{\mathcal{N}_{k_n}(x), \quad x \in \mathcal{C}\}| \leq (n+1)^{d+2}$, which implies in turn that there exists $\{x_1, \dots, x_J\}$, with $J \leq (n+1)^{d+2}$ such that

$$\mathbb{E}\left[\left(\sup_{x\in\mathcal{C}}\left|\frac{1}{k_n}\sum_{i=1}^{k_n} Y_{(i,n)}(x) - f^*(X_{(i,n)}(x))\right|\right)^p\right] \leq \mathbb{E}\left[\left(\max_{j\in\{1,\dots,J\}}\left|\frac{1}{k_n}\sum_{i=1}^{k_n} Y_{(i,n)}(x_j) - f^*(X_{(i,n)}(x_j))\right|\right)^p\right].$$

Notice that conditional on $X_1, \dots, X_n$ the random variables $(Y_{(i,n)}(x_j) - f^*(X_{(i,n)}(x_j))_{i=1,\dots,k_n}$ are independent with zero mean (see Proposition 8.1 in [9]). Besides from Equation (3.19) they are uniformly sub-exponential over $\mathcal{C}$, then we deduce from the Bernstein Inequality (see [92]) that for all $t \geq 0$ and $j = 1, \dots, J$,

$$\mathbb{P}\left(\left|\frac{1}{k_n}\sum_{i=1}^{k_n} Y_{(i,n)}(x_j) - f^*(X_{(i,n)}(x_j))\right| \geq t\right) \leq \exp\left(-ck_n \min\left(\frac{t^2}{K^2}, \frac{t}{K}\right)\right) \ ,$$

where $c > 0$ is an absolute constant and $K > 0$ depends on c_0 in Eq. (3.19). Set $v_n = \sqrt{\frac{(d+2)\log(n+1)}{ck_n}}$. Our choice of k_n ensures that $v_n \leq 1$, and then we deduce from the union bound that for $t \in (Kv_n, K)$,

$$\mathbb{P}\left(\max_{j\in\{1,\dots,J\}}\left|\frac{1}{k_n}\sum_{i=1}^{k_n} Y_{(i,n)}(x_j) - f^*(X_{(i,n)}(x_j))\right| \geq t\right) \leq \exp\left((d+2)\log(n+1) - ck_n t^2/K^2\right) \ ,$$

and for $t > K$,

$$\mathbb{P}\left(\max_{j\in\{1,\dots,J\}}\left|\frac{1}{k_n}\sum_{i=1}^{k_n} Y_{(i,n)}(x_j) - f^*(X_{(i,n)}(x_j))\right| \geq t\right) \leq \exp\left((d+2)\log(n+1) - ck_n t/K\right) \ .$$

Considering these two cases, we can derive an exponential bound on the term

$$\mathbb{P}\left(\max_{j\in\{1,\dots,J\}}\left|\frac{1}{k_n}\sum_{i=1}^{k_n} Y_{(i,n)}(x_j) - f^*(X_{(i,n)}(x_j))\right| \geq t\right)$$

for all $t \geq Kv_n$, therefore we can use similar arguments as in Lemma 14 and conclude that

$$\mathbb{E}\left[\left(\sup_{x\in\mathcal{C}}\left|\frac{1}{k_n}\sum_{i=1}^{k_n} Y_{(i,n)}(x) - f^*(X_{(i,n)}(x))\right|\right)^p\right] \leq \mathbb{E}\left[\left(\max_{j\in\{1,\dots,J\}}\left|\frac{1}{k_n}\sum_{i=1}^{k_n} Y_{(i,n)}(x_j) - f^*(X_{(i,n)}(x_j))\right|\right)^p\right] \leq C\left(\frac{\log(n)}{k_n}\right)^{p/2} \ . \quad (3.26)$$

Combining the above inequality, Equation (3.25), and Lemma 11, gives the desired result. □

To conclude this section, we also provide the rate of convergence of the kNN estimator in L_2-norm

Theorem 14. *Assume Assumption 10 is satisfied and let $k_n \propto n^{2/(d+2)}$, then*

$$\mathbb{E}\left[\left(\hat{f}(X) - f^*(X)\right)^2\right] \leq Cn^{-2/(d+2)} \ ,$$

where $C > 0$ is a constant which depends on f^*, c_0, $\mathcal{C}$, and d.

The proof of this result is provided in [36] for $d \geq 3$ (see Theorem 6.2). However, a small change implies that the same proof holds for all d under Assumption 10.

Conditional variance function estimation

We provide the rate of convergence of the kNN estimator of σ^2. This proof is largely inspired by [9], though we are interested here in finite sample bounds.

Proposition 9. *Grant Assumptions 9 and 10. Let $k_n \propto n^{2/(d+2)}$, the following holds*

$$\mathbb{E}\left[\left(\sup_{x\in\mathcal{C}}\left|\hat{\sigma}^2(x) - \sigma^2(x)\right|\right)^{1+\alpha}\right] \leq C\log(n)^{(\alpha+1)}n^{-(\alpha+1)/(d+2)} \ ,$$

for all $\alpha \geq 0$, where $C > 0$ is a constant which depends on f^*, σ^2, c_0, $\mathcal{C}$, and on the dimension d.

Proof. First, we define the function $\tilde{\sigma}^2$ by

$$\tilde{\sigma}^2(x) = \frac{1}{k_n}\sum_{i=1}^{k_n}\left(Y_{(i,n)}(x) - f^*(X_{(i,n)}(x))\right)^2, \quad \forall x \in \mathbb{R}^d \ .$$

The function $\tilde{\sigma}^2$ is the pseudo-estimator of σ^2 that would be used in the case where the function f^* is known. By the triangle inequality, we have that for all $x \in \mathcal{C}$,

$$\left|\hat{\sigma}^2(x) - \sigma^2(x)\right| \leq \left|\hat{\sigma}^2(x) - \tilde{\sigma}^2(x)\right| + \left|\tilde{\sigma}^2(x) - \sigma^2(x)\right| \ .$$

Now, we observe that

$$\hat{\sigma}^2(x) - \tilde{\sigma}^2(x) =$$
$$\frac{1}{k_n}\sum_{i=1}^{k_n}\left(f^*(X_{(i,n)}(x)) - \hat{f}(X_{(i,n)}(x))\right)\left(2(Y_{(i,n)}(x) - f^*(X_{(i,n)}(x))) + f^*(X_{(i,n)}(x)) - \hat{f}(X_{(i,n)}(x))\right) \ .$$

Therefore, we deduce

$$\sup_{x\in\mathcal{C}}\left|\hat{\sigma}^2(x) - \sigma^2(x)\right| \leq \sup_{x\in\mathcal{C}}\left|\tilde{\sigma}^2(x) - \sigma^2(x)\right| + \left(\sup_{x\in\mathcal{C}}\left|\hat{f}(x) - f^*(x)\right|\right)^2 +$$
$$2\sup_{x\in\mathcal{C}}\left|\hat{f}(x) - f^*(x)\right|\sup_{x\in\mathcal{C}}\left|\frac{1}{k_n}\sum_{i=1}^{k_n}(Y_{(i,n)}(x) - f^*(X_{(i,n)}(x))\right| \ .$$

From the above inequality, using the fact that $(a+b+c)^p \leq 3^{p-1}(a^p + b^p + c^p)$ for $p \geq 1$, a, b, $c \in \mathbb{R}$ and applying the Cauchy-Schwartz Inequality, we obtain

$$\mathbb{E}\left[\left(\sup_{x\in\mathcal{C}}\left|\hat{\sigma}^2(x) - \sigma^2(x)\right|\right)^{1+\alpha}\right] \leq$$
$$C_1\mathbb{E}\left[\left(\sup_{x\in\mathcal{C}}\left|\tilde{\sigma}^2(x) - \sigma^2(x)\right|\right)^{1+\alpha}\right] + C_2\mathbb{E}\left[\left(\sup_{x\in\mathcal{C}}\left|\hat{f}(x) - f^*(x)\right|\right)^{2(1+\alpha)}\right]$$
$$+C_3\left\{\mathbb{E}\left[\left(\sup_{x\in\mathcal{C}}\left|\hat{f}(x) - f^*(x)\right|\right)^{2(1+\alpha)}\right]\right\}^{1/2}\left\{\mathbb{E}\left[\left(\sup_{x\in\mathcal{C}}\left|\frac{1}{k_n}\sum_{i=1}^{k_n}(Y_{(i,n)}(x) - f^*(X_{(i,n)}(x))\right|\right)^{2(1+\alpha)}\right]\right\}^{1/2} \ ,$$

where C_1, C_2 and C_3 are non negative reals. We finish the proof of the proposition by bounded the above l.h.s. This relies on controls of estimation error of kNN for the regression function f^* and the conditional variance function σ^2. Observe that when Y is either bounded or satisfies the model conditions in Eq. (3.5), we have that the random variables $Y - f^*(X)$ and $(Y - f^*(X))^2 - \sigma^2(X)$ satisfy the uniform noise condition (3.19). Indeed, while this fact is clear for $Y - f^*(X)$, it also holds true for $(Y - f^*(X))^2 - \sigma^2(X)$ since, conditionally on X, this random variable is either bounded (since σ^2 is bounded as well) or sub-exponential. Therefore, the result of Theorem 13 applies for the kNN estimators $\tilde{\sigma}^2$ and $\hat{f}$. Furthermore, using the result in Eq. (3.26), we deduce from the above inequality

$$\mathbb{E}\left[\left(\sup_{x\in\mathcal{C}}|\hat{\sigma}^2(x)-\sigma^2(x)|\right)^{1+\alpha}\right] \leq C\log(n)^{(\alpha+1)}n^{-(\alpha+1)/(d+2)} \ ,$$

where $C > 0$ is a constant which depends on f^*, σ^2, c_0, $\mathcal{C}$, and the dimension d. □

3.7.5 Technical tools

In this section, we state several results that may help for readability of the chapter. The first result is a direct application of the classical peeling argument of [2].

Lemma 12 (Lemma 1 in [23]). *Let X be a real random variable, $(X_n)_{n\geq 1}$ be a sequence of real random variables and $t_0 \in \mathbb{R}$. Assume that there exist $C_1 > 0$ and $\gamma_0 > 0$ such that*

$$\mathbb{P}_X\left(|X - t_0| \leq \delta\right) \leq C_1\delta^{\gamma_0}, \quad \forall\delta > 0 \ ,$$

and a sequence of positive numbers a_n tends towards infinity, C_2, C_3 some positive constants such that

$$\mathbb{P}_{X_n}\left(|X_n - X| \geq \delta|X\right) \leq C_2\exp\left(-C_3a_n\delta^2\right), \quad \forall\delta > 0, \quad \forall n \in \mathbb{N}.$$

Then, there exists $C > 0$ depending only on C_1, C_2 and C_3, such that

$$|\mathbb{E}\left[\mathbb{1}_{X_n\geq t_0} - \mathbb{1}_{X\geq t_0}\right]| \leq Ca_n^{-\gamma_0/2}.$$

The next result describes the representation of ∞-Wasserstein distance (W_∞) on the real line. Let $Z_\infty(\mathbb{R})$ be the collection of all compactly supported probability measures on $\mathbb{R}$.

Lemma 13 (Theorem 2.12 in [11]). *Let μ and ν be probability measures in $Z_\infty(\mathbb{R})$ with respective distribution functions F and G. Then, $W_\infty(\mu,\nu) := \sup_{0<t<1}|F^{-1}(t) - G^{-1}(t)|$ is the infimum over all $h \geq 0$ such that*

$$G(x - h) \leq F(x) \leq G(x + h) \ \text{ for all } \ x \in \mathbb{R}.$$

The following result provides a bound on moments of a positive random variable provided a tail control.

Lemma 14. *Let $a \geq 1$, let b, c be two non negative real numbers, and let $m \in \mathbb{N}$. Consider Z a positive random variable such that*

$$\mathbb{P}\left(Z \geq t\right) \leq c\exp\left(a - bt^m\right) \ ,$$

for all $t \geq (a/b)^{1/m}$. Then for all $p \geq 1$, there exists a constant $C > 0$ such that

$$\mathbb{E}\left[Z^p\right] \leq C(a/b)^{p/m} \ .$$

Proof. Using the following equality which holds for any positive random variable Z, and any $p \geq 1$

$$\mathbb{E}\left[Z^p\right] = \int_0^{+\infty}\mathbb{P}\left(Z \geq t\right)pt^{p-1}\mathrm{d}t \ , \tag{3.27}$$

and the condition in Lemma 6, we deduce

$$\mathbb{E}\left[Z^p\right] \leq \int_0^u pt^{p-1}\mathrm{d}t + c\int_u^{+\infty}\exp\left(a - bt^m\right)pt^{p-1}\mathrm{d}t \ , \tag{3.28}$$

where $u = (a/b)^{1/m}$ and where we used the trivial inequality $\mathbb{P}\left(Z \geq t\right) \leq 1$ to bound the first term in the r.h.s. Since $(a')^m - (b')^m \geq (a' - b')^m$ for all a', $b' \in \mathbb{R}$ such that $a' \geq b' \geq 0$, we can write that

$$\exp\left(a - bt^m\right) \leq \exp\left(-(t-u)^m b\right) \quad ,$$

which yields

$$\begin{aligned}
\int_u^{+\infty} \exp\left(a - bt^m\right) pt^{p-1} \mathrm{d}t &\leq \int_u^{+\infty} \exp\left(-(t-u)^m b\right) pt^{p-1} \mathrm{d}t \\
&\leq \frac{1}{u} \int_u^{+\infty} \exp\left(-(t-u)^m b\right) pt^{p} \mathrm{d}t \\
&= \frac{p}{u} \left(\frac{1}{b}\right)^{1/m} \int_0^{+\infty} e^{-v^m} \left(v \left(\frac{1}{b}\right)^{1/m} + u\right)^p \mathrm{d}v \quad ,
\end{aligned}$$

where we consider the changing of variable $v = ((t-u)^m b)^{1/m}$ in the last equality. Finally, using that $(a' + b')^p \leq 2^{p-1}((a')^p + (b')^p)$ for all $p \geq 1$, a', $b' \in \mathbb{R}$ and given that $u \geq (1/b)^{1/m}$, we show from the above inequality that

$$\begin{aligned}
\int_u^{+\infty} \exp\left(a - bt^m\right) pt^{p-1} \mathrm{d}t &\leq C_1 \left(\frac{1}{b}\right)^{p/m} \int_0^{+\infty} v^p e^{-v^m} \mathrm{d}v + C_2 u^p \int_0^{+\infty} e^{-v^m} \mathrm{d}v \\
&\leq C_3 u^p \quad ,
\end{aligned}$$

for positive constants C_1, C_2, C_3. Inject this into Eq.(3.28) leads to the result. □

Chapter 4

Prediction interval with fixed expected length in the Gaussian heteroscedastic regression

Abstract : We tackle the problem of building a prediction interval in heteroscedastic Gaussian regression. We focus on prediction intervals with constrained expected length in order to guarantee interpretability of the output. In this framework, we derive a closed form expression of the optimal prediction interval that allows for the development a data-driven prediction interval based on plug-in. The construction of the proposed algorithm is based on two samples, one labeled and another unlabeled. Under mild conditions, we show that our procedure is asymptotically as good as the optimal prediction interval both in terms of expected length and error rate. In particular, the control of the expected length is distribution-free. We also derive rates of convergence under smoothness and the Tsybakov noise conditions. We conduct a numerical analysis that exhibits the good performance of our method. It also indicates that even with a few amount of unlabeled data, our method is very effective in enforcing the length constraint.

Contents

4.1 Introduction

Prediction is one of the main goals in supervised learning, it consists in building, given historical data, a candidate output for a new observation. One common practice thereafter is to carry out inference on the output and then to ask for confidence in the predicted value, therefore, *prediction interval (PI)* appears as appropriate tools to handle this problem in the regression setting. A typical application is the prediction in the linear regression case when the data are assumed Gaussian with common variance. In this context, the notion of PI is well studied and well understood both from practice and theory.

However, in the general case, inference as a post-processing step may produce irrelevant conclusions due to the stochastic nature of the data-driven prediction procedure (see for instance [7]). Therefore, in order to guarantee the theoretical validity of the prediction intervals, it is suitable to process at once both aspects of the problem, that is, one might design a data-driven procedure directly devoted to the *prediction interval* purpose.

In a classical setting of PI, one often asks for a pre-specified level of confidence for the predicted range of values (says 95% or 99% according to the problem). This is for instance the approach that is considered in the *conformal prediction* literature [95, 94, 52, 50]. However, this strategy may suffer from interpretability issues for problems where prediction task is difficult or when classical assumptions on the noise are not satisfied. Specifically, for relatively restrictive values of the confidence level, the resulting output might be so large that it becomes useless.

In contrast, our purpose is to produce for future observation a prediction interval with a predetermined expected length. This framework is completely different from the previous one since it does not ensure any coverage guarantee but rather ensures the interpretability of the predicted output. Indeed, since the length of the output interval is controlled, we do not expect for a given input instance $\mathbf{x} \in \mathbb{R}^d$ a too large set of candidate values.

Generally speaking, the range of values that we would output with PI has no reason to be an interval. However, in a Gaussian model, this range of values indeed forms an interval (or a union of it). In this chapter, we investigate the problem of PI under expected length constraint in the Gaussian regression setup. We aim at providing a general device that outputs a PI for a new feature. Our procedure relies on the plug-in principle and we propose in the present contribution a statistical analysis of it in this setting.

Main contributions. Denote by $\Gamma : \mathbb{R}^d \to \mathcal{P}(\mathbb{R})$ a given prediction set, where $\mathcal{P}(\mathbb{R})$ is the set of subsets of $\mathbb{R}$. One of the main inputs of the present work is the introduction of a novel framework for PI in the regression setting, taking sides of controlling the expected size $\mathbb{E}\left[L(\Gamma(\mathbf{X}))\right]$ of the output predictor Γ while minimizing its error rate $\mathbb{P}\left(Y \notin \Gamma(\mathbf{X})\right)$, where $L(\Gamma(\mathbf{X})) = \int_{\mathbb{R}} \mathbb{1}_{y \in \Gamma(\mathbf{X})} dy$ stands for the Lebesgue measure of Γ. We derive the optimal rule for this problem which is defined as

$$\Gamma_\ell^* \in \underset{\Gamma : \mathbb{E}[L(\Gamma(\mathbf{X}))] \leq \ell}{\operatorname{argmin}} \mathbb{P}\left(Y \notin \Gamma(\mathbf{X})\right) \quad ,$$

where $\ell > 0$ is a preset length chosen by the practitioner.

In the Gaussian framework, based on the plug-in principle, we then build a general procedure that estimate the optimum and prove that the resulting empirical predictor performs as well as Γ_ℓ^* both in term of expected length and error rate. Notably, the control on the expected length of the proposed estimator is distribution-free. Furthermore, our algorithm has two appealing properties. It can benefit from a semi-supervised setting and can be applied to any off-the-shelf machine learning algorithm.

On the other hand, we evaluate the performance of our estimator with respect to the symmetric difference distance and a risk measure which properly balances the expected length and the error rate. Specifically, we establish the consistency for our procedure under mild assumptions and provide rates of convergence under suitable assumptions on the distribution of the data.

We additionally conduct a numerical study that confirms our theoretical findings and shows how effective our method is in controlling the length, an important aspect to ensure the interpretability of the output. Finally, we provide a numerical comparison with the strategy which consists in building

PI under expected coverage constraint. Our numerical experiment highlights that our proposed approach produced significantly more stable PI. In particular, our algorithm seems to be more adapted when the sample size of the training sample is moderate.

Related works. A first line of work related to PI is *confidence intervals*. This is one of the most popular tools in statistical inference and differs from PI by the fact that the purpose there is to output a range of values for a given parameter of the model such that the mean, while our goal is here the prediction. The spectrum of applications of confidence interval is extremely wide and from some perspective PI can be seen as part of the confidence interval literature where we focus on building a confidence interval for the output of a new observation.

Probably the closest direction of works to ours is *conformal prediction* [95, 52, 50]. The main difference relies on the way the expected length and the error rate of the prediction interval is considered. The goal there is to produce a PI with a pre-specified level of accuracy. The connection of PI with controlled expected size is important to figure out since, *at the population level*, each PI with controlled accuracy corresponds to a PI with controlled expected size. From practice however, the two approaches start to differ. We defer this discussion to Sections 4.4.2 and 4.5.2 where a complete comparison to *PI with controlled accuracy* is conducted.

Providing an output with a pre-defined length has rarely been considered. Probably the first reference that deals with such notion is [45]. There, the authors build confidence intervals for the mean and variance in a Gaussian problem that reach given confidence level while being of size L. In contrast to that work, we deal with prediction intervals, our control on the length is in expectation which offers more flexibility on "hard" points, we do not focus on a pre-specified level of confidence but rather minimize the error under a size constraint, and we derive a statistical and a numerical analysis of our method.

Finally, let us notify that constraining the expected length is not novel. It has already been considered in the multi-class classification setting [22, 20]. There, the control of the length is interpreted as the desired average number of output labels. Similar to the present work, the goal is to focus on a set of values for prediction while maintaining the interpretability of the output. The main difference with earlier work is that we deal here with real valued output which is more tricky. From this perspective the present chapter is a generalization of these previous works to the Gaussian regression setting.

Outline of the chapter. Section 4.2 provides the main notation and describes the framework of prediction intervals under expected length constraint in the Gaussian regression. In particular, the explicit form of the optimal rule is provided. Section 4.3 introduces our data-driven procedure as well as its statistical analysis. This theoretical analysis is complemented with a numerical study presented in Section 4.5. Additional considerations beyond the Gaussian assumption and other frameworks of prediction intervals are considered in Section 4.4. A conclusion is provided in Section 4.6, while the proofs of our results are postponed to the appendix.

4.2 General framework

In the present contribution we focus on the Gaussian model, that is, we assume that $(\mathbf{X}, Y) \in \mathbb{R}^d \times \mathbb{R}$ are such that

$$Y = f^*(\mathbf{X}) + \sigma(\mathbf{X})\,\varepsilon \ , \tag{4.1}$$

where $\varepsilon \sim \mathcal{N}(0,1)$ is independent of $\mathbf{X}$. In this expression, $f^* : \mathbb{R}^d \to \mathbb{R}$ is the regression function and $\sigma : \mathbb{R}^d \to \mathbb{R}_+^*$ is the conditional variance function, both of them assumed to be unknown. The main assumptions that we consider throughout the chapter are presented in Section 4.2.1. The characterization of the optimal prediction interval under expected length constraint is provided in Section 4.2.2. Finally, we define the measure of performance dedicated to asses the quality of a prediction interval in Section 4.2.3.

4.2.1 Assumptions

Given an observation $\mathbf{X} \in \mathbb{R}^d$, our goal is to produce the most accurate, in a certain sense to be specified later, a range of predicted values where the corresponding label $Y \in \mathbb{R}$ lies. Such predictions will be describe a set of $\mathcal{P}(\mathbb{R})$ and denoted by $\Gamma(\mathbf{x})$ for each $\mathbf{x} \in \mathbb{R}^d$. In other words, the predictor Γ is a mapping from $\mathbb{R}^d$ onto $\mathcal{P}(\mathbb{R})$.

Throughout the chapter we denote by $p(\cdot|\mathbf{x})$ the conditional density of Y given $\mathbf{x}$, that is, for all $y \in \mathbb{R}$

$$p(y|\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma(\mathbf{x})} \exp\left(-\frac{(y - f^*(\mathbf{x}))^2}{2\sigma^2(\mathbf{x})}\right) \ ,$$

that is, we focus on the heteroscedastic Gaussian regression model. In order to avoid pathological situations, we impose the following mild assumptions on the regression and conditional variance functions.

Assumption 12. *There exist $0 < \sigma_0 < \sigma_1 < \infty$ such that for all $\mathbf{x} \in \mathbb{R}^d$*

$$\sigma_0 \leq \sigma(\mathbf{x}) \leq \sigma_1 \ .$$

Assumption 13. *There exists $C_1 > 0$ such that*

$$\mathbb{E}\left[|f^*(\mathbf{X})|\right] \leq C_1 \ .$$

In addition, we consider an assumption which is PI context-specific. It ensures in particular the existence and uniqueness and the optimal PI. Note that similar assumption is considered in the set-valued classification framework [20].

Assumption 14 (Continuity)**.** *For all $y \in \mathbb{R}$, the mapping $t \mapsto \mathbb{P}_{\mathbf{X}}(p(y|\mathbf{X}) \geq t)$ is continuous on $\mathbb{R}_+^*$.*

In other word, we assume that $p(y|\mathbf{X})$ is atomless.

4.2.2 Prediction interval with expected length

For a given predictor Γ two features are of interest, its error rate $\mathbb{P}(Y \notin \Gamma(\mathbf{X}))$ and its expected Lebesgue measure defined as

$$\mathcal{L}(\Gamma) := \mathbb{E}\left[L(\Gamma(\mathbf{X})\right] = \mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\{y \in \Gamma(\mathbf{X})\}} \mathrm{d}y\right] \ .$$

Given $\ell > 0$, we focus on the following problem

$$\Gamma_\ell^* \in \arg\min\{\mathbb{P}\left(Y \notin \Gamma(\mathbf{X})\right) \ : \ \Gamma \text{ such that } \mathcal{L}(\Gamma) \leq \ell\} \ . \tag{4.2}$$

The next proposition provides the characterization of the optimal predictor under Assumption 14.

Proposition 10. *Let $\ell > 0$, under Assumption 14, the optimal predictor Γ_ℓ^* can be expressed as*

$$\Gamma_\ell^*(\mathbf{X}) = \{y \in \mathbb{R} : \ p(y|\mathbf{X}) \geq \lambda_\ell^*\} \ ,$$

where $\lambda_\ell^* = G^{-1}(\ell)$ with $G(t) := \int_{\mathbb{R}} \mathbb{P}(p(y|\mathbf{X}) \geq t)dy$ for all $t > 0$[1]*.*

The parameter λ_ℓ^*, which corresponds to the value of the generalized inverse function G^{-1} at ℓ, plays a crucial role in our study since it fully determines the optimal predictor Γ^*. This being said, let us comment on Proposition 10. First, an important consequence of the above proposition is that the predictor Γ_ℓ^* is an interval of length ℓ, that is $\mathcal{L}(\Gamma_\ell^*) = \ell$ and we additionally can express Γ_ℓ^* as

$$\Gamma_\ell^*(\mathbf{X}) = \left[f^*(\mathbf{X}) - \sqrt{2\sigma^2(\mathbf{X}) \log\left(\frac{1}{\sqrt{2\pi}\lambda_\ell^*\sigma(\mathbf{X})}\right)} \ , \ f^*(\mathbf{X}) + \sqrt{2\sigma^2(\mathbf{X}) \log\left(\frac{1}{\sqrt{2\pi}\lambda_\ell^*\sigma(\mathbf{X})}\right)}\right] \ .$$

[1]When $t = 0$, we have $G(t) = +\infty$ and then we will use the convention $G^{-1}(+\infty) = 0$.

Second, the function G defined in Proposition 10 is the extension to the regression case of the function G defined in [22] in the multi-class setting. Note that the function G is always well-defined and continuous for $t > 0$, since by Markov Inequality and Fubini Theorem,

$$G(t) = \int_{\mathbb{R}} \mathbb{P}(p(y|\mathbf{X}) \geq t) dy \leq \frac{1}{t} \int_{\mathbb{R}} \mathbb{E}\left[p(y|\mathbf{X})\right] \mathrm{d}y \leq \frac{1}{t} \mathbb{E}\left[\int_{\mathbb{R}} p(y|\mathbf{X}) \mathrm{d}y\right] \leq \frac{1}{t} \ .$$

Finally, we highlight that parameter λ_ℓ^* is simply the Lagrange multiplier of the minimization problem defined by Equation (4.2). Therefore, Γ_ℓ^* can be expressed as the minimizer of the unconstrained problem

$$\Gamma_\ell^* \in \arg\min_{\Gamma} \mathbb{P}\left(Y \notin \Gamma(\mathbf{X})\right) + \lambda_\ell^* \mathbb{E}\left[L(\Gamma(\mathbf{X})\right] \ . \tag{4.3}$$

4.2.3 Measures of performance

In this paragraph we introduce two ways to quantify the quality of a given prediction interval Γ . The first one, suggested by Equation (4.3), balances the error rate and the expected length of the predictor

$$R_\ell(\Gamma) = \mathbb{P}\left(Y \notin \Gamma(\mathbf{X})\right) + \lambda_\ell^* \mathbb{E}\left[L(\Gamma(\mathbf{X})\right] \ ,$$

with $\lambda_\ell^* = G^{-1}(\ell)$. This risk is particularly important from our perspective since minimizing it over all predictors lead to the optimal predictor Γ_ℓ^*, which reaches the requested expected length. A natural "distance" to the optimal predictor is then evaluated through the excess risk

$$\mathcal{E}_\ell\left(\Gamma\right) = R_\ell(\Gamma) - R_\ell(\Gamma_\ell^*) \ .$$

The following proposition provides a closed formula for this term.

Proposition 11. *Let $\ell \geq 0$. For any predictor Γ*

$$\mathcal{E}_\ell(\Gamma) = \mathbb{E}\left[\int_{\Gamma(\mathbf{X}) \triangle \Gamma_\ell^*(\mathbf{X})} |p(y|\mathbf{X}) - \lambda_\ell^*| \ dy\right] \ .$$

Interestingly, the above result shows that the performance of a predictor Γ is directly linked to the behavior of the conditional density $p(y|\mathbf{x})$ around the threshold λ_ℓ^* on the symmetric difference $\{\Gamma(\mathbf{X}) \triangle \Gamma_\ell^*(\mathbf{X})\}$.

A second measure of performance arises naturally when we deal with predictors that are intervals. It is the expectation of symmetric difference between the considered predictor Γ and optimal predictor Γ_ℓ^* defined for all predictor Γ as

$$\mathcal{H}\left(\Gamma\right) = \mathbb{E}\left[L\left(\Gamma(\mathbf{X}) \triangle \Gamma_\ell^*(\mathbf{X})\right)\right] = \mathbb{E}\left[\int_{\Gamma(\mathbf{X}) \triangle \Gamma_\ell^*(\mathbf{X})} dy\right] \ .$$

In some sense, we note that the measure $\mathcal{H}$ provides a stronger guarantee than the excess risk since $\mathcal{E}_\ell(\Gamma) \leq C_2 \mathcal{H}(\Gamma)$ where C_2 is a positive constant which depends on σ_0. Besides, $\mathcal{H}(\Gamma) = 0$ implies that $\Gamma = \Gamma_\ell^*$ while this property does not necessarily hold for the excess risk.

4.3 Data-driven procedure

In this section, we provide a general data-driven procedure to estimate the optimal predictor Γ_ℓ^*. Two key features are expected from the resulting empirical prediction interval. The expected length should be of order ℓ while keeping its error rate close to one obtained by the oracle predictor. The estimation procedure is presented in the Section 4.3.1, and its main properties are provided in Section 4.3.2. Finally, Section 4.3.3 is dedicated to the study of rates of convergence.

4.3.1 Empirical prediction interval

The result provided in Proposition 10 suggests that an empirical prediction interval can be obtained through the plug-in principle by considering estimators of the conditional density p and the parameter $\lambda_\ell^* = G^{-1}(\ell)$. From a theoretical perspective, this learning task requires two independent samples.

First, in order to build an estimator of the conditional density p, we estimate the functions f^* and σ. Hence, we exploit a labeled sample $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ and build based on it estimators $\hat{f}$ and $\tilde{\sigma}$ of these two functions by the means of any machine learning algorithm. However, to establish theoretical guarantees, we require that the estimator $\tilde{\sigma}$ satisfies similar assumption as Assumption 12. To this end, we consider a thresholded version of the estimator $\tilde{\sigma}$ denoted by $\hat{\sigma}$ and define for $s > 0$ as

$$\hat{\sigma}^2(\mathbf{x}) = \tilde{\sigma}^2(\mathbf{x})\mathbb{1}_{\{s^{-1} \le \tilde{\sigma}^2(\mathbf{x}) \le s\}} + s^{-1}\mathbb{1}_{\{\tilde{\sigma}^2(\mathbf{x}) < s^{-1}\}} + s\mathbb{1}_{\{\tilde{\sigma}^2(\mathbf{x}) > s\}} \ .$$

A straightforward consequence of the definition of $\hat{\sigma}$ is that $\frac{1}{s} \le \hat{\sigma}^2(\mathbf{x}) \le s$. Furthermore, if s satisfies $\frac{1}{s} \le \sigma_0^2 \le \sigma_1^2 \le s$, we have for all $\mathbf{x}$

$$\left|\hat{\sigma}^2(\mathbf{x}) - \sigma^2(\mathbf{x})\right| \le \left|\tilde{\sigma}^2(\mathbf{x}) - \sigma^2(\mathbf{x})\right| \ ,$$

Hence consistency of $\tilde{\sigma}^2$ would imply the consistency of $\hat{\sigma}^2$.

Based on $\hat{f}$ and $\hat{\sigma}$, an estimator $\tilde{p}$ of the conditional density p naturally derives and can be written for all $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ as

$$\tilde{p}(y|\mathbf{x}) = \frac{1}{\sqrt{2\pi}\hat{\sigma}(\mathbf{x})} \exp\left(-\frac{\left(y - \hat{f}(\mathbf{x})\right)^2}{2\hat{\sigma}^2(\mathbf{x})}\right) \ .$$

The second step is devoted to the estimation of the parameter λ_ℓ^* and requires an *unlabeled* sample $\mathcal{D}_N = \{\mathbf{X}_{n+1}, \ldots, \mathbf{X}_{n+N}\}$ which consists of i.i.d. observations of $\mathbf{X}$ and is independent of $\mathcal{D}_n$. Since λ_ℓ^* depends on the function G, it is suitable to consider the empirical counterpart of the function G, that we build based on $\hat{p}$ and define for all $t \in [0, 1]$ as

$$\tilde{G}(t) = \int_{\mathbb{R}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{\hat{p}(y|\mathbf{X}_{n+i}) > t\}} \mathrm{d}y \ .$$

As a result, the empirical prediction interval is defined[2] point-wise as

$$\tilde{\Gamma}(\mathbf{x}) = \{y \in \mathbb{R} : \ \tilde{p}(y|\mathbf{x}) \ge \tilde{G}^{-1}(\ell)\} \ .$$

The predictor $\tilde{\Gamma}$ is very natural but has a few limitations: i) because Y is unbounded, the study of the theoretical properties of the estimator $\tilde{\Gamma}$ might be difficult; ii) in addition, establishing a theoretical analysis on $\tilde{\Gamma}$ involves similar assumption to Assumption 14 for $\tilde{G}$. More precisely, it requires that conditional on $\mathcal{D}_n$ the cumulative distribution of $\tilde{p}(y|\mathbf{X})$ is atomless; iii) furthermore, the above expression of $\tilde{\Gamma}(\mathbf{x})$ is explicit but relies on computing an integral in order to evaluate the function $\tilde{G}$. This integral should be approximated. To circumvent all these issues, we consider the following modifications of the initial estimator $\tilde{\Gamma}$.

For i) – Thresholding. Let $s > 0$, we consider a thresholded version of p given by

$$\hat{p}(y|\mathbf{x}) = \frac{1}{\sqrt{2\pi}\hat{\sigma}(\mathbf{x})} \exp\left(-\frac{\left(y - \hat{f}(\mathbf{x})\right)^2}{2\hat{\sigma}^2(\mathbf{x})}\right) \mathbb{1}_{\{|y| \le s\}} \ . \tag{4.4}$$

[2]Here again, we use the convention $\tilde{G}^{-1}(+\infty) = 0$.

For ii) – Randomization. To ensure the continuity of the conditional C.D.F. of $\hat{p}(y|\mathbf{X})$ for $y \in [-s, s]$, we introduce a random perturbation ζ distributed according to a Uniform distribution on $[0, u]$, for $u > 0$ and independent of $(\mathbf{X}, Y)$. We then define the randomized version of $\hat{p}$ as

$$\hat{p}(y|\mathbf{X}, \zeta) = \hat{p}(y|\mathbf{X}) + \zeta \mathbb{1}_{\{|y| \leq s\}} \ . \tag{4.5}$$

For iii) – Discretization. To approximate $\tilde{G}$, we simply consider the Riemann sum based on the regular grid $\mathcal{G} = \{y_1, \ldots, y_M\}$ of $[-s, s]$ for some $M \geq 1$. To this end, we introduce $(\zeta_1, \ldots, \zeta_N)$ i.i.d. copies of ζ and then define

$$\hat{G}(t) = \frac{2s}{MN} \sum_{k=1}^{M} \sum_{i=1}^{N} \mathbb{1}_{\{\hat{p}(y_k|\mathbf{X}_{n+i}, \zeta_i) > t\}} \ .$$

Finally, the resulting empirical prediction interval writes as

$$\hat{\Gamma}(\mathbf{X}, \zeta) = \{y \in \mathbb{R} : \ \hat{p}(y|\mathbf{X}, \zeta) \geq \hat{G}^{-1}(\ell)\} \ . \tag{4.6}$$

4.3.2 Theoretical guarantees

In this section, we provide the main properties of the empirical prediction interval $\hat{\Gamma}$. We first illustrate that the prediction interval $\hat{\Gamma}$ has an expected length equal to the requested value ℓ. This is one of the main striking feature of our data-driven procedure.

Proposition 12. *Assume that $M > 4\sqrt{N}$, then*

$$\mathbb{E}\left[\left|\mathcal{L}(\hat{\Gamma}) - \ell\right|\right] \leq C \frac{s}{\sqrt{N}} \ ,$$

where $C > 0$ is an absolute constant.

The above result states that our methodology is able to produce a prediction interval with an expected length ℓ, irrespectively of the distribution of the data and of whether or not we have build accurate estimates for f^* and σ. Importantly, Proposition 12 holds even if $(\mathbf{X}, Y)$ does not satisfy Equation (4.1). From this perspective the control on the expected length of the produced prediction interval is *distribution-free*. Notice in particular that the stated bound depends only on the parameter s which should be specified by the practitioner (this choice is discussed later) and on the number N of unlabeled data. In some semi-supervised applications, the amount of these data can be very large so that we can get a good approximation of the marginal distribution $\mathbb{P}_{\mathbf{X}}$ and then we can expect a good control of the expected length almost for free. Let us also add that Proposition 12 is a fundamental step to show the following bound on the excess risk:

Proposition 13. *Let Assumption 14 be satisfied. For $M > 4\sqrt{N}$, we have*

$$\mathbb{E}\left[\mathcal{E}_\ell\left(\hat{\Gamma}\right)\right] \leq C \left(\mathbb{E}\left[\int_{\mathbb{R}} |\hat{p}(y|\mathbf{X}) - p(y|\mathbf{X})| \, \mathrm{d}y\right] + su + \frac{s}{\sqrt{N}}\right) \ ,$$

where $C > 0$ is an absolute constant.

The above result shows that the excess-risk of $\hat{\Gamma}$ is mainly controlled by the L_1-risk of the estimator of the conditional density. The residual terms are related to the randomization on the one hand and to the control of the expected length of $\hat{\Gamma}$, given in Proposition 12, on the other hand. Proposition 13 is an intermediate step to establish consistency of the proposed prediction interval as well as to build explicit rates of convergence for the excess-risk of $\hat{\Gamma}$. This is the purpose of the next paragraph and Section 4.3.3 respectively.

Consistency result. Proposition 13 shows that the consistency of $\hat{\Gamma}$ with respect to the excess-risk relies to the consistency of the estimator $\hat{p}(y|\mathbf{x})$. In view of Equation (4.4), it is clear that the performance of $\hat{p}$ is directly linked to the statistical properties of $\hat{f}$ and $\hat{\sigma}$. More precisely, we obtain the following result.

Theorem 15. *Let Assumptions 12, 13, and 14. Consider* $s = \log(\min(n, N))$, $M > 4\sqrt{N}$, *and* $u = u_n \to 0$. *Assume that*

$$\sqrt{s}\mathbb{E}\left[(\hat{f}(X) - f^*(X))^2\right] \to 0, \quad \text{and} \quad s^{5/2}\mathbb{E}\left[|\hat{\sigma}^2(X) - \sigma(X)|\right] \to 0 \ ,$$

then the following holds

$$\mathbb{E}\left[\mathcal{E}_\ell\left(\hat{\Gamma}\right)\right] \leq C_2\mathbb{E}\left[\mathcal{H}\left(\hat{\Gamma}\right)\right] \to 0 \ .$$

Let us make several comments on this theorem. First, under suitable assumptions, both excess-risk and expected symmetric difference of $\hat{\Gamma}$ converge to 0. Notably, since $\mathbb{E}\left[\mathcal{E}_\ell\left(\hat{\Gamma}\right)\right] \leq C_2\mathbb{E}\left[\mathcal{H}\left(\hat{\Gamma}\right)\right]$, consistency *w.r.t.* the expected symmetric difference implies consistency *w.r.t.* the excess-risk. From this perspective, symmetric difference control is a more difficult problem that excess-risk control. In particular, $\mathbb{E}\left[\mathcal{H}\left(\hat{\Gamma}\right)\right] \to 0$ indicates that $\hat{\Gamma} = \Gamma_\ell^*$ asymptotically. Another aspect that need to be discussed is the assumptions that are requested for the proof of Theorem 15. More specifically, consistency of $\hat{f}$, and $\hat{\sigma}^2$ are naturally required to ensure that $\hat{p}$ is a consistent estimator of p. In particular, convergence of $\hat{f}$ and $\hat{\sigma}$ can be made possible by several learning algorithms such as kernel methods, local polynomials, regularized least-squares among many others.

4.3.3 Rates of convergence

Theorem 15 establishes the consistency of the prediction interval $\hat{\Gamma}$ under mild assumptions. In this section, we focus on rates of convergence. More structural assumptions are then required. We borrow conditions from [24] introduced in the framework of regression with abstention. We assume that $\mathbf{X}$ belongs to a compact $\mathcal{C}$, and we consider the following assumptions.

Assumption 15 (Regularity)**.** *The functions* f^* *and* σ^2 *are Lipschitz.*

Assumption 16 (Strong density assumption)**.** *The marginal distribution* $\mathbb{P}_\mathbf{X}$ *satisfies the strong density assumption*

- $\mathbb{P}_\mathbf{X}$ *is supported on a compact regular set* $\mathcal{C} \subset \mathbb{R}^d$,
- $\mathbb{P}_\mathbf{X}$ *admits a density* μ *w.r.t. to the Lebesgue measure such that* $0 < \mu_{\min} \leq \mu(\mathbf{x}) \leq \mu_{\max} < \infty$, *for all* $\mathbf{x} \in \mathcal{C}$.

Assumption 17 (α-Margin assumption)**.** *We say that* $p(\cdot|X)$ *satisfies Margin assumption with parameter* $\alpha \geq 0$ *at level* λ_ℓ *with respect to* $\mathbb{P}_X$ *if there exist constants* $c_0 > 0$ *and* $t_0 > 0$ *such that for all* $0 < t \leq t_0$,

$$\int_{\mathbb{R}} \mathbb{P}_X\left(|p(y|\mathbf{X}) - \lambda_\ell| \leq t\right) dy \leq c_0 t^\alpha \ .$$

The above first two assumptions are rather classical when we deal with rates of convergence in nonparametric statistic. We refer the reader to the book [36] for a more detail discussion. In addition, Assumption 17, also known as Tsybakov noise condition [85], has been introduced in the binary classification setting to get fast rates of convergence [2]. In our setting, we notice that the Tsybakov noise condition is required around the threshold λ_ℓ. Moreover, since we extend this assumption to the case of regression, we need to integrate it *w.r.t.* $y \in \mathbb{R}$. Based on the above conditions, we can establish the following result.

Proposition 14. *Let Assumptions 12, 15, 16, and 17 be satisfied. For $s = \log(\min(n, N))$, and $M > 4\sqrt{N}$, we have that*

$$\mathbb{E}\left[\mathcal{E}_\ell\left(\hat{\Gamma}\right)\right] \leq C\left(\mathbb{E}\left[\left(\sup_{(x,y)\in\mathcal{C}\times[-s,s]} |\hat{p}(y|x) - p(y|x)|\right)^{1+\alpha}\right] + \frac{1}{\min(n,N)^{1+\alpha}} + u^{1+\alpha} + \frac{\log(N)}{\sqrt{N}}\right),$$

where $C > 0$ is a constant which depends on f^*, σ^2, c_0, α, and $\mathcal{C}$.

As compared to the upper-bound that we get in Proposition 13, the bound here is better because of the exponent $1 + \alpha$ against 1. However, it is obtained under stronger assumptions.

Estimators of regression and variance function. The framework that we have described so far is quite general and allows to use any off-the-shelf machine learning algorithms to estimate the regression and the variance functions. In what follows, we propose a more concrete illustration of our approach by considering empirical prediction intervals $\hat{\Gamma}$ where both regression and variance functions are estimated with the kNN algorithm. Hereafter, we briefly recall the definition of the estimators that are based on the $l$abeled sample $\mathcal{D}_n$. For any $\mathbf{x} \in \mathbb{R}^d$, we denote by $(\mathbf{X}_{(i,n)}(\mathbf{x}), Y_{(i,n)}(\mathbf{x})), i = 1, \dots n$ the reordered data according to the ℓ_2 distance in $\mathbb{R}^d$, meaning that

$$\|\mathbf{X}_{(i,n)}(\mathbf{x}) - \mathbf{x}\| < \|\mathbf{X}_{(j,n)}(\mathbf{x}) - \mathbf{x}\| \ ,$$

for all $i < j$ in $\{1, \dots, n\}$. For simplicity, we assume that ties occur with probability 0. Let $k = k_n$ be an integer. The kNN estimator of f^* and σ^2 are then defined, for all $\mathbf{x} \in \mathbb{R}^d$, as follows

$$\hat{f}(\mathbf{x}) = \frac{1}{k_n}\sum_{i=1}^{k_n} Y_{(i,n)}(\mathbf{x}) \ \text{ and } \ \tilde{\sigma}^2(\mathbf{x}) = \frac{1}{k_n}\sum_{i=1}^{k_n}\left(Y_{(i,n)}(\mathbf{x}) - \hat{f}(X_{(i,n)}(\mathbf{x}))\right)^2 \ . \tag{4.7}$$

The properties of these estimator are provided in [36] for the regression function and in [24] for the variance function. In particular, the authors in [24] establish rates of convergence *w.r.t.* the sup-norm for the estimator $\hat{\sigma}$.

Rates of convergence. The next result, which is an adaption of Proposition 3.1 in [20], is useful to derive upper-bound on the measure of risk $\mathcal{H}$ of $\hat{\Gamma}$ thanks to a control on the excess-risk.

Proposition 15. *Let Assumptions 17 be satisfied. There exists an absolute constant $C_3 > 0$ such that*

$$\mathbb{E}\left[\mathcal{H}(\hat{\Gamma})\right] \leq C_3\left(\mathbb{E}\left[\mathcal{E}_\ell\left(\hat{\Gamma}\right)\right]\right)^{\alpha/\alpha+1} \ .$$

Importantly, this proposition, together with the inequality $\mathcal{E}_\ell(\Gamma) \leq C_2\mathcal{H}(\Gamma)$ for all Γ, shows that under appropriate regularity condition consistency of $\hat{\Gamma}$ *w.r.t.* the distance $\mathcal{H}$ and the excess-risk are equivalent. The only difference is in the rates of convergence. The above result highlights the link between them under Assumption 17. In particular, we only have to establish rates of convergence *w.r.t.* $\mathcal{E}$. Let us introduce the following notation. When $a \propto b$, it means that the quantities a and b are equal up to a constant. Moreover $\lesssim_{\log(n)}$ says that the inequality holds up to some constants and logarithmic factors. Now, we state the main result of this section.

Theorem 16. *Let Assumptions 12 and 15-17 be satisfied. Let $k_n \propto n^{2/d+2}$, $s = \log(\min(n, N))$, $M > 4\sqrt{N}$, and $u_n = \frac{1}{n}$. The following holds*

$$\mathbb{E}\left[\mathcal{E}_\ell(\hat{\Gamma})\right] \lesssim_{\log(n)} n^{-(1+\alpha)/(d+2)} + \min(n, N)^{-(1+\alpha)} + N^{-1/2} \ .$$

Several comments can be made from the above result. The first term is the classical nonparametric fast rate of convergence for the excess-risk under the Margin assumption and the Lipschitzness of the regression function. The last two terms that are related to the problem of PI estimation have different behavior according to the interplay between n and N. In particular, as soon as $N \leq n$,

the limiting term is $N^{-1/2}$ and the rate becomes slow if $n^{-(1+\alpha)/(d+2)}$ goes faster to 0. On the other hand, if the number of unlabeled data N is large with $N \gg n^{1+\alpha}$ we recover the fast rate of convergence $n^{-(1+\alpha)/(d+2)}$. Between these two extremes, $N^{-1/2}$ can still be the limiting term. However, we hope that in our semi-supervised setting, enough data are available to make this term negligible as compared to the others.

4.4 Extension and other approach

In this section, we discuss some points beyond the considered framework in this chapter. The extension of our results to other regression models is presented in Section 4.4.1. Another approach to build prediction interval based on the control of the expected error rate [52] is described in Section 4.4.2. In particular, we exhibit the main differences with our considered procedure.

4.4.1 Beyond Gaussian setting

In the present work, we study prediction intervals under expected length constraint in the heteroscedastic Gaussian regression setup. The appealing aspect of this framework lies in the form of the optimal predictor

$$\Gamma_\ell^*(\mathbf{X}) = \{y \in \mathbb{R} : \ p(y|\mathbf{X}) \geq \lambda_\ell^*\} \ , \tag{4.8}$$

with $\lambda_\ell^* = G^{-1}(\ell)$ and $G(t) := \int_{\mathbb{R}} \mathbb{P}(p(y|\mathbf{X}) \geq t)dy$. Furthermore, the density $p(y|\mathbf{X})$ has an explicit expression that exclusively depends on the regression and the conditional variance functions f and σ. Therefore, our proposed algorithm only involves estimators of f and σ to estimate the conditional density p. In particular, we do not consider any general procedure for density estimation.

In this paragraph, we discuss possible extensions outside the Gaussian framework but still considering the regression framework $Y = f^*(\mathbf{X}) + \sigma(\mathbf{X})\,\varepsilon$. In order to make sure that the optimal predictor is well defined, we require the following assumption.

Assumption 18. *We assume that the variable Y given $\mathbf{X}$ has density $p(\cdot|\mathbf{X})$.*

If we do not assume that $Y|\mathbf{X}$ belongs to a given family of distribution, the characterization of the prediction interval (4.8) still holds but the expression of the conditional density can not be simplify. Therefore, a data-driven predictor, based on the plug-in principle, must rely on estimates $\hat{p}(\cdot|\mathbf{x})$ of the conditional density $p(\cdot|\mathbf{x})$. The way to build the estimator $\hat{\Gamma}$ does not differ from the Gaussian case ones $\hat{p}$ is obtained (see Section 4.3). From the theoretical perspective, general properties such as Propositions 10 and 11 still hold and the question here is to investigate consistency results of the algorithm $\hat{\Gamma}$. The control on the expected length of the prediction interval $\mathbb{E}\left[\left|\mathcal{L}(\hat{\Gamma}) - \ell\right|\right] \leq C\frac{s}{\sqrt{N}}$ given in Proposition 12 is also still valid since this result is distribution-free. On the other hand, consistency for the excess-risk requires conditions. In the case where Y is bounded, if the estimator of the conditional probabilities is such that $\mathbb{E}\left[\int_{\mathbb{R}} |\hat{p}(y|\mathbf{X}) - p(y|\mathbf{X})|dy\right] \underset{n\to+\infty}{\longrightarrow} 0$, we can establish under Assumptions 12, 13, 14, and 18 that

$$\mathbb{E}\left[\mathcal{H}\left(\hat{\Gamma}\right)\right] \underset{n,N\to+\infty}{\longrightarrow} 0 \ .$$

Essentially, this result says that the estimation procedure that we study in this chapter extends beyond the Gaussian setting. In particular, we still manage to get consistency for bounded random variable. It is worth mentioning that consistency might also be obtained as soon as $Y|\mathbf{X}$ is sub-Gaussian. Then our method is statistically valid for general settings.

4.4.2 Prediction interval under expected coverage constraint

In this section, we present the approach which focuses on the construction of prediction interval under expected coverage. This method consists in minimizing the length of the prediction interval under a constraint on its expected error rate. This approach is for instance studied in [52].

More precisely, let $\beta > 0$. We consider the following problem

$$\Gamma^*_\beta = \arg \min_{\mathbb{P}(Y \notin \Gamma(\mathbf{X}) \leq \beta} \mathbb{E}\left[L(\Gamma(\mathbf{X})]\right] \ .$$

Under Assumptions 14 and 18 we can derive an expression of Γ^*_β based on thresholding of the conditional densities:

$$\Gamma^*_\beta = \{y \in \mathbb{R}, \ \ p(y|\mathbf{x}) \geq t_\beta\} \ ,$$

with t_β defined as solution of

$$\mathbb{E}\left[\mathbb{1}_{\{p(Y|\mathbf{X}) \geq t_\beta\}}\right] = \int_{\mathbb{R}} \mathbb{1}_{\{p(y|\mathbf{x}) \geq t_\beta\}} p(y|\mathbf{x}) \, \mathrm{d}y = 1 - \beta \ .$$

Therefore, from the above equation, we deduce that

$$t_\beta = H^{-1}(1 - \beta) \ ,$$

where $H(t) = \mathbb{E}\left[\mathbb{1}_{\{p(Y|\mathbf{X}) \geq t\}}\right]$. Similarly to the procedure described in Section 4.3.1, we are able to provide a randomized prediction interval $\hat{\Gamma}_\beta$ based on the estimator $\hat{p}$. We point out that an important difference between the construction of estimators $\hat{\Gamma}_\beta$ and $\hat{\Gamma}$ is the estimation of the function H. Indeed, this step require a *labeled* and not an *unlabeled* dataset, but do not request the discretization step. More formally, considering a *labeled* dataset $\mathcal{D}_K = \{(\mathbf{X}_i, Y_i), \ i = 1, \ldots, K\}$, and $(\zeta_1, \ldots, \zeta_K)$ the vector of perturbation, the estimator $\hat{H}$ of the function H is defined for each $t > 0$, as follows

$$\hat{H}(t) = \frac{1}{K} \sum_{i=1}^{K} \mathbb{1}_{\{\hat{p}(Y_i|\mathbf{X}_i, \zeta_i) \geq t\}} \ .$$

Although a theoretical comparison with our proposed method is not our purpose, using similar arguments as in [52], we can establish the consistency of $\hat{\Gamma}_\beta$ under same assumptions as in Theorem 15.

$$\mathbb{E}\left[\mathcal{H}\left(\hat{\Gamma}_\beta\right)\right] \to 0 \ .$$

In Section 4.5, we focus on a comparison between our method and the expected coverage approach from a numerical perspective.

4.5 Numerical experiments

This section is devoted to a numerical study of the performance of our procedure. More precisely, we analyze our approach on synthetic data in Section 4.5.1 and provide a comparison with the expected coverage approach described in Section 4.5.2.

4.5.1 Simulation study

We illustrate the performance of our procedure on the following model

$$Y = \exp(-\|\mathbf{X}\|_2) + \frac{d\varepsilon}{2 + 4\|\mathbf{X}\|_2}, \ \ \mathbf{X} \in \mathbb{R}^d \ , \tag{4.9}$$

where $\mathbf{X} = (X^1, \ldots, X^d)$ is such that for $j = 1, \ldots, d$, the X^j are i.i.d. simulated according to a Uniform distribution on $[0, 1]$ and are independent from $\varepsilon \sim \mathcal{N}(0, 1)$. Note that the considered model satisfies Equation 4.1, and that Assumptions 12, and 13 are fulfilled.

For our numerical experiments, we choose reasonable dimensions of the features space $d \in \{1, 5\}$. Before going further in our investigations, we display the boxplots of the output variable Y in Figure 4.1. We see that the range of values of Y is much larger for $d = 5$ and is included in $[-5, 5]$ for both $d = 1, 5$. Besides, we choose to focus on $\ell \in \{0.1, 0.5, 1, 2\}$ which seems to be relevant values according to Figure 4.1 in order to still get interpretation of the output. For $\ell \in \{0.1, 0.5, 1, 2\}$, we provide the evaluation of the expected length and the error rate for the oracle prediction set Γ^*_ℓ. To this end, we repeat 100 times the following scheme.

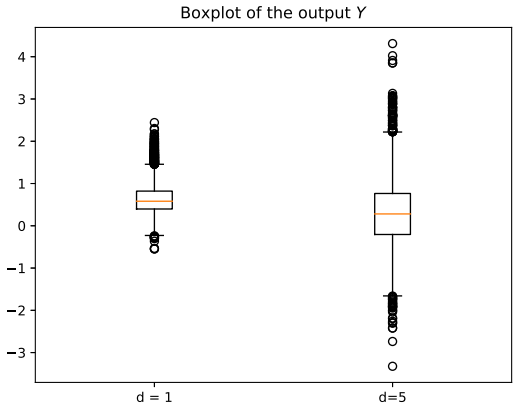


Figure 4.1 – Boxplot of the output Y for $d = 1, 5$

	Expected length		Error rate	
ℓ	$d = 1$	$d = 5$	$d = 1$	$d = 5$
0.1	0.1 (0.01)	0.1 (0.01)	0.81 (0.01)	0.94 (0.01)
0.5	0.49 (0.01)	0.49 (0.01)	0.34 (0.01)	0.71 (0.01)
1	0.99 (0.01)	0.99 (0.01)	0.07 (0.01)	0.48 (0.01)
2	1.99 (0.03)	1.99 (0.01)	0.00 (0.00)	0.17 (0.01)

Table 4.1 – Performance of the Oracle PI for $\ell \in \{0.1, 0.5, 1, 2\}$.

i) estimate λ_ℓ^* from an *unlabeled* dataset of size $N = 1000$ on a regular grid of size $M = 1000$ of the interval $[-5, 5]$;

ii) derive the resulting prediction interval on the same grid over a test set of size $T = 1000$;

iii) based on the test set, compute the expected length and the error rate.

From these repetitions, we compute the mean and standard deviation of the estimates. The obtained results are provided in Table 4.1.

Simulation scheme. To assess the performance of our procedure, we consider the following scheme. For $d \in \{1, 5\}$ and $\ell \in \{0.1, 0.5, 1, 2\}$, we repeat 100 the following steps.

i) estimate f^* and σ^2 from a training test of size $n = 500$. We consider the residual-based method [37]. The estimation of f^* and σ^2 relies on the random forests algorithm from `python` library `sklearn`. We also choose $u = 10^{-5}$ for the parameter of the perturbation ζ (see Eq. (4.5));

ii) compute $\hat{G}^{-1}(\ell)$ using an *unlabeled* dataset of size $N = 100$ on a regular grid of size $M = 100$ of the interval $[-s, s]$, where $s = \max(-\min(Y_{train}); \max(Y_{train})$,

iii) derive the resulting prediction interval on a regular grid of size 1000 of $[-s, s]$ over a test set of size $T = 1000$;

iv) based on the test set, compute the expected length and the error rate.

From these experiments, we compute the empirical means and standard deviations expected length and the error rate. The results are provided in Table 4.2. A visual description of the behavior of our PI is also given in Figure 4.2.

Notice that the value of s that we consider here is different from the one suggested by the theory in 16. This is a minor point. The parameter s in the theory is set such that most of the labels lie in $[-s, s]$ with high probability. This happens when n and M grow since $s = \log(\min(n, N))$. Our choice in practice ensures that this property holds regardless the values of n and N.

	Expected length		Error rate	
ℓ	$d = 1$	$d = 5$	$d = 1$	$d = 5$
0.1	0.1 (0.01)	0.1 (0.02)	0.81 (0.02)	0.94 (0.01)
0.5	0.50 (0.01)	0.50 (0.02)	0.34 (0.02)	0.72 (0.02)
1	1.00 (0.02)	1.00 (0.02)	0.07 (0.01)	0.48 (0.01)
2	2.01 (0.06)	2.01 (0.01)	0.00 (0.00)	0.17 (0.01)

Table 4.2 – Performance of $\hat{\Gamma}$ for $\ell \in \{0.1, 0.5, 1, 2\}$.

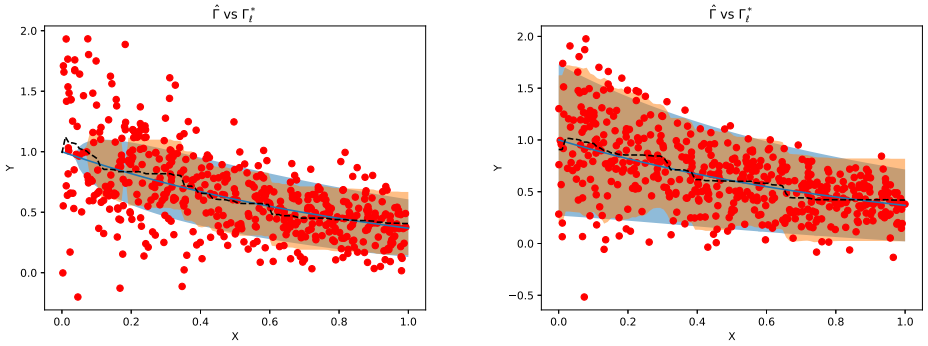


Figure 4.2 – Visual description of the empirical PI $\hat{\Gamma}$ and its oracle counterpart Γ^*_ℓ, with $\ell = 0.5$ on the left and $\ell = 1$ on the right for $d = 1$. The scatter plot of data is displayed and the graph of both regression function f^* and estimator $\hat{f}$ is represented (solid line for f^*, dashed line for $\hat{f}$. The oracle PI Γ^*_ℓ (empirical PI $\hat{\Gamma}$, respectively) is given in blue (orange, respectively).

Results. Two conclusions can be made from this first numerical study. First Tables 4.1 and 4.2 highlight how effective our method is in producing PI with (almost) exactly the right length. This is an important point and suggests that our strategy succeeds to enforce the constraint on the length prescribed by the optimization problem. Second, let us focus on a comparison between Γ^*_ℓ, the oracle PI, and its empirical counterpart $\hat{\Gamma}$. Table 4.1 and Table 4.2 show how close are the performance of these two PI both in terms of expected length and of error rate. Interestingly, the performance of $\hat{\Gamma}$ is obtained with a moderate size N of the unlabeled sample that is used to estimate the threshold. These results also suggest that $n = 500$ is enough to have good estimations of the regression and variance functions. The closeness between Γ^*_ℓ and $\hat{\Gamma}$ is also illustrated in Figure 4.2.

4.5.2 Numerical comparison with expected coverage approach

In this section, we numerically compare our procedure to the approach that constraint the expected coverage described in Section 4.4.2. We consider the model defined in Equation 4.9 with $d = 5$ and focus on the estimation of Γ^*_ℓ for $\ell = 2$. With this expected length, the oracle predictor Γ^*_ℓ reaches an error rate of $\beta = 0.17$. Therefore, for this learning task, we are able to provide empirical PI for both approaches. That is to say, we compute $\hat{\Gamma}$ with $\ell = 2$ as expected length and $\hat{\Gamma}_\beta$ with $\beta = 0.17$ as expected error. In order to get a fair comparison of the methods, we repeat 20 times the following steps. For both approaches, we use a training set of size $n = 500$ to estimate the density p and we estimate the threshold of the considered procedure with a dataset of size $N \in \{10, 30, 50, 70, 100, 150, 200, 500, 1000\}$. Finally, we compute the expected length and error rate of both empirical PI over a test set of size $T = 1000$. From these repetitions, we compute empirical means and standard deviations. The results are displayed in Figure 4.3.

As expected, in average, both methods behaves similarly. However there are important differences in favor of our approach. First, the convergence of our method is much faster to the mean value both for the expected length and the error rate. We notice that $N = 10$ is already enough for our method while more than 500 samples are needed for the method that focus on the coverage as constraint. Second, it seems that our construction is much more stable, in particular for length calibration. It illustrates the efficiency of our procedure to build prediction interval with the right

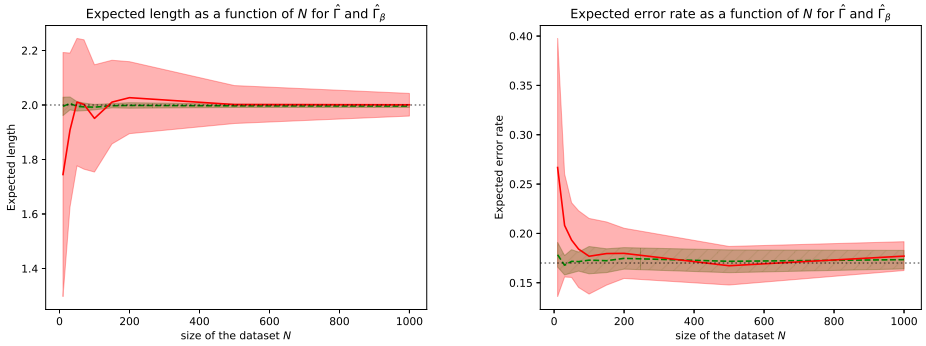


Figure 4.3 – Comparison between $\hat{\Gamma}$ and $\hat{\Gamma}_\beta$. We plot the expected length (on the left) and the expected coverage (on the right) as a function of N over 20 repetitions for $\hat{\Gamma}$ (dashed) and $\hat{\Gamma}_\beta$ (solid line in red). The true value of the parameter is given by the dotted line.

expected length.

The two approaches are definitively not comparable in terms of objectives. Indeed, if we are really focused on constraining the error rate, then the length constraint appears (at first sight) sub-optimal and vice versa if we ask for interpretable outputs. However, our numerical analysis clearly suggests that our methodology is more stable: it induces a procedure with a lower variance.

4.6 Conclusion

In this chapter, we provide a general methodology to build *prediction intervals with controlled expected length* in the Gaussian regression. Our proposed algorithm is very effective in controlling the expected length of the output and then ensure the interpretability of the outcome. The theoretical analysis indicates that our method mimics the optimal rule *w.r.t.* the expected length and, under appropriate properties on the base estimators of the regression function, it is also efficient *w.r.t.* the symmetric difference distance and the excess-risk. Furthermore, a numerical study supports our theoretical results. Notably, it highlights good stability properties as compared to prediction intervals that focus on expected coverage constraints.

Our numerical comparison to PI under expected coverage constraint additionally opens a very significant door to the use of our method. Because of the stability of our method, one may think to the following two-stage procedure to produce a PI with error rate β.

- *Step 1.* Build the PI with error rate β and evaluate its length $\tilde{\ell}$;
- *Step 2.* Build our PI with average length $\tilde{\ell}$.

While we do not expect a significant improvement in average, the resulting prediction interval might be more stable. This will be the purpose of future investigation.

On the other hand, inference in the high-dimensional setting is a crucial challenge with modern data. Several successful studies consider the Gaussian *homoscedastic* linear regression [57, 89, 62, 6]. An important direction for future research is to carry out PI i) for non Gaussian models; ii) and that can handle heteroscedastic model. Both of these questions have their applications in the high dimensional setting.

4.7 Appendix

This appendix is devoted to the proof of our main results. The proofs related to Section 4.2 are provided in Section 4.7.2, while Section 4.7.3 is devoted to the proofs of Section 4.3. Finally, Section 4.7.1 gathers useful results. In particular, we give rates of convergence for KNN estimates for both regression and variance function. Notice that in the whole appendix, C is a positive constant that may change from one line to another.

4.7.1 Technical results

In this section, we provide some useful properties that are used for the proof of our main results

Technical lemmas

The first tool we introduce is a generalization of the classical inverse transform theorem [90, Lemma 21.1] to the continuous case. Let $a > 0$. We consider a random process $(Z_y)_{y\in[-a,a]}$ such that the function H defined by

$$H(t) = \frac{1}{2a}\int_{-a}^{a} \mathbb{P}(Z_y \geq t)dy \ ,$$

is continuous on $\mathbb{R}_+$.

Lemma 15. *Let T uniformly distributed on $[-a,a]$ and independent of $(Z_y)_{y\in[-a,a]}$. We consider the random variable Z_T and let U be distributed according to the uniform distribution on $[0,1]$. Then*

$$H(Z_T) \overset{\mathcal{L}}{=} U \ \text{ and } \ H^{-1}(U) \overset{\mathcal{L}}{=} Z_T \ .$$

Proof. For every $t \geq 0$, we have $\mathbb{P}(H(Z_T) \leq t) = \mathbb{P}(Z_T \geq H^{-1}(t))$. Denote by $d\mathbb{P}_T$ the marginal distribution of T. Since the variable T is independent of $(Z_y)_{y\in[-a,a]}$ and H is continuous, one gets

$$\begin{aligned}
\mathbb{P}(H(Z_T) \leq t) &= \int \mathbb{P}(Z_T \geq H^{-1}(t)|T=y))\, d\mathbb{P}_T(y) \\
&= \frac{1}{2a}\int_{-a}^{a} \mathbb{P}(Z_y \geq H^{-1}(t)|T=y))\, dy \\
&= \frac{1}{2a}\int_{-a}^{a} \mathbb{P}(Z_y \geq H^{-1}(t))\, dy = H(H^{-1}(t)) = t \ ,
\end{aligned}$$

and we deduce that $H(Z_T) \overset{\mathcal{L}}{=} U$. For the second point of the Lemma, we observe that

$$\mathbb{P}(H^{-1}(U) \leq t) \ = \ \mathbb{P}(U \geq H(t)) = \frac{1}{2a}\int_{-a}^{a} \mathbb{P}(Z_y \leq t)dy = \frac{1}{2a}\int_{-a}^{a} \mathbb{P}(Z_y \leq t|T=y))dy = \mathbb{P}(Z_T \leq t) \ .$$

□

Rates of convergence for K-NN estimators

In this section, we gather the results we use for K-NN estimators of both regression and variance function. The proof of this result is provided in [24].

Theorem 17. *Grants Assumptions 15, 16, for $k_n \propto n^{-2/d+2}$, and all $\alpha > 0$, the K-NN estimators defined in Equation (4.7) satisfy*

$$\begin{aligned}
\mathbb{E}\left[\left(\sup_{\mathbf{x}\in\mathcal{C}} |\hat{f}(\mathbf{x}) - f^*(\mathbf{x})|\right)^{1+\alpha}\right] &\leq C\log(n)^{1+\alpha} n^{-(1+\alpha)/(2+d)} \ , \\
\mathbb{E}\left[\left(\sup_{\mathbf{x}\in\mathcal{C}} |\hat{\sigma}^2(\mathbf{x}) - \sigma(\mathbf{x})|\right)^{1+\alpha}\right] &\leq C\log(n)^{1+\alpha} n^{-(1+\alpha)/(2+d)} \ .
\end{aligned}$$

4.7.2 Proof of Section 4.2

In this section, we provide proofs related to the optimal confidence and to the excess-risk formula

Proof of Proposition 10. First, let us consider the Lagrangian of the optimization problem 4.2. It can be written as

$$H(\Gamma, \lambda) = \mathbb{P}\left(Y \notin \Gamma(\mathbf{X})\right) + \lambda\left(\mathbb{E}_{\mathbf{X}}[L(\Gamma(\mathbf{X})] - \ell\right) \ ,$$

where $\lambda \geq 0$ is a dual variable of the problem. Since,

$$\mathbb{P}\left(Y \in \Gamma(\mathbf{X})\right) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}\left[\mathbb{1}_{\{Y\in\Gamma(\mathbf{X})\}}|\mathbf{X}\right]\right] = \mathbb{E}_{\mathbf{X}}\left[\int_{\mathbb{R}} p(y|\mathbf{X})\mathbb{1}_{\{y\in\Gamma(\mathbf{X})\}}dy\right] \ ,$$

the Lagrangian reads as

$$H(\Gamma,\lambda) = 1 - \lambda\ell - \mathbb{E}_{\mathbf{X}}\left[\int_{\mathbb{R}} (p(y|\mathbf{X}) - \lambda)\mathbb{1}_{\{y\in\Gamma(\mathbf{X})\}}dy\right] \ . \tag{4.10}$$

Minimizing *w.r.t.* Γ leads to an optimal solution that can be written for all $\lambda \geq 0$ and all $\mathbf{x} \in \mathbb{R}^d$ as

$$\Gamma^*(\lambda,\mathbf{x}) = \{y \in \mathbb{R} : p(y|\mathbf{X}) \geq \lambda\} \ .$$

Injecting this value into 4.10 gives

$$H(\Gamma^*(\lambda,\mathbf{X}),\lambda) = 1 - \lambda\ell - \mathbb{E}_{\mathbf{X}}\left[\int_{\mathbb{R}} (p(y|\mathbf{X}) - \lambda)_+\mathbb{1}_{\{y\in\Gamma^*(\lambda,\mathbf{X})\}}dy\right] \ ,$$

where $(\cdot)_+$ stands for the positive part. First order optimality conditions for convex non-smooth minimization problems implies $0 \in \partial H(\Gamma^*(\lambda_\ell^*,\mathbf{X}),\lambda_\ell^*)$ where ∂H is the sub-differential of H. Therefore, using the Fundamental Theorem of Calculus, we get $\mathbb{E}_{\mathbf{X}}\left[\int_{\mathbb{R}} \mathbb{1}_{\{y\in\Gamma^*(\lambda_\ell^*,\mathbf{X})\}}dy\right] = \ell$. But, using the above definition of Γ^* we can write by Fubini's theorem the left hand side term as $\mathbb{E}_{\mathbf{X}}\left[\int_{\mathbb{R}} \mathbb{1}_{\{y\in\Gamma^*(\lambda_\ell^*,\mathbf{X})\}}dy\right] = \int_{\mathbb{R}} \mathbb{P}\left((p(y|\mathbf{X}) \geq \lambda_\ell^*\right)dy = G(\lambda_\ell^*)$. We then conclude that $\lambda_\ell^* = G^{-1}(\ell)$. Notice that for this value, we have

$$\mathcal{L}(\Gamma^*) = \mathbb{E}_{\mathbf{X}}[L(\Gamma^*(\lambda_\ell^*,\mathbf{X})] = \mathbb{E}_{\mathbf{X}}\left[\int \mathbb{1}_{\{y\in\Gamma^*(\lambda_\ell^*,\mathbf{X})\}}dy\right] = G(\lambda_\ell^*) = \ell \ .$$

□

Proof of Proposition 11. Let $\ell \geq 0$. Considering a similar decomposition as in the proof of Proposition 10, we can write the error rate of a predictor Γ as

$$R_\ell(\Gamma) = 1 - \mathbb{E}_{\mathbf{X}}\left[\int_{\mathbb{R}} (p(y|\mathbf{X}) - \lambda_\ell^*)\mathbb{1}_{\{y\in\Gamma(\mathbf{X})\}}dy\right] \ . \tag{4.11}$$

Therefore, we deduce

$$\mathcal{E}_\ell\left(\Gamma\right) = \mathbb{E}_{\mathbf{X}}\left[\int_{\mathbb{R}} (p(y|\mathbf{X}) - \lambda_\ell^*)\left(\mathbb{1}_{\{y\in\Gamma_\ell^*(\mathbf{X})\}} - \mathbb{1}_{\{y\in\Gamma(\mathbf{X})\}}\right)dy\right] \ ,$$

and the result follows from the fact that $\mathbb{1}_{\{y\in\Gamma_\ell^*(\mathbf{X})\}} - \mathbb{1}_{\{y\in\Gamma(\mathbf{X})\}} = \operatorname{sgn}(p(y|\mathbf{X}) - \lambda_\ell^*)$ since we have the equality between events $\{y \in \Gamma_\ell^*(\mathbf{X})\} = \{p(y|\mathbf{X}) - \lambda_\ell^* \geq 0\}$, where $\operatorname{sgn} : \mathbb{R} \to \{-1,1\}$ stands for the sign. □

4.7.3 Proof of Section 4.3

We now consider the theoretical properties of the prediction interval $\hat{\Gamma}$. We first consider its expected length and then derive a finite sample bound on its excess-risk.

Length control

Proof of Proposition 12. To show this result, we need to introduce some pseudo-oracle predictor that has expected length ℓ. Let us then define the randomized predictor

$$\bar{\Gamma}(\mathbf{X},\zeta) = \{y \in \mathbb{R} : \ \hat{p}(y|\mathbf{X},\zeta) \geq \bar{G}^{-1}(\ell)\} \ , \tag{4.12}$$

where $\bar{G}(t) := \int_{\mathbb{R}} \mathbb{P}_{\mathbf{X},\zeta}(\hat{p}(y|\mathbf{X},\zeta) \geq t)dy$ for all $t > 0$. Here again, the property $\mathcal{L}(\bar{\Gamma}) := \mathbb{E}_{\mathbf{X},\zeta}\left[L\left(\bar{\Gamma}(\mathbf{X},\zeta)\right)\right] = \ell$ is due to the fact that the conditional on the data $\mathcal{D}_n$ the r.v. $\hat{p}(y|\mathbf{X},\zeta)$ has no atoms since it is randomized.

Let us now consider the purpose of the proposition. We need to bound $\mathbb{E}\left[\left|\mathcal{L}(\hat{\Gamma}) - \ell\right|\right]$. We can write

$$\begin{aligned}
\left|\mathcal{L}(\hat{\Gamma}) - \ell\right| = \left|\mathcal{L}(\hat{\Gamma}) - \mathcal{L}(\bar{\Gamma})\right| &= \left|\mathbb{E}\left[\int_{\mathbb{R}} \left(\mathbb{1}_{\{\hat{G}(\hat{p}(y|\mathbf{X},\zeta))\leq\ell\}} - \mathbb{1}_{\{\bar{G}(\hat{p}(y|\mathbf{X},\zeta))\leq\ell\}}\right) dy\right]\right| \\
&\leq \mathbb{E}\left[\int_{\mathbb{R}} \left|\mathbb{1}_{\{\hat{G}(\hat{p}(y|\mathbf{X},\zeta))\leq\ell\}} - \mathbb{1}_{\{\bar{G}(\hat{p}(y|\mathbf{X},\zeta))\leq\ell\}}\right| dy\right] \\
&\leq \mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\{|\hat{G}(\hat{p}(y|\mathbf{X},\zeta))-\bar{G}(\hat{p}(y|\mathbf{X},\zeta))|\geq|\bar{G}(\hat{p}(y|\mathbf{X},\zeta))-\ell|\}} dy\right] \\
&= \int_{\mathbb{R}} \mathbb{P}\left(|\hat{G}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq |\bar{G}(\hat{p}(y|\mathbf{X},\zeta)) - \ell|\right) dy \ ,
\end{aligned} \tag{4.13}$$

where we use Fubini's theorem at last. Now notice that the above integral is limited to the compact $[-s, s]$ since, this is the support of the function $\hat{p}(\cdot|\mathbf{x}, z)$ for all $(\mathbf{x}, z) \in \mathbb{R}^d \times [0, u]$. To bound this integral, we make use of the peeling technique of [2]. That is, we consider for $\delta > 0$ and $y \in [-s, s]$

$$\begin{aligned}
A_0(y) &= \left\{0 \leq |\bar{G}(\hat{p}(y|\mathbf{X},\zeta)) - \ell| \leq \delta\right\} \\
A_j(y) &= \left\{2^{j-1}\delta \leq |\bar{G}(\hat{p}(y|\mathbf{X},\zeta)) - \ell| \leq 2^j\delta\right\}, \qquad \text{for} \quad j \geq 1 \ .
\end{aligned}$$

Since for $y \in [-s, s]$, the events $(A_j(y))_{j\geq 0}$ are mutually exclusive, we deduce

$$\begin{aligned}
&\int_{-s}^{s} \mathbb{P}\left(|\hat{G}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq |\bar{G}(\hat{p}(y|\mathbf{X},\zeta)) - \ell|\right) dy = \\
&\qquad \int_{-s}^{s} \sum_{j\geq 0} \mathbb{P}\left(|\hat{G}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq |\bar{G}(\hat{p}(y|\mathbf{X},\zeta)) - \ell| \ , \ A_j(y)\right) dy \ .
\end{aligned} \tag{4.14}$$

Controlling this term relies on a bound on $\int_{-s}^{s} \mathbb{P}(A_j(y))dy$. It is clear that $0 \leq \bar{G}(t) = \int_{-s}^{s} \mathbb{P}_X(\hat{p}(y|\mathbf{X},\zeta) \geq t|\mathcal{D}_n)dy \leq 2s$ for all $t \in [0, 1]$. We can apply Lemma 15 to say that $\bar{G}(Z_T)$ is uniformly distributed on $[0, 2s]$ and then, for all $j \geq 0$ and $\delta > 0$, we deduce that

$$\begin{aligned}
\int_{-s}^{s} \mathbb{P}(A_j(y))dy &= 2s\frac{1}{2s}\int_{-s}^{s} \mathbb{P}\left(|\bar{G}(\hat{p}(y|\mathbf{X},\zeta)) - \ell| \leq 2^j\delta \mid \mathcal{D}_n\right) dy \\
&= 2s \times \mathbb{P}\left(|\bar{G}(Z_T) - \ell| \leq 2^j\delta|\mathcal{D}_n\right) \leq 2s\frac{2^{j+1}\delta}{2s} = 2^{j+1}\delta \ .
\end{aligned} \tag{4.15}$$

Next, let us consider (4.14). We observe that for all $j \geq 1$

$$\begin{aligned}
&\int_{-s}^{s} \mathbb{P}\left(|\hat{G}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq |\bar{G}(\hat{p}(y|\mathbf{X},\zeta)) - \ell| \ , \ A_j(y)\right) dy \\
&\qquad \leq \int_{-s}^{s} \mathbb{P}\left(|\hat{G}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq 2^{j-1}\delta \ , \ A_j(y)\right) dy \\
&\qquad \leq \int_{-s}^{s} \mathbb{E}_{(\mathcal{D}_n,\mathbf{X},\zeta)}\left[\mathbb{P}_{\mathcal{D}_N}\left(|\hat{G}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq 2^{j-1}\delta\right)\mathbb{1}_{A_j(y)}\right] dy.
\end{aligned} \tag{4.16}$$

In Section 4.3.1, we have presented the predictor $\hat{\Gamma}$ that relies on the function $\hat{G}$ which is discretized. On the other hand, $\bar{G}$ is not discretized. Because of this difference, it is convenient, in order to control (4.16), to provide some additional notation. Let us define

$$\hat{\bar{G}}(t) := \frac{1}{N}\sum_{i=1}^{N} \int_{-s}^{s} \mathbb{1}_{\{\hat{p}(y|\mathbf{X}_{n+i},\zeta_i)\geq t\}} dy \ .$$

Then for all $y \in [-s, s]$, conditional on $(\mathcal{D}_n, \mathbf{X}, \zeta)$, the probability in Eq. (4.16) is bounded as follows

$$\mathbb{P}_{\mathcal{D}_N}\left(|\hat{G}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq 2^{j-1}\delta\right) \leq$$
$$\mathbb{P}_{\mathcal{D}_N}\left(|\hat{\bar{G}}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq 2^{j-1}\frac{\delta}{2}\right) + \mathbb{P}_{\mathcal{D}_N}\left(|\hat{\bar{G}}(\hat{p}(y|\mathbf{X},\zeta)) - \hat{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq 2^{j-1}\tfrac{\delta}{2}\right) \quad (4.17)$$

These two last terms are treated in different ways. For the first one, we observe that for all $t \in [0,1]$

$$\begin{aligned}|\hat{G}(t) - \hat{\bar{G}}(t)| &= \left|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{M}\left(\int_{y_k}^{y_{k+1}} \mathbb{1}_{\{\hat{p}(y|\mathbf{X}_{n+i},\zeta_i)\geq t\}} - \mathbb{1}_{\{\hat{p}(y_k|\mathbf{X}_{n+i},\zeta_i)\geq t\}}\right)dy\right| \\ &\leq \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{M}\left(\int_{y_k}^{y_{k+1}} \left|\mathbb{1}_{\{\hat{p}(y|\mathbf{X}_{n+i},\zeta_i)\geq t\}} - \mathbb{1}_{\{\hat{p}(y_k|\mathbf{X}_{n+i},\zeta_i)\geq t\}}\right|\right)dy \ .\end{aligned}$$

We recall that for all $|y| \leq s$, we have $\hat{p}(y|\mathbf{x},\zeta) = \hat{p}(y|\mathbf{x}) + \zeta$. Because, conditional on $\mathcal{D}_n$, the function $\hat{p}(\cdot|\mathbf{x})$ is a Gaussian density and since the perturbation ζ acts on each y in the same way, it turns out that the function $\hat{p}(.|\mathbf{x},\zeta)$ is continuously increasing and then decreasing with a maximum at $y = \hat{f}(\mathbf{x})$. Therefore, for any fixed t the indicators $\mathbb{1}_{\{\hat{p}(y|\mathbf{X}_{n+i},\zeta_i)\geq t\}}$ and $\mathbb{1}_{\{\hat{p}(y_k|\mathbf{X}_{n+i},\zeta_i)\geq t\}}$ differ at most in 2 intervals of the form $[y_k, y_{k+1}]$. Then we deduce that

$$|\hat{G}(t) - \hat{\bar{G}}(t)| \leq 2 \times \frac{2s}{M} \ .$$

Injecting this inequality to (4.17) gives

$$\mathbb{P}_{\mathcal{D}_N}\left(|\hat{G}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq 2^{j-1}\delta\right) \leq$$
$$\mathbb{P}_{\mathcal{D}_N}\left(|\hat{\bar{G}}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq 2^{j-1}\frac{\delta}{2}\right) + \mathbb{1}_{\{4s/M \geq 2^{j-2}\delta\}} \ . \quad (4.18)$$

Let us now consider the second term. Conditional on $(\mathcal{D}_n, \mathbf{X}, \zeta)$, the random variable $\hat{\bar{G}}(\hat{p}(y|\mathbf{X},\zeta))$ is an empirical mean of i.i.d. random variables of common mean $\bar{G}(\hat{p}(y|X,\zeta)) \in [0, 2s]$, we deduce from Hoeffding's inequality that

$$\mathbb{P}_{\mathcal{D}_N}\left(|\hat{\bar{G}}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq 2^{j-2}\delta|\mathcal{D}_n, \mathbf{X}\right) \leq 2\exp\left(\frac{-N\delta^2 2^{2j-1}}{16s^2}\right) \ .$$

Therefore, from Inequalities (4.14), (4.15), (4.16), and (4.18) one gets for $\delta = \frac{4s}{\sqrt{N}}$ and $M > 4\sqrt{N}$

$$\begin{aligned}&\int_{-s}^{s}\mathbb{P}\left(|\hat{G}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq |\bar{G}(\hat{p}(y|\mathbf{X},\zeta)) - \ell|\right)dy \\ &\qquad \leq \int_{-s}^{s}\mathbb{P}(A_0(y))dy + \sum_{j\geq 1} 2\exp\left(\frac{-N\delta^2 2^{2j-1}}{16s^2}\right)\int_{-s}^{s}\mathbb{P}(A_j(y))dy \\ &\qquad \leq 2\delta + \delta\sum_{j\geq 1} 2^{j+2}\exp\left(\frac{-N\delta^2 2^{2j-1}}{16s^2}\right) \leq \frac{Cs}{\sqrt{N}} \ . \quad (4.19)\end{aligned}$$

□

Excess-risk control

Proof of Proposition 13. Throughout the proof, we denote $\bar{\lambda}_\ell := \bar{G}^{-1}(\ell)$, where $\bar{G}$ is defined in Equation (4.12). We start with the following decomposition.

$$\mathcal{E}_\ell\left(\hat{\Gamma}\right) = \mathcal{E}\left(\bar{\Gamma}\right) + \left(R_\ell(\hat{\Gamma}) - R_\ell(\bar{\Gamma})\right) \ . \quad (4.20)$$

For the second term of the *r.h.s.* in the above equation, thanks to Equation (4.11), we have that

$$R_\ell(\hat{\Gamma}) - R_\ell(\bar{\Gamma}) = \mathbb{E}_{\mathbf{X},\zeta}\left[\int_{\mathbb{R}} (p(y|\mathbf{X}) - \lambda_\ell)\left(\mathbb{1}_{\{y\in\bar{\Gamma}(\mathbf{X},\zeta)\}} - \mathbb{1}_{\{y\in\hat{\Gamma}(\mathbf{X},\zeta)\}}\right) dy\right] \ .$$

From Assumption 12, we have that $|p(y|\mathbf{X}) - \lambda_\ell|$ is bounded by $C_1 > 0$ which depends on σ_0. Hence, we deduce that

$$\mathbb{E}\left[\left|R_\ell(\hat{\Gamma}) - R_\ell(\bar{\Gamma})\right|\right] \leq C_1\mathbb{E}\left[\int_{\mathbb{R}} \left|\mathbb{1}_{\{y\in\bar{\Gamma}(\mathbf{X},\zeta)\}} - \mathbb{1}_{\{y\in\hat{\Gamma}(\mathbf{X},\zeta)\}}\right| dy\right] \ .$$

This last inequality can be rewritten as

$$\mathbb{E}\left[\left|R_\ell(\hat{\Gamma}) - R_\ell(\bar{\Gamma})\right|\right] \leq C_1\mathbb{E}\left[\int_{\mathbb{R}} \left|\mathbb{1}_{\{\hat{G}(\hat{p}(y|\mathbf{X},\zeta))\leq\ell\}} - \mathbb{1}_{\{\bar{G}(\hat{p}(y|\mathbf{X},\zeta))\leq\ell\}}\right| dy\right] \ .$$

Therefore, from Equation 4.13, and (4.19), we deduce

$$\mathbb{E}\left[\left|R_l(\hat{\Gamma}) - R_l(\bar{\Gamma})\right|\right] \leq C\frac{s}{\sqrt{N}} \ . \tag{4.21}$$

Now we bound the first term in the *r.h.s.* in Equation (4.20). Thanks to Proposition 11, we have that

$$\mathcal{E}_\ell(\bar{\Gamma}) = \mathbb{E}_{\mathbf{X},\zeta}\left[\int_{\bar{\Gamma}(\mathbf{X},\zeta)\triangle\Gamma^*_\ell(\mathbf{X})} |p(y|\mathbf{X}) - \lambda^*_\ell| \ dy\right] \ .$$

Now, we consider the following cases

- If $y \in \bar{\Gamma}(\mathbf{X},\zeta) \setminus \Gamma^*_\ell(\mathbf{X})$, we have that $p(y|\mathbf{X}) < \lambda^*_\ell$ and $\hat{p}(y|\mathbf{X},\zeta) \geq \bar{\lambda}_\ell$. Therefore,

  $$|p(y|\mathbf{X}) - \lambda^*_\ell| = (\lambda^*_\ell - \bar{\lambda}_\ell) + (\bar{\lambda}_\ell - \hat{p}(y|\mathbf{X},\zeta)) + (\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})) \ .$$

  Using the fact that $\bar{\lambda}_\ell - \hat{p}(y|\mathbf{X},\zeta) \leq 0$, we get

  $$\int |p(y|\mathbf{X}) - \lambda^*_\ell|\mathbb{1}_{\{y\in\bar{\Gamma}(\mathbf{X},\zeta)\setminus\Gamma^*_\ell(\mathbf{X})\}}dy \leq \int \left((\lambda^*_\ell - \bar{\lambda}_\ell) + \left|\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})\right|\right)\mathbb{1}_{\{y\in\bar{\Gamma}(\mathbf{X},\zeta)\setminus\Gamma^*_\ell(\mathbf{X})\}}dy \ .$$

- If $y \in \Gamma^*_\ell(\mathbf{X}) \setminus \bar{\Gamma}(\mathbf{X},\zeta)$, we have that $p(y|\mathbf{X}) \geq \lambda^*_\ell$ and $\hat{p}(y|\mathbf{X},\zeta) < \bar{\lambda}_\ell$. Therefore,

  $$|p(y|\mathbf{X}) - \lambda^*_\ell| = (p(y|\mathbf{X}) - \hat{p}(y|\mathbf{X},\zeta)) + (\hat{p}(y|\mathbf{X},\zeta) - \bar{\lambda}_\ell) + (\bar{\lambda}_\ell - \lambda^*_\ell) \ .$$

  Using the fact that $\hat{p}(y|\mathbf{X},\zeta) - \bar{\lambda}_\ell < 0$, we get

  $$\int |p(y|\mathbf{X}) - \lambda_\ell|\mathbb{1}_{\{y\in\Gamma^*_\ell(\mathbf{X})\setminus\bar{\Gamma}(\mathbf{X},\zeta)\}}dy \leq \int \left((\bar{\lambda}_\ell - \lambda^*_\ell) + \left|\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})\right|\right)\mathbb{1}_{\{y\in\Gamma^*_\ell(\mathbf{X})\setminus\bar{\Gamma}(\mathbf{X},\zeta)\}}dy \ .$$

From the above considerations, we deduce the following inequality

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X},\zeta}\left[\int_{\bar{\Gamma}(\mathbf{X},\zeta)\triangle\Gamma^*_\ell(\mathbf{X})} |p(y|\mathbf{X}) - \lambda^*_\ell| \ dy\right] \\
&\qquad \leq |\bar{\lambda}_\ell - \lambda^*_\ell| \ \mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\{y\in\bar{\Gamma}(\mathbf{X},\zeta)\triangle\Gamma^*_\ell(\mathbf{X})\}}\mathrm{d}y\right] + \mathbb{E}\left[\int_{\mathbb{R}} |\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \, \mathrm{d}y\right] \\
&\qquad\qquad \leq |\bar{\lambda}_\ell - \lambda^*_\ell| \times (\mathcal{L}(\bar{\Gamma}) - \mathcal{L}(\Gamma^*)) + \mathbb{E}\left[\int_{\mathbb{R}} |\hat{p}(y|\mathbf{X}) - p(y|\mathbf{X})| \, \mathrm{d}y\right] + 2su \ ,
\end{aligned}$$

where the last inequality is due to the fact that $\hat{p}(y|\mathbf{X},\zeta) = \hat{p}(y|\mathbf{X}) + \zeta\mathbb{1}_{y\in[-s,s]}$ with $|\zeta| \leq u$. But $\mathcal{L}(\bar{\Gamma}) = \mathcal{L}(\Gamma^*) = \ell$ by construction. Then,

$$\mathbb{E}\left[\mathcal{E}_\ell(\bar{\Gamma})\right] \leq \mathbb{E}\left[\int_{\mathbb{R}} |\hat{p}(y|\mathbf{X}) - p(y|\mathbf{X})| \, \mathrm{d}y\right] + 2su \ .$$

Injecting this last inequality and (4.21) into (4.20) gives the announced result. □

Consistency Result

This section is devoted to the proof of Theorem 15. We first provide a result on the L_1-integrated estimation error of $\hat{p}$.

Proposition 16. *Under Assumption 12, we have that*

$$\mathbb{E}\left[\int_{\mathbb{R}} |\hat{p}(y|\mathbf{X}) - p(y|\mathbf{X})|\,\mathrm{d}y\right] \leq C\left(\sqrt{s}\mathbb{E}\left[(\hat{f}(\mathbf{X}) - f^*(\mathbf{X}))^2\right] + \mathbb{E}\left[\left|\hat{f}(\mathbf{X}) - f^*(\mathbf{X})\right|\right]\right) + Cs^{5/2}\mathbb{E}\left[|\hat{\sigma}^2(\mathbf{X}) - \sigma^2(\mathbf{X})|\right] \ ,$$

where $C > 0$ is a constant which depends on σ_0 and σ_1 in Assumption 12.

Proof. To build this proof, we use the triangle inequality to split the term $|\hat{p}(y|\mathbf{X}) - p(y|\mathbf{X})|$ into 3. We then have to consider each of these terms consecutively. The first of these terms can be bounded as follows:

$$\begin{aligned}\left|\frac{1}{\sqrt{2\pi\hat{\sigma}^2(\mathbf{X})}}\exp\left(-\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right) - \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\right| \\ \leq \frac{1}{\sqrt{2\pi\hat{\sigma}^2(\mathbf{X})}}\exp\left(-\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\left|1 - \frac{\hat{\sigma}(\mathbf{X})}{\sigma(\mathbf{X})}\right| \\ = \frac{1}{\sqrt{2\pi\hat{\sigma}^2(\mathbf{X})}}\exp\left(-\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\left|\frac{\sigma(\mathbf{X}) - \hat{\sigma}(\mathbf{X})}{\sigma(\mathbf{X})}\right| \ . \quad (4.22)\end{aligned}$$

This upper-bound consists of two parts. One part which is the density of a Gaussian random variable (whose integral *w.r.t.* y is 1) and a second term which is independent of y. Observe that this second term $|\sigma(\mathbf{X}) - \hat{\sigma}(\mathbf{X})|$ is of the same order as $|\sigma^2(\mathbf{X}) - \hat{\sigma}^2(\mathbf{X})|$. Indeed, notice that when $\sigma(\mathbf{X}) > \hat{\sigma}(\mathbf{X})$

$$\sigma^2(\mathbf{X}) - \hat{\sigma}^2(\mathbf{X}) = (\sigma(\mathbf{X}) - \hat{\sigma}(\mathbf{X}))(\sigma(\mathbf{X}) + \hat{\sigma}(\mathbf{X})) \geq \left(\sigma_0 + \frac{1}{\sqrt{s}}\right)(\sigma(\mathbf{X}) - \hat{\sigma}(\mathbf{X})) \ ,$$

where in the last inequality, we use Assumption 12 and the fact that $\hat{\sigma}(X) \geq 1/\sqrt{s}$. Written differently, this means that

$$|\sigma(\mathbf{X}) - \hat{\sigma}(\mathbf{X})| \leq \frac{\sqrt{s}}{1+\sigma_0\sqrt{s}}|\sigma^2(\mathbf{X}) - \hat{\sigma}^2(\mathbf{X})| \leq \frac{\sigma_0\sqrt{s}}{1+\sigma_0\sqrt{s}}\frac{|\sigma^2(\mathbf{X}) - \hat{\sigma}^2(\mathbf{X})|}{\sigma_0} \leq C|\sigma^2(\mathbf{X}) - \hat{\sigma}^2(\mathbf{X})| \ ,$$

since $1/\sigma_0 \leq C$. The same reasoning holds in the case where $\sigma(\mathbf{X}) < \hat{\sigma}(\mathbf{X})$ and then we conclude that

$$|\sigma(\mathbf{X}) - \hat{\sigma}(\mathbf{X})| \leq C|\sigma^2(\mathbf{X}) - \hat{\sigma}^2(\mathbf{X})| \ .$$

Injecting this bound into (4.22) and using again Assumption 12, we deduce that

$$\int_{\mathbb{R}}\left|\frac{1}{\sqrt{2\pi\hat{\sigma}^2(\mathbf{X})}}\exp\left(-\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right) - \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\right| dy \leq C\left|\sigma(\mathbf{X}) - \hat{\sigma}(\mathbf{X})\right| \leq C\left|\sigma^2(\mathbf{X}) - \hat{\sigma}^2(\mathbf{X})\right| \ . \quad (4.23)$$

Let us now consider the second term in the decomposition of $|\hat{p}(y|\mathbf{X}) - p(y|\mathbf{X})|$. Since $x \mapsto \exp(-x)$

is 1-Lipschitz on $\mathbb{R}_+$, from Assumption 12 we have that in the case where $(y-\hat{f}(\mathbf{X}))^2 \geq (y-f^*(\mathbf{X}))^2$

$$\begin{aligned}
&\left|\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)-\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\right| \\
&\quad=\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\left|\exp\left(-\left(\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\right)-1\right| \\
&\quad\leq\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}\times 2\hat{\sigma}^2(\mathbf{X})}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\left|(y-\hat{f}(\mathbf{X}))^2-(y-f^*(\mathbf{X}))^2\right| \\
&\quad\leq\frac{C}{\sqrt{2\pi\hat{\sigma}^2(\mathbf{X})}\hat{\sigma}(\mathbf{X})}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\left|(y-\hat{f}(\mathbf{X}))^2-(y-f^*(\mathbf{X}))^2\right| \ .
\end{aligned}$$

Using the following decomposition

$$(y-\hat{f}(\mathbf{X}))^2-(y-f^*(\mathbf{X}))^2=(\hat{f}(\mathbf{X})-f^*(\mathbf{X}))^2+2(y-f^*(\mathbf{X})(f^*(\mathbf{X})-\hat{f}(\mathbf{X})) \ ,$$

we deduce

$$\begin{aligned}
&\left|\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)-\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\right| \\
&\leq\frac{C}{\sqrt{2\pi\hat{\sigma}^2(\mathbf{X})}\hat{\sigma}(\mathbf{X})}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\left(\left(\hat{f}(\mathbf{X})-f^*(\mathbf{X})\right)^2+|y-f^*(\mathbf{X})||\hat{f}(\mathbf{X})-f^*(\mathbf{X})|\right) \ .
\end{aligned} \tag{4.24}$$

In the case where $(y-\hat{f}(\mathbf{X}))^2 \leq (y-f^*(\mathbf{X}))^2$, we obtain similar bound as in the above equation by switching the role of $\hat{f}$ by f^*. Notice that $\int_{\mathbb{R}}\frac{1}{\sqrt{2\pi\hat{\sigma}^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\times|y-f^*(\mathbf{X})|dy$ is the expectation of the r.v. $|Y-f^*(\mathbf{X})|$ where Y is Gaussian with expectation $f^*(\mathbf{X})$ and variance $\hat{\sigma}^2(\mathbf{X})$. Therefore, using the fact that $\mathbb{E}\left[|Z-\mathbb{E}\left[Z\right]|\right]\leq\sqrt{\mathrm{Var}(Z)}$ for any real valued random variable Z, we get

$$\begin{aligned}
&\int_{\mathbb{R}}\left|\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)-\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\right|dy \\
&\qquad\leq\frac{C}{\hat{\sigma}(\mathbf{X})}\left(\left(\hat{f}(\mathbf{X})-f^*(\mathbf{X})\right)^2+\hat{\sigma}(\mathbf{X})|\hat{f}(\mathbf{X})-f^*(\mathbf{X})|\right) \ .
\end{aligned}$$

Finally, using that $\hat{\sigma}(\mathbf{X})\geq 1/\sqrt{s}$, we deduce

$$\begin{aligned}
&\int_{\mathbb{R}}\left|\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)-\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\right|dy \\
&\qquad\leq C\left(\sqrt{s}\left(\hat{f}(\mathbf{X})-f^*(\mathbf{X})\right)^2+|\hat{f}(\mathbf{X})-f^*(\mathbf{X})|\right) \ .
\end{aligned} \tag{4.25}$$

The remaining term in the decomposition of $|\hat{p}(y|\mathbf{X})-p(y|\mathbf{X})|$ is

$$\begin{aligned}
&\left|\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)-\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\sigma^2(\mathbf{X})}\right)\right| \\
&\quad=\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\sigma^2(\mathbf{X})}\right)\left|\exp\left(-\left(\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}-\frac{(y-f^*(\mathbf{X}))^2}{2\sigma^2(\mathbf{X})}\right)\right)-1\right| \ .
\end{aligned}$$

Hence, if $\sigma^2(\mathbf{X}) \geq \hat{\sigma}^2(\mathbf{X})$, since $x \mapsto \exp(-x)$ is 1-Lipschitz on $\mathbb{R}_+$, we deduce from the above inequality that

$$\left| \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right) - \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\sigma^2(\mathbf{X})}\right) \right| \leq \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\sigma^2(\mathbf{X})}\right) \left| \frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})} - \frac{(y-f^*(\mathbf{X}))^2}{2\sigma^2(\mathbf{X})} \right| .$$

Therefore, from Assumption 12, and since $\hat{\sigma}^2(\mathbf{X}) \geq 1/s$, we get in the case where $\sigma^2(\mathbf{X}) \geq \hat{\sigma}^2(\mathbf{X})$

$$\left| \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right) - \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\sigma^2(\mathbf{X})}\right) \right| \leq \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\sigma^2(\mathbf{X})}\right) Cs(y-f^*(\mathbf{X}))^2 \left|\hat{\sigma}^2(\mathbf{X}) - \sigma^2(\mathbf{X})\right| . \quad (4.26)$$

In the case where $\sigma^2(\mathbf{X}) \leq \hat{\sigma}^2(\mathbf{X})$, using same arguments and additionally the fact that $\hat{\sigma}^2(\mathbf{X}) \leq s$, we can obtain

$$\left| \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right) - \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\sigma^2(\mathbf{X})}\right) \right| \leq \frac{\sqrt{s}}{\sigma_0\sqrt{2\pi\hat{\sigma}^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right) Cs(y-f^*(\mathbf{X}))^2 \left|\hat{\sigma}^2(\mathbf{X}) - \sigma^2(\mathbf{X})\right| . \quad (4.27)$$

Therefore, from Equation (4.26), and (4.27), we get

$$\int_{\mathbb{R}} \left| \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right) - \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\sigma^2(\mathbf{X})}\right) \right| dy \leq Cs^{5/2} \left|\hat{\sigma}^2(\mathbf{X}) - \sigma^2(\mathbf{X})\right| , \quad (4.28)$$

where we used the fact that the integral *w.r.t.* y is the variance of Gaussian r.v. with variance $\hat{\sigma}^2(\mathbf{X})$ and is then such that $\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\hat{\sigma}^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right) (y-f^*(\mathbf{X}))^2 dy = \hat{\sigma}^2(\mathbf{X}) \leq s$. The combination of Equations (4.23), (4.25), and (4.28) yields the result. □

Now, we provide the proof of Theorem 15.

Proof of Theorem 15. We prove the consistency $\hat{\Gamma}$ *w.r.t.* the symmetric difference distance $\mathcal{H}$. We have that

$$\mathcal{H}(\hat{\Gamma}) \leq \mathbb{E}\left[\int_{\hat{\Gamma}(\mathbf{X},\zeta)\triangle\bar{\Gamma}(\mathbf{X},\zeta)} dy\right] + \mathbb{E}\left[\int_{\bar{\Gamma}(\mathbf{X},\zeta)\triangle\Gamma^*(\mathbf{X})} dy\right] . \quad (4.29)$$

We bound the first term in the *r.h.s.* in the above inequality.

$$\mathbb{E}\left[\int_{\hat{\Gamma}(\mathbf{X},\zeta)\triangle\bar{\Gamma}(\mathbf{X},\zeta)} dy\right] = \mathbb{E}\left[\int_{\mathbb{R}} \left|\mathbb{1}_{\{y\in\hat{\Gamma}(\mathbf{X},\zeta)\}} - \mathbb{1}_{\{y\in\bar{\Gamma}(\mathbf{X},\zeta)\}}\right|\right] = \mathbb{E}\left[\int_{\mathbb{R}} \left|\mathbb{1}_{\{\hat{G}(\hat{p}(y|\mathbf{X},\zeta))\leq\ell\}} - \mathbb{1}_{\{\bar{G}(\hat{p}(y|\mathbf{X},\zeta))\leq\ell\}}\right| dy\right] .$$

Therefore, from Equations (4.13) and (4.19), we deduce

$$\mathbb{E}\left[\int_{\hat{\Gamma}(\mathbf{X},\zeta)\triangle\bar{\Gamma}(\mathbf{X},\zeta)} dy\right] \leq \frac{Cs}{\sqrt{N}} . \quad (4.30)$$

Now, we study the second term in the *r.h.s.* of Equation (4.29). We observe that if $y \in \bar{\Gamma}(\mathbf{X},\zeta) \setminus \Gamma^*_\ell(\mathbf{X})$ the following holds

- on the event $\{\bar{G}^{-1}(\ell) \geq G^{-1}(\ell)\}$, $|\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \geq \left|p(y|\mathbf{X}) - G^{-1}(\ell)\right|$,
- on the event $\{\bar{G}^{-1}(\ell) < G^{-1}(\ell)\}$,

$$\text{either } \ \hat{p}(y|\mathbf{X},\zeta) \in (\bar{G}^{-1}(\ell), G^{-1}(\ell)) \ \text{ or } \ |\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \geq \left|p(y|\mathbf{X}) - G^{-1}(\ell)\right| \ .$$

Note that similar reasoning holds if $y \in \Gamma^*_\ell(\mathbf{X}) \setminus \bar{\Gamma}(\mathbf{X},\zeta)$. Therefore, we deduce that conditional on $\mathcal{D}_n$,

$$\begin{aligned}\mathbb{E}\left[\int_{\bar{\Gamma}(\mathbf{X},\zeta)\triangle\Gamma^*(\mathbf{X})} dy\right] \leq \mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\{|\hat{p}(y|\mathbf{X},\zeta)-p(y|\mathbf{X})|\geq|p(y|\mathbf{X})-G^{-1}(\ell)|\}} dy\right] \\ + \mathbb{1}_{\{\bar{G}^{-1}(\ell)<G^{-1}(\ell)\}}\mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\{\hat{p}(y|\mathbf{X},\zeta)\in(\bar{G}^{-1}(\ell),G^{-1}(\ell))\}} dy\right] \\ + \mathbb{1}_{\{\bar{G}^{-1}(\ell)\geq G^{-1}(\ell)\}}\mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\{\hat{p}(y|\mathbf{X},\zeta)\in(G^{-1}(\ell),\bar{G}^{-1}(\ell))\}} dy\right] \ .\end{aligned}$$

Using first the definition of $\bar{G}$ and then the fact that $\bar{G}(\bar{G}^{-1}(\ell)) = G(G^{-1}(\ell)) = \ell$ in this last inequality, we deduce the following

$$\begin{aligned}\mathbb{E}\left[\int_{\bar{\Gamma}(\mathbf{X},\zeta)\triangle\Gamma^*(\mathbf{X})} dy\right] \leq \mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\{|\hat{p}(y|\mathbf{X},\zeta)-p(y|\mathbf{X})|\geq|p(y|\mathbf{X})-G^{-1}(\ell)|\}} dy\right] \\ + \mathbb{E}\left[\left|G(G^{-1}(\ell)) - \bar{G}(G^{-1}(\ell))\right|\right] \ .\end{aligned}$$

Now, we observe that

$$\begin{aligned}\mathbb{E}\left[\left|G(G^{-1}(\ell)) - \bar{G}(G^{-1}(\ell))\right|\right] &\leq \mathbb{E}\left[\int_{\mathbb{R}} \left|\mathbb{1}_{\{p(y|\mathbf{X})\geq G^{-1}(\ell)\}} - \mathbb{1}_{\{\hat{p}(y|\mathbf{X},\zeta)\geq G^{-1}(\ell)\}}\right| dy\right] \\ &\leq \mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\{|\hat{p}(y|\mathbf{X},\zeta)-p(y|\mathbf{X})|\geq|p(y|\mathbf{X})-G^{-1}(\ell)|\}} dy\right] \ .\end{aligned}$$

Therefore, we have obtained

$$\begin{aligned}\mathbb{E}\left[\int_{\bar{\Gamma}(\mathbf{X},\zeta)\triangle\Gamma^*(\mathbf{X})} dy\right] &\leq 2\mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\{|\hat{p}(y|\mathbf{X},\zeta)-p(y|\mathbf{X})|\geq|p(y|\mathbf{X})-G^{-1}(\ell)|\}} dy\right] \\ &\leq 2\int_{\mathbb{R}} \mathbb{P}\left(|\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \geq \left|p(y|\mathbf{X}) - G^{-1}(\ell)\right|\right) dy \ .\end{aligned} \tag{4.31}$$

Let us consider the term in the *r.h.s* of Equation (4.31). Let $\delta > 0$, we have that

$$\begin{aligned}\int_{\mathbb{R}} \mathbb{P}\left(|\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \geq \left|p(y|\mathbf{X}) - G^{-1}(\ell)\right|\right) dy \leq \int_{\mathbb{R}} \mathbb{P}\left(|\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \geq \delta\right) dy \\ + \int_{\mathbb{R}} \mathbb{P}\left(\left|p(y|\mathbf{X}) - G^{-1}(\ell)\right| \leq \delta\right) dy \ .\end{aligned}$$

From Markov's inequality, we deduce

$$\begin{aligned}\int_{\mathbb{R}} \mathbb{P}\left(|\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \geq \left|p(y|\mathbf{X}) - G^{-1}(\ell)\right|\right) dy \leq \frac{1}{\delta}\mathbb{E}\left[\int_{\mathbb{R}} |\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})|\, dy\right] \\ + G(G^{-1}(\ell) - \delta) - G(G^{-1}(\ell) + \delta) \ .\end{aligned} \tag{4.32}$$

Since $\hat{p}$ is supported on $[-s, s]$, we observe that

$$\begin{aligned}\mathbb{E}\left[\int_{\mathbb{R}} |\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})|\, \mathrm{d}y\right] &= \mathbb{E}\left[\int_{[-s,s]} |\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})|\, \mathrm{d}y\right] + \mathbb{E}\left[\int_{|y|\geq s}^{+\infty} p(y|\mathbf{X})\mathrm{d}y\right] \\ &\leq \mathbb{E}\left[\int_{[-s,s]} |\hat{p}(y|\mathbf{X}) - p(y|\mathbf{X})|\, \mathrm{d}y\right] + 2su + \mathbb{E}\left[\int_{|y|\geq s}^{+\infty} p(y|\mathbf{X})\mathrm{d}y\right] \ .\end{aligned} \tag{4.33}$$

Now, we observe that

$$\mathbb{E}\left[\int_s^{+\infty} p(y|\mathbf{X})\mathrm{d}y\right] = \mathbb{E}\left[\mathbb{1}_{\{|f^*(X)|\leq \frac{s}{2}\}}\int_s^{+\infty} p(y|\mathbf{X})\mathrm{d}y\right] + \mathbb{E}\left[\mathbb{1}_{\{|f^*(X)|> \frac{s}{2}\}}\int_s^{+\infty} p(y|\mathbf{X})\mathrm{d}y\right] \ .$$

From Markov inequality and Assumption 13, the second term of the r.h.s. in the above inequality is bounded by

$$\mathbb{E}\left[\mathbb{1}_{\{|f^*(\mathbf{X})|> \frac{s}{2}\}}\int_s^{+\infty} p(y|\mathbf{X})\mathrm{d}y\right] \leq \mathbb{E}\left[\mathbb{1}_{\{|f^*(\mathbf{X})|> \frac{s}{2}\}}\right] \leq \frac{2C_1}{s} \ .$$

On the other hand, we observe that

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{\{|f^*(\mathbf{X})|\leq \frac{s}{2}\}}\int_s^{+\infty} p(y|\mathbf{X})\mathrm{d}y\right] &\leq \mathbb{E}\left[\int_s^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-s/2)^2}{2\sigma^2(\mathbf{X})}\right)\mathrm{d}y\right] \\ &= \mathbb{E}\left[\int_{s/2}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{y^2}{2\sigma^2(\mathbf{X})}\right)\mathrm{d}y\right] \ .\end{aligned}$$

Therefore, Assumption 12 and standard result on Gaussian tails yields for $s \geq 1$

$$\mathbb{E}\left[\mathbb{1}_{\{|f^*(\mathbf{X})|\leq \frac{s}{2}\}}\int_s^{+\infty} p(y|\mathbf{X})\mathrm{d}y\right] \leq \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{s^2}{8\sigma_1^2}\right) \ .$$

Hence combining the above inequalities, we get for $s \geq 1$

$$\mathbb{E}\left[\int_s^{+\infty} p(y|\mathbf{X})\mathrm{d}y\right] \leq C'\left[\exp\left(-\frac{s^2}{8\sigma_1^2}\right) + \frac{C}{s}\right] \ ,$$

where C and C' are two positive constants. Note that similar arguments yields

$$\mathbb{E}\left[\int_{-\infty}^{-s} p(y|\mathbf{X})\mathrm{d}y\right] \leq C'\left[\exp\left(-\frac{s^2}{8\sigma_1^2}\right) + \frac{C}{s}\right] \ .$$

Therefore considering Equation (4.33) and Proposition 16 and defining $s = \log(\min(n, N))$, we get

$$\lim_n \frac{1}{\delta}\mathbb{E}\left[\int_{\mathbb{R}} |\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})|\, dy\right] = 0 \ .$$

Hence we obtain from Equation (4.32) that for all $\delta > 0$

$$\limsup_n \int_{\mathbb{R}} \mathbb{P}\left(|\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \geq \left|p(y|\mathbf{X}) - G^{-1}(\ell)\right|\right) dy \leq G(G^{-1}(\ell) - \delta) - G(G^{-1}(\ell) + \delta) \ .$$

Since G is continuous, with $\delta \to 0$, we get

$$\lim_n \int_{\mathbb{R}} \mathbb{P}\left(|\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \geq \left|p(y|\mathbf{X}) - G^{-1}(\ell)\right|\right) dy = 0 \ .$$

The above equation together with Equation (4.29), (4.30), (4.31) yields the desired result. □

Rates of convergence

We start this section with a result on the estimation error of $\hat{p}$ *w.r.t.* the sup-norm.

Proposition 17. *Let $s = \log(\min(n, N))$. Under Assumptions 12, 15, and 16, we have that*

$$\sup_{(\mathbf{x},y)\in\mathcal{C}\times[-s,s]} |\hat{p}(y|\mathbf{x}) - p(y|\mathbf{x})| \leq$$
$$C\left(s\sup_{\mathbf{x}\in\mathcal{C}}\left(\hat{f}(\mathbf{x}) - f^*(\mathbf{x})\right)^2 + s^2\sup_{\mathbf{x}\in\mathcal{C}}\left|\hat{f}(\mathbf{x}) - f^*(\mathbf{x})\right| + s^3\sup_{\mathbf{x}\in\mathcal{C}}\left|\hat{\sigma}^2(\mathbf{x}) - \sigma^2(\mathbf{x})\right|\right) \ ,$$

where $C > 0$ is a constant which depends on f^*, σ^2, and on the set $\mathcal{C}$.

Proof. We consider the same decomposition into 3 that we used in the proof of Proposition 16. Using the fact that $\hat{\sigma}(\mathbf{x}) \geq \frac{1}{\sqrt{s}}$ and Assumption 12, we get for all $\mathbf{x} \in \mathcal{C}$, and $y \in [-s, s]$ (*c.f.*, Eq. (4.22)), the first term is controlled as follows:

$$\left| \frac{1}{\sqrt{2\pi\hat{\sigma}^2(\mathbf{x})}} \exp\left(-\frac{(y - \hat{f}(\mathbf{x}))^2}{2\hat{\sigma}^2(\mathbf{x})} \right) - \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{x})}} \exp\left(-\frac{(y - \hat{f}(\mathbf{x}))^2}{2\hat{\sigma}^2(\mathbf{x})} \right) \right| \leq Cs \sup_{x \in \mathcal{C}} \left| \hat{\sigma}^2(\mathbf{x}) - \sigma^2(\mathbf{x}) \right| \ . \tag{4.34}$$

According to the second term, from Assumptions 15 and 16, and since f^* is a Lipschitz function on the compact $\mathcal{C}$, we have that $|f^*(\mathbf{x})| \leq s$ for n, N large enough. Therefore, using the fact that $x \mapsto \exp(-x)$ is 1-Lipschitz on $\mathbb{R}_+$ and that $\frac{1}{s} \leq \hat{\sigma}^2(\mathbf{x})$, we get (*c.f.*, Eq. (4.24))

$$\left| \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{x})}} \exp\left(-\frac{(y - \hat{f}(\mathbf{x}))^2}{2\hat{\sigma}^2(\mathbf{x})} \right) - \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{x})}} \exp\left(-\frac{(y - f^*(\mathbf{x}))^2}{2\hat{\sigma}^2(\mathbf{x})} \right) \right| \leq C \left(s \sup_{\mathbf{x} \in \mathcal{C}} \left(\hat{f}(\mathbf{x}) - f^*(\mathbf{x}) \right)^2 + s^2 \sup_{\mathbf{x} \in \mathcal{C}} \left| \hat{f}(\mathbf{x}) - f^*(\mathbf{x}) \right| \right) \ . \tag{4.35}$$

Finally, considering the last term, we deduce from (4.26) and (4.27) that

$$\left| \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{x})}} \exp\left(-\frac{(y - f^*(\mathbf{x}))^2}{2\hat{\sigma}^2(\mathbf{x})} \right) - \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{x})}} \exp\left(-\frac{(y - f^*(\mathbf{x}))^2}{2\sigma^2(\mathbf{x})} \right) \right| \leq Cs^3 \sup_{\mathbf{x} \in \mathcal{C}} \left| \hat{\sigma}^2(\mathbf{x}) - \sigma^2(\mathbf{x}) \right| \ . \tag{4.36}$$

The combination of Equations (4.34), (4.35), and (4.36) gives the proposition. □

Proof of Proposition 14. We recall that

$$\mathcal{E}_\ell\left(\bar{\Gamma}\right) = \mathbb{E}\left[\int_{\bar{\Gamma}(\mathbf{X},\zeta) \triangle \Gamma_\ell^*(\mathbf{X})} |p(y|\mathbf{X}) - \lambda_\ell^*| \, dy \right] \ .$$

Now, we observe that for $y \in \bar{\Gamma}(\mathbf{X}, \zeta) \triangle \Gamma_\ell^*(\mathbf{X})$

$$|p(y|\mathbf{X}) - \lambda_\ell^*| \leq |\hat{p}(y|\mathbf{X}, \zeta) - p(y|\mathbf{X}) \,| + \left| \bar{\lambda}_\ell - \lambda_\ell^* \right| \ ,$$

where we recall that $\bar{\lambda}_\ell := \bar{G}^{-1}(\ell)$, with $\bar{G}$ defined in Eq. (4.12). Using similar arguments as those used in the proof of Theorem 4.4 in [24] that is inspired by Theorem 2.12 in [11], it is not difficult to see that conditional on $\mathcal{D}_n$,

$$\left| \bar{\lambda}_\ell - \lambda_\ell^* \right| \leq \sup_{(\mathbf{x},y) \in \mathcal{C} \times \mathbb{R}} |\hat{p}(y|\mathbf{x}) - p(y|\mathbf{x})| + u := \hat{m}(u) \ .$$

Therefore, we deduce that

$$\mathbb{E}\left[\int_{\bar{\Gamma}(\mathbf{X},\zeta) \triangle \Gamma_\ell^*(\mathbf{X})} |p(y|\mathbf{X}) - \lambda_\ell^*| \, dy \right] \leq 2\hat{m}(u) \mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\{|p(y|\mathbf{X}) - \lambda_\ell^*| \leq 2\hat{m}(u)\}} \mathrm{d}y \right] \ .$$

Hence from the above inequality, and Assumption 17 we get

$$\mathbb{E}\left[\mathcal{E}_\ell\left(\bar{\Gamma}\right) \right] \leq 2^{1+\alpha} c_0 \mathbb{E}\left[\hat{m}(u)^{1+\alpha} \right] \ .$$

Therefore, from Equations (4.20) and (4.21), we obtain the following with $s = \log(\min(n, N))$

$$\mathbb{E}\left[\mathcal{E}_\ell\left(\hat{\Gamma}\right) \right] \leq C \left(\mathbb{E}\left[\left(\sup_{(x,y) \in \mathbb{R} \times \mathcal{C}} |\hat{p}(y|\mathbf{x}) - p(y|\mathbf{x})| \right)^{1+\alpha} \right] + u^{1+\alpha} + \frac{\log(N)}{N} \right) \ .$$

Finally, since $\hat{p}$ is supported on $[-s, s]$, we have

$$\sup_{(\mathbf{x},y) \in \mathcal{C} \times \mathbb{R}} |\hat{p}(y|\mathbf{x}) - p(y|\mathbf{x})| \leq \sup_{(\mathbf{x},y) \in \mathcal{C} \times [-s,s]} |\hat{p}(y|\mathbf{x}) - p(y|\mathbf{x})| + \sup_{(\mathbf{x},y) \in \mathcal{C} \times \mathbb{R} \setminus [-s,s]} p(y|\mathbf{x}) \ .$$

For n, N large enough, we can assume, since f^* is bounded, that $|f^*(\mathbf{X})| \leq s/2$. From Assumption 12, we have for n, N large enough

$$\sup_{(\mathbf{x},y)\in\mathcal{C}\times\mathbb{R}\setminus[-s,s]} p(y|\mathbf{x}) \leq C\exp\left(-\frac{s^2}{8\sigma_1^2}\right) \leq \exp(-s) \leq \frac{C}{\min(n,N)} \quad ,$$

which yields the desired result.

□

Proof of Theorem 16. The proof is a straightforward application of Proposition 14, 17, and Theorem 17, where we also use the fact that for n, N large enough

$$\left|\hat{\sigma}^2(\mathbf{x}) - \sigma^2(\mathbf{x})\right| \leq \left|\tilde{\sigma}^2(\mathbf{x}) - \sigma^2(\mathbf{x})\right| \quad .$$

□

Bibliography

[1] T.G. Anderson and J. Lund. Estimating continuous-time stochastic volatility models of the short-term interest rate. *Journal of Econometrics*, 77(2):343–377, 1997. 16, 36

[2] J.-Y. Audibert and A. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35(2):608–633, 2007. 8, 26, 64, 68, 79, 86, 96, 105

[3] J.Y. Audibert. Aggregated estimators and empirical complexity for least square regression. *Ann. Inst. H. Poincaré Probab. Statist.*, 40(6):685–736, 2004. 39

[4] J.Y. Audibert. Robust linear least squares regression. *Annals of Statistics*, 37(4):1591–1646, 2009. 37, 38

[5] P. Bartlett and H. Wegkamp. Classification with a Reject Option using a Hinge Loss . *Journal of Machine Learning Research*, 9:1823–1840, 2008. 64

[6] A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.*, 81(2):608–650, 2014. 102

[7] R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *Ann. Statist.*, 41(2):802–837, 2013. 90

[8] G. Biau and L. Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101(10):2499–2518, 2010. 7, 14

[9] G. Biau and L. Devroye. *Lectures on the Nearest Neighbor Method*. Springer Series in the Data Sciences. Springer New York, 2015. 5, 6, 7, 36, 68, 81, 84, 85

[10] S. A. Billings and K.L. Lee. Time series prediction using support vector machines, the orthogonal and the regularized orthogonal least- squares algorithms. *International Journal of Systems Science*, 33(10):811–821, 2002. 9

[11] S. Bobkov and M. Ledoux. One-dimensional empirical measures, order statistics and Kantorovich transport distances. 2016. to appear in the Memoirs of the American Mathematical Society. 79, 86, 113

[12] L. Breiman. Bagging predictors . *Machine Learning*, 24:123–140, 1996. 12, 14

[13] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. 11, 14

[14] L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor Francis, 1984. 11

[15] L.D. Brown and M. Levine. Variance estimation in nonparametric regression via the difference sequence method. *Annals of statistics*, 35(5):2219–2232, 2007. 22, 36

[16] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for gaussian regression. *Annals of Statistics*, 35(4):1674–1697, 2007. 14, 15, 37, 38, 39, 42

[17] T. Cai, M Levine, and L. Wang. Variance function estimation in multivariate nonparametric regression. *Journal of Multivariate Analysis*, 100(1):126–136, 2009. 22

[18] C. Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, (4):247–254, 1957. 22, 64

[19] C. Chow. On optimum error and reject trade-off. *IEEE Transactions on Information Theory*, 16:41–46, 1970. 22, 23, 64

[20] E. Chzhen, C. Denis, and M. Hebiri. Minimax semi-supervised set-valued approach to multi-class classification. *Bernoulli*, 2021. 91, 92, 97

[21] C. Cortes, G. DeSalvo, and M. Mohri. Boosting with abstention. *In Advances in Neural Information Processing Systems*, pages 1660–1668, 2016. 22

[22] C. Denis and M. Hebiri. Confidence sets with expected sizes for multiclass classification. *J. Mach. Learn. Res.*, 18(102):1–28, 2017. 91, 93

[23] C. Denis and M. Hebiri. Consistency of plug-in confidence sets for classification in semi-supervised learning. *Journal of Nonparametric Statistics*, 2019. 22, 23, 30, 64, 66, 68, 86

[24] C. Denis, M. Hebiri, and A. Zaoui. Regression with reject option and application to knn. NeurIPS, 2020. iii, 36, 37, 96, 97, 103, 113

[25] C. Denis, M. Hebiri, and A. Zaoui. Prediction intervals with controlled length in the heteroscedastic gaussian regression. Preprint, 2022. iii

[26] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996. 84

[27] S.S. Durbha, R.L. King, and N.H. Younan. Support vector machines regression for retrieval of leaf area index from multiangle imaging spectroradiometer. *Remote Sensing of Environment*, 107(1):348–361, 2007. 9

[28] J. Fan. Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 87(420):998–1004, 1992. 16, 36

[29] J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London., 1996. 16

[30] J. Fan and Q. Yao. Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 85(3):645–660, 1998. 17, 20, 22, 37, 68

[31] Y. Freund. Boosting a weak learning algorithm by majority . *Information and Computation*, 121(2):256–285, 1995. 14

[32] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2010. 44

[33] G. Fumera and F. Roli. Support vector machines with embedded reject option. *In Pattern recognition with support vector machines*, pages 68–82, 2002. 22

[34] R. Genuer and J.-M Poggi. Arbres CART et Forêts aléatoires, Importance et sélection de variables. *Dans Apprentissage Statistique et Données Massives, Maumy-Bertrand M., Saporta G. et Thomas Agnan C. (eds), Technip*, pages 295–342, 2018. 11

[35] S. Gey. Bornes de risque, détection de ruptures, boosting : trois thèmes statistiques autour de cart en régression. In *PhD thesis, Paris 11, Orsay*, 2002. 11

[36] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002. 5, 6, 7, 8, 9, 26, 36, 68, 69, 81, 82, 85, 96, 97

[37] P. Hall and R.J. Carroll. Variance function estimation in regression: the effect of estimating the mean. *Journal of the Royal Statistical Society. Series B (Methodological)*, 51(1):3–14, 1989. 36, 67, 68, 100

[38] W. Härdle and A.B. Tsybakov. Local polynomial estimators of the volatility function in nonparametric autoregression. *Journal of Econometrics*, 81(1):223–242, 1997. 17, 36, 37, 68

[39] J. Hart. *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer Series in Statistics. 1997. 16

[40] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2009. 9, 12

[41] M.E. Hellman. The nearest neighbor classification rule with a reject option . *IEEE Transactions on Systems Science and Cybernetics*, 6(3):179–185, 1970. 22

[42] R. Herbei and M. Wegkamp. Classification with reject option. *The Canadian Journal of Statistics*, 34(4):709–721, 2006. 22, 23, 64, 65

[43] A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric regression. *Annals of Statistics*, 28(3):681–712, 2000. 14, 15, 37, 38

[44] R. Koenker. *Quantile Regression*. Cambridge University Press, 2005. 36, 47

[45] L.H. Koopmans and Clifford Qualls. Fixed length confidence intervals for parameters of the normal distribution based on two-stage sampling procedures. *Rocky Mountain J. Math.*, 1(4):587–602, 1971. 91

[46] R. Kulik and C. Wichelhaus. Nonparametric conditional variance and error density estimation in regression models with dependent errors and predictors. *Electron. J. Statist.*, 5:856–898, 2011. 20, 37, 68

[47] G. Lecué. Empirical risk minimization is optimal for the convex aggregation problem. *Bernoulli*, 19(5B):2153–2166, 2013. 15, 42

[48] G. Lecué and S. Mendelson. Aggregation via empirical risk minimization. *Probability theory and related fields*, 145(3-4):591–613, 2009. 15, 42

[49] J. Lei. Classification with confidence. *Biometrika*, 101(4):755–769, 2014. 23, 64

[50] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018. 24, 25, 90, 91

[51] J. Lei, J. Robins, and L. Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013. 24

[52] J. Lei and L. Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society, Series B*, 76(1):71–96, 2014. 16, 24, 25, 90, 91, 98, 99

[53] S. Li. Fnn: Fast nearest neighbor search algorithms and applications. 2019. 44

[54] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2:18–22, 2002. 44

[55] Z. Lin, S. Trivedi, and J. Sun. Locally valid and discriminative prediction intervals for deep learning models. NeurIPS, 2020. 24, 25

[56] C. Loader. *Local Regression and Likelihood*. New York: Springer, 1999. 16

[57] S. Lu, Y. Liu, L. Yin, and K. Zhang. Confidence intervals and regions for the lasso by using stochastic variational inequality techniques in optimization. *Journal of the Royal Statistical Society Series B*, 79(2):589–611, 2017. 102

[58] S. Mallat. *A wavelet tour of signal processing*. Academic Press, San Diego. Third edition, 2008. 12

[59] E. Mammen, J.P. Nielsen, M. Scholz, and S. Sperlich. Conditional variance forecasts for long-term stock returns. *Machine learning in insurance*, 7(4), 2019. 16, 36

[60] E. Mammen and A. Tsybakov. Smooth discrimination analysis. *Annals of Statistics*, 27(6):1808–1829, 1999. 68

[61] P. Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *Ann. Probab.*, 18(3):1269–1283, 1990. 68, 78

[62] J. Minnier, L. Tian, and T. Cai. A perturbation method for inference on regularized regression estimates. *J. Amer. Statist. Assoc.*, 106(496):1371–1382, 2011. 102

[63] H.G. Müller and U. Stadtmüller. Estimation of heteroscedasticity in regression analysis. *The Annals of Statistics*, 15(2):610–625, 1987. 16, 21, 22, 36

[64] M. Naadeem, J.D. Zucker, and B. Hanczar. Accuracy-rejection curves (ARCs) for comparing classification methods with a reject option. In *MLSB*, pages 65–81, 2010. 64

[65] E.A. Nadaraya. On estimating regression. *Theory of Probability and Its Applications*, 9:141–142, 1964. 8

[66] E.A. Nadaraya. Remarks on nonparametric estimates for density functions and regression curves. *Theory of Probability and Its Applications*, 15:134–137, 1970. 8

[67] M.H. Neumann. Fully data-driven nonparametric variance estimators. *Statistics*, 25:189–212, 1994. 37

[68] J.D. Opsomer, D. Ruppert, M.P Wand, U. Holst, and O. Hossjer. Kriging with nonparametric variance function estimation. *Biometrics*, 55(3):704–710, 1999. 36

[69] A. Pagan and A. Ullah. *Nonparametric Econometrics*. Cambridge University Press, Cambridge, 1999. 9

[70] E. Parzen. On the estimation of a probability density function and the mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962. 8

[71] W. Polonik. Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *Annals of Statistics*, 23(3):855–881, 1995. 64, 68

[72] B. Ripley. tree: Classification and regression trees. 2019. 44

[73] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956. 8

[74] D. Ruppert, M.P. Wand, U. Holst, and O. HöSJER. Local polynomial variance function estimation. *Technometrics*, 39(3):262–273, 1997. 36, 37

[75] R.E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990. 12

[76] E. Scornet. On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146:72–83, 2016. 11, 12

[77] E. Scornet. Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62:1485–1500, 2016. 11, 12

[78] E. Scornet, G. Biau, and J.-P. Vert. Consistency of random forests. *Annals of Statistics*, 43(4):1716–1741, 08 2015. 5, 6, 11, 12, 36, 68

[79] G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008. 24, 25

[80] K. Shan and Y. Yang. Combining regression quantile estimators. *Statistica Sinica*, 19(3):1171–1191, 2009. 36, 47

[81] Y. Shen, G. Gao, D. Witten, and F. Han. Optimal estimation of variance in nonparametric regression with random design. 2019. 68

[82] C. Stone. Consistent nonparametric regression. *Annals of Statistics*, pages 595–620, 1977. 5, 6, 36, 68

[83] I Takeuchi, Q.v. Le, T.D. Sears, and A.J. Smola. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7(45):1231–1264, 2006. 47, 48

[84] T. Tony Cai, M. G. Low, and Y. Xia. Adaptive confidence intervals for regression functions under shape constraints. *The Annals of Statistics*, 41(2):722–750, 2013. 24

[85] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004. 96

[86] A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2008. 5, 6, 9, 26, 36, 68

[87] A.B. Tsybakov. Optimal rates of aggregation. *Learning Theory and Kernel Machines*, pages 303–313, 2003. 14, 15, 37, 38, 39, 42

[88] A.B. Tsybakov. Aggregation and minimax optimality in high-dimensional estimation. *Proceedings of International Congress of Mathematicians*, 3:225–246, 2014. 15, 29, 37, 38, 55

[89] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 2014. 102

[90] A.W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998. 103

[91] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995. 9

[92] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. 81, 84

[93] N. Verzelen and E. Gassiat. Adaptive estimation of high-dimensional signal-to-noiseratios. *Bernoulli*, 24(4B):3683–3710, 2018. 16, 36

[94] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer, New York, 2005. 24, 25, 64, 90

[95] V. Vovk, I. Nouretdinov, and A. Gammerman. On-line predictive linear regression. *The Annals of Statistics*, 37(3):1566–1590, 2009. 24, 90, 91

[96] G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia, 1990. 12

[97] L. Wang, L. D.Brown, T.Tony Cai, and M. Levine. Effect of mean on variance function estimation in nonparametric regression. *Annals of Statistics*, 36(2):646–664, 2008. 22, 36

[98] G.S. Watson. Smooth regression analysis. *Sankhya, Series A*, pages 359–372, 1964. 8

[99] M. Wegkamp and M. Yuan. Support vector machines with a reject option. *Bernoulli*, 17(4):1368–1385, 2011. 22

[100] Y. Wiener and R. El-Yaniv. Pointwise tracking the optimal regression function. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2042–2050. Curran Associates, Inc., 2012. 23, 64

[101] K. Xu and P.C. B Phillips. Tilted nonparametric estimation of volatility functions with empirical applications. *Journal of Business & Economic Statistics*, 29(4):518–528, 2011. 21, 37

[102] K. Yu and M. Jones. Likelihood based-local linear estimation of the conditional variance function. *Journal of the American Statistical Association*, 99(465):139–144, 2004. 20, 37

[103] M. Yuan and M. Wegkamp. Classification methods with reject option based on convex risk minimization. *J. Mach. Learn. Res.*, 11:111–130, 2010. 22, 23

[104] Y. Yuhong. Aggregating regression procedures to improve performance. *Bernoulli*, 10(1):25–47, 2004. 15, 37, 38

[105] A. Zaoui. Variance function estimation in regression model via aggregation procedures. *Journal of Nonparametric Statistics*, 2022. iii, 71

[106] Flavio A. Ziegelmann. Nonparametric estimation of volatility functions: The local exponential estimator. *Econometric Theory*, 18:985–991, 2002. 20, 37