Hold-out and Aggregated hold-out

Guillaume Maillard

université
PARIS-SACLAY

Hold-out and Aggregated hold-out

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 547 Mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques appliquées
Unité de recherche: Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France.
Référent: Faculté des sciences d'Orsay

Thèse présentée et soutenue à Orsay, le 29/09/2020, par

Guillaume MAILLARD

Composition du jury:

Pascal Massart Professeur, Université Paris-Saclay	Président
Fabienne Comte Professeure, Université Paris-Descartes	Rapportrice & Examinatrice
Yuhong Yang Professeur, University of Minnesota	Rapporteur & Examinateur
Stéphane Boucheron Professeur, Université Paris-Diderot	Examinateur
Oleg Lepski Professeur, Université d'Aix-Marseille	Examinateur
Adrien Saumard Enseignant-Chercheur, ENSAI/CREST	Examinateur
Sylvain Arlot Professeur, Université Paris-Saclay	Directeur
Matthieu Lerasle Chargé de Recherche CNRS, ENSAE/CREST	Codirecteur

Thèse de doctorat

NNT: 2020UPASM005

Foreword

The subject of this dissertation is *selection* and *aggregation* of estimators, using the hold-out, also known as *simple validation*, and its aggregated version, called aggregated hold-out. The thesis includes articles that have been submitted to a journal (Chapters 3 and 4) and additional work that has not yet been made public (Chapters 5 and 6). Chapter 3 is joint work with my advisors, Sylvain Arlot and Matthieu Lerasle, while all other chapters are my own work. For the submitted material, I am still awaiting an answer from the publishers.

The contents of this thesis are laid out as follows.

- Chapter 2 is the Introduction. It presents the topic of this thesis, gives a state-of-the-art and situates the results of the thesis in their scientific context.
- Chapter 1 is a somewhat shorter introduction in French, more focused on the contents of the thesis.
- Chapter 3 consists of the article [73], which studies the hold-out and aggregated hold-out applied to kernel (RKHS) methods. It also contains a general definition of aggregated hold-out (Agghoo) and general theorems (Theorems 3.7.2 and 3.7.3) that are used throughout the manuscript.
- Chapter 4 consists of the article [72]. It studies the hold-out (and Agghoo) applied to sparse linear regression with a robust loss function.
- Chapter 5 is a detailed asymptotic analysis of the hold-out risk estimator applied to Fourier series estimators in least-squares regression.
- In the same setting as chapter 5, chapter 6 contains a precise analysis of the risks of the hold-out and its aggregated version (Agghoo).
- Chapter 7 concludes the thesis and gives some perspectives for future work on the hold-out and Agghoo.

The dependencies between the chapters are as follows. Chapter 3 states two general theorems (Theorems 3.7.2 and 3.7.3) that are used throughout the thesis. Apart

from these theorems, Chapters 3 and 4 can be read independently from each other and from Chapters 5 and 6. In contrast, chapter 6 uses the notation, assumptions and the main Theorem of Chapter 5, as well as some other results proved in Chapter 5. The two chapters should therefore be read in succession.

Remerciements

Le thésard devant explorer par lui-même un sujet de recherche est semblable à un voyageur s'aventurant sans carte en terre inconnue: il a besoin pour réussir que quelqu'un du coin lui indique les passages les plus praticables et lui évite de s'aventurer trop loin dans des voies sans issues.

Mes directeurs de thèse, Sylvain Arlot et Matthieu Lerasle, ont su jouer ce rôle à chaque fois que j'en avais besoin, et j'ai toujours senti que je pouvais compter sur leur avis.

Je tiens à les remercier de la bonne volonté dont ils ont fait preuve en relisant et corrigeant ligne à ligne mes premiers jets mals rédigés, autant de fois que nécessaire. Je les remercie aussi pour la patience avec laquelle ils ont supporté ou remédié à mon manque d'organisation pour les choses administratives.

Une fois terminée, la thèse a été relue par Yuhong Yang et Fabienne Comte, tâche dont j'étais bien placé pour savoir qu'elle ne serait pas des plus faciles. Je les remercie de l'avoir acceptée et de s'en être acquittés dans des délais très raisonnables.

Merci aussi à Yannick Barraud et à l'Université du Luxembourg de m'avoir accueilli pendant que je préparais ma soutenance.

En dehors du cadre strictement professionnel, je tiens à remercier tous ceux qui ont contribué à rendre agréables ces trois années de thèse, à l'université comme en dehors. La menée à bien de ce travail de longue haleine aurait été compromise si je n'avais bénéficié dans le même temps d'un cadre de vie stable, ainsi que de l'affection et du soutien moral de mes amis et de ma famille.

Il y a souvent eu au cours de cette thèse des moments de frustration et de doute: simulations qui ne marchent pas, obstacles imprévus dans une démonstration, etc. Heureusement, au bureau 2R21, on avait rarement l'occasion de se morfondre seul très longtemps. François et Cyril venaient régulièrement me distraire par leurs blagues qui dépassent les bornes dans plusieurs directions à la fois; Benjamin avait, lui, l'art de ponctuer ces propos par des silences et des rougissements éloquents. Je les remercie pour ces moments de franche rigolade, ainsi qu'Hugo pour la patience avec laquelle il nous a supportés. En plus des conversations, les échecs constituaient un des divertissements possibles en 2R21, et je remercie Eliot pour nous avoir

permis d'assister à ses parties, Hugo pour ses commentaires avisés et Vincent pour m'avoir fait jouer.

Au laboratoire, il régnait généralement une bonne ambiance entre doctorants, et je tiens notamment à en remercier Jeanne, qui a contribué à la cohésion du groupe en rassemblant et intégrant les nouveaux venus; Guillaume Lachaussée, pour sa capacité à lancer la conversation à table là où tous se seraient perdus dans leurs pensées; Martin, qui animait les conversations du midi par son esprit de contradiction; Louise, Camille et Gabriel, qui ont créé et animé un groupe WhatsApp de doctorants pour que l'on reste en contact pendant le confinement.

Quand j'ai voulu moi-même m'impliquer dans la vie sociale de l'Université en créant un club de lecture, de nombreux doctorants ont répondu à l'appel. Je souhaite remercier François, Vincent, Jean, Pierre-Louis, Camille, Pierre, Céline et Lucien d'avoir accepté de participer à ces séances, y compris en préparant des exposés.

En quittant l'université après une dure journée de réflexion, il était agréable de se poser pour bavarder autour d'un verre et d'un bon repas. Je remercie Adrien P., Samuel et Tran pour avoir si souvent répondu présent à de telles occasions. Merci aussi à Adrien K. et Assaf, qui m'ont souvent fait bénéficier de leurs talents de cuisinier et d'oenologue quand nous étions colocataires à Bourg-la-Reine.

Quand le besoin se faisait sentir de quitter le labyrinthe mental de mes recherches mathématiques, ainsi que le cadre familier d'Orsay, je pouvais compter sur Nicolas, Alice, Adrien K, Alexandre, Ségolène et Louis pour me faire prendre de grands bols d'air des montagnes. Merci en particulier à Nicolas pour la logistique, à Adrien et Alexandre pour la planification de nos expéditions, et à Louis pour le vin, la musique et la bonne humeur.

Enfin, je voudrais remercier ma famille, qui m'a toujours soutenu moralement et matériellement au cours de ces trois années de thèse. Merci à mes parents, à l'éducation desquels je dois d'être là où je suis, et chez qui j'étais toujours le bienvenu. Merci aussi à ma soeur Julia, qui m'a gentiment prêté son appartement au moment où je devais quitter le mien: c'est chez elle que j'ai eu l'idée qui devait mener au premier chapitre de cette thèse.

Contents

Chapter 1

Introduction et résumé (en Français)

Les statistiques paramétriques élémentaires sont conçues pour traiter des situations où un modèle connu, de dimension finie, est donné et où la taille de l'échantillon statistique est suffisamment grande pour qu'on puisse la considérer comme infinie (cadre asymptotique).

Les développements de l'informatique et la prolifération des données ont augmenté les possibilités et les ambitions des statistiques et ont conduit les statisticiens à considérer des problèmes plus généraux. Des problèmes de régression où, par exemple, il n'y a pas de "vrai modèle" unique et où la fonction de régression doit être cherchée dans une classe de fonctions vérifiant seulement certaines conditions de régularité (régression non-paramétrique), ou des problèmes de régression linéaire où le nombre de coordonnées n'est pas négligeable par rapport à la taille de l'échantillon (statistiques en grande dimension), si bien qu'il est impossible de faire de l'inférence statistique sur le modèle complet (de grande dimension ou de dimension infinie).

Ces situations obligent le statisticien à trouver un compromis entre la généralité du modèle statistique et la difficulté d'estimer ses paramètres. L'espoir est qu'un modèle simple suffise à décrire la réalité en première approximation, de sorte qu'une estimation statistique précise soit possible. Comme la complexité requise est inconnue en général, une pratique répandue consiste à introduire une famille de modèles de tailles et complexités variées. Le statisticien doit alors choisir un modèle, ou estimateur dans cette famille, de façon à s'adapter au mieux à la complexité intrinsèque du problème.

Pour des familles de modèles spécifiques, il est possible d'utiliser des méthodes ad-hoc reposant sur des calculs théoriques. Cependant, il y a des cas où cette approche ne fonctionne pas, soit parce que les calculs théoriques requis ne sont pas faisables, soit parce qu'ils font intervenir des quantités dépendant de la loi inconnue des données, telles que le niveau de bruit. Il est donc important de disposer de méthodes "boîte noire" qui ne nécessitent aucune information à priori,

ni sur la loi des données, ni même sur la famille d'estimateurs. De telles méthodes se fondent généralement sur la "validation", c'est-à-dire qu'elles réservent une partie de l'échantillon à l'estimation du risque des estimateurs afin d'accéder à une évaluation indépendante de leur risque. Ces estimées du risque peuvent être utilisées soit pour *sélectionner* un seul estimateur de la famille - souvent en minimisant le risque estimé - soit pour pondérer les différents estimateurs au sein d'une combinaison convexe.

Dans cette thèse, j'étudie une méthode (Agghoo) qui combine des éléments des deux approches. La subdivision des données en deux échantillons est faite plusieurs fois. Chaque fois, l'échantillon de validation est utilisé pour estimer le risque et sélectionner un estimateurs, entraîné sur les données restantes. A la fin, les différents estimateurs correspondant aux différentes subdivisions sont agrégés.

1.1 Cadre de l'apprentissage statistique

En statistique, il est très fréquent que l'on définisse des fonctions de risque afin de mesurer la qualité d'un estimateur. En conséquence, de nombreux problèmes statistiques peuvent s'exprimer comme la minimisation d'une fonction de risque sur un certain ensemble $\mathcal{S}$.

Agghoo, à l'instar de la validation croisée, s'applique aux problèmes de minimisation de risque pour lesquels le *risque* $R(\hat{s})$ d'un estimateur $\hat{s}$ peut s'exprimer comme une espérance $P\gamma(\hat{s}) = \mathbb{E}_Z[\gamma(\hat{s}, Z)]$, où γ est une fonction de contraste connue, Z suit la loi inconnue P et est indépendante de $\hat{s}$. Ce cadre inclus la *régression* et la *classification*. En régression, $\mathcal{S}$ est un espace de fonctions mesurables réelles, $Z = (X, Y)$ où Y est une variable aléatoire réelle et $\gamma(t, (x, y)) = \varphi(y - t(x))$, pour une fonction φ - souvent convexe - telle que $\varphi(0) = 0$. En classification, $\mathcal{S}$ est un ensemble de fonctions mesurables à valeur dans un ensemble fini $\mathcal{Y}$ de *labels*, les données sont des couples (X, Y) où $Y \in \mathcal{Y}$ est un *label*, et la fonction de contraste $\gamma(t, (x, y)) = \mathbb{I}_{t(x) \neq y}$ indique une erreur dans la *classification* de x par la fonction t, par rapport au label observé y.

Quand le risque provient d'un problème d'estimation, et qu'il mesure à quel point $\hat{s}$ s'approche d'un paramètre s de la loi sous-jacente P, le risque doit logiquement être positif et atteindre un minimum de 0 en la cible s. Il n'est parfois pas possible de trouver une fonction de contraste γ telle que $P\gamma(t) = R(t)$, mais seulement telle que $P\gamma(t) = R(t) + P\gamma(s)$, où $P\gamma(s)$ est une constante non-nulle qui dépend uniquement de s et P, si bien qu'il est équivalent de minimiser R ou $P\gamma(t)$. R s'exprime alors comme l' *excès de risque* relatif à l'optimum s:

$$\ell(s, t) = P\gamma(t) - P\gamma(s) = P\gamma(t) - \operatorname*{argmin}_{t' \in \mathcal{S}} P\gamma(t').$$

Un exemple de ce cas de figure est l' *estimation de densité* L^2. Dans ce problème, on observe des variables aléatoires i.i.d $(Z_i)_{1\leqslant i\leqslant n}$ distribuées selon une loi P ayant pour densité s par rapport à une mesure connue μ (souvent la mesure de Lebesgue), et l'objectif est de construire un estimateur $\hat{s}$ approchant s au sens de la distance L^2 au carré, $R(\hat{s}) = \|\hat{s} - s\|^2_{L^2(\mu)}$. En définissant $\gamma(t, Z) = \|t\|^2_{L^2(\mu)} - 2t(Z)$, on obtient $\mathbb{E}[\gamma(t, Z)] = \|t - s\|^2_{L^2(\mu)} - \|s\|^2_{L^2(\mu)}$ et $\ell(s, t) = \|t - s\|^2_{L^2(\mu)}$.

L'intérêt du formalisme présenté ci-dessus est qu'il permet de définir naturellement une large gamme d'estimateurs aux propriétés intéressantes. Supposons que l'on dispose d'un échantillon $Z_1, \ldots, Z_n$ de loi P. Comme le risque de $t \in \mathcal{S}$ s'exprime comme l'espérance $P\gamma(t) = E_Z[\gamma(t, Z)]$, il peut être estimé par la moyenne

$$P_n\gamma(t) = \frac{1}{n}\sum_{i=1}^{n} \gamma(t, Z_i).$$

Le but étant de minimiser le risque $P\gamma(t)$, une idée naturelle est de remplacer le risque $P\gamma$ par son approximation $P_n\gamma$ et de minimiser cet estimateur empirique du risque (*minimisation du risque empirique*). Dans le cadre non-paramétrique de l'apprentissage statistique, la classe $\mathcal{S}$ est trop grande pour que cette stratégie soit valable telle quelle: en régression par exemple, il existe une infinité de fonctions mesurables qui reproduisent exactement les données $(t(X_i) = Y_i)$ et ont donc un risque empirique nul: il est alors impossible de dire quoi que ce soit des valeurs de $t(x)$ pour x n'appartenant pas à l'ensemble fini $X_1, \ldots, X_n$.

Une option consiste à remplacer l'ensemble $\mathcal{S}$ par un sous-ensemble m appelé *modèle*, sur lequel la minimisation du risque empirique fonctionne: on obtient un estimateur

$$\hat{s}_m = \operatorname*{argmin}_{t\in m} P_n\gamma(t).$$

Par exemple, en régression des moindres carrés, m est généralement un espace vectoriel de fonctions, telles que des fonctions constantes par morceaux (régressogrames) ou polynômiales par morceaux, des ondelettes ou d'autres espaces classiques utilisés pour l'approximation de fonctions.

Une deuxième façon d'adapter la minimisation du risque empirique au cadre non-paramétrique est de pénaliser les éléments t trop grands ou trop complexes, en utilisant une pénalité $\Omega(t)$. Pour tout $\lambda > 0$, on obtient un estimateur

$$\hat{s}_\lambda = \operatorname*{argmin}_{t\in\mathcal{S}} \left\{P_n\gamma(t) + \lambda\Omega(t)\right\}.$$

λ est appelé *paramètre de régularisation*. Un exemple important de ce type de méthode est celui des *méthodes à noyaux*, où la pénalité est $\Omega(t) = \|t\|^2_{\mathcal{H}}$, pour un *espace de Hilbert à noyau reproduisant* $\mathcal{H}$ (voir Scholkopf et Smola [95]).

Les deux méthodes soulèvent des problèmes similaires. Quelle doit être la taille du modèle m ou du paramètre de régularisation λ? Dans le cas des modèles, le risque de l'estimateur $\hat{s}_m$ dépend de deux facteurs: l'erreur d'approximation induite par la restriction de $P\gamma$ de $\mathcal{S}$ à m, et l' *erreur d'estimation* due à l'estimation du risque $P\gamma$ par le risque empirique $P_n\gamma$ sur le modèle m. Plus le modèle m est grand, plus l'*erreur d'approximation* sera faible, mais plus l'*erreur d'estimation* risque d'être importante: on parle de *compromis biais variance*. Pour un ERM pénalisé, choisir λ trop grand biaise trop l'estimation en faveur des petites valeurs de $\Omega(t)$, l'erreur d'approximation $\operatorname{argmin}_t P\gamma(t) + \lambda\Omega(t) - \operatorname{argmin}_{t'} P\gamma(t')$ devient trop grande. Pour les méthodes à noyaux, $\hat{s}_\lambda \to 0$ quand $\lambda \to +\infty$. En revanche, prendre $\lambda = 0$ redonne l'ERM sur tout l'ensemble $\mathcal{S}$, qui ne converge pas. Pour que l'erreur d'approximation et l'erreur d'estimation convergent simultanément vers 0, il faut donc choisir une suite de modèles m de taille croissante, ou une suite de paramètres de régularisation qui tend vers 0, mais pas trop vite. Pour cette raison, les statisticiens considèrent généralement des familles de pénalités $(\lambda\Omega)_{\lambda>0}$ ou de modèles $m \in \mathcal{M}$. Etant donné un tel ensemble d'estimateurs, le choix optimal du paramètre de régularisation ou du modèle dépend de la "complexité" de la cible s, et notamment des considérations suivantes: quelle est la qualité de l'approximation de s par les modèles m? Quelle est la taille de $\Omega(s)$? Le problème devient alors de construire un estimateur $\hat{s}$ qui fasse aussi bien que le choix optimal théorique du modèle m ou du paramètre de régularisation λ (inconnu, car dépendant de P).

Dans un cadre asymptotique, ce critère peut être formalisé par l'équation

$$\frac{\ell(s,\hat{s})}{\inf_{m\in\mathcal{M}} \ell(s,\hat{s}_m)} \to 1,$$

où $\to$ désigne une notion de convergence classique (L^p, convergence en probabilité, presque sûrement, etc), une propriété appelée "consistence pour la sélection de modèles". Dans un cadre non-asymptotique, un objectif typique est de démontrer des *inégalités d'oracle* de la forme générale:

$$\mathbb{E}\ell(s,\hat{s}) \leqslant C\mathbb{E}\left[\inf_{m\in\mathcal{M}} \ell(s,\hat{s}_m)\right] + r_n, \tag{1.1}$$

où C est une constante et r_n un terme de reste "petit" (dans l'idéal négligeable par rapport aux autres termes de l'équation). L'estimateur $\hat{s}_{\hat{m}_*}$ qui réalise l'infimum dans l'équation (1.1) est appelé *oracle*. C'est le meilleur estimateur dans la famille $(\hat{s}_m)_{m\in\mathcal{M}}$.

1.2 Hold-out

Afin de construire un estimateur $\hat{s}$ qui vérifie une inégalité d'oracle, une première catégorie de méthodes consiste à estimer le mieux possible le risque des estima-

teurs $\hat{s}_m$ afin de sélectionner l'un d'entre eux dont le risque estimé est minimal: c'est la *sélection de modèles*. L'estimateur du risque empirique sur l'échantillon d'entraînement ne peut pas être utilisé, car il est beaucoup trop biaisé: en effet, dans le cas des pénalités, il est facile de voir que minimiser le risque empirique par rapport à λ revient en fait à sélectionner la plus petite valeur de λ. Si un *échantillon de validation* $Z_1, \ldots, Z_{n_v}$ est donné, et qu'il est indépendant des estimateurs $\hat{s}_m$, le risque des $\hat{s}_m$ peut tout simplement être estimé par la moyenne empirique, comme dans la définition de l'ERM:

$$P\gamma(\hat{s}_m) \approx \frac{1}{n_v} \sum_{i=1}^{n_v} \gamma(\hat{s}_m, Z_i).$$

Cet estimateur est lui, sans biais. Il est alors possible de choisir un modèle $\hat{m}$ par minimisation du risque empirique sur l'ensemble $\{\hat{s}_m : m \in \mathcal{M}\}$.

A défaut d'un *échantillon de validation*, il est possible de subdiviser l'échantillon donné en un sous-échantillon d'entraînement et un sous-échantillon de validation, et de procéder comme ci-dessus, avec des estimateurs entraînés sur une partie des données seulement: c'est le *hold-out*. Plus précisément, pour tout sous-ensemble $T \subset \{1, \ldots, n\}$, soit $D_n^T = (Z_i)_{i \in T}$ le sous-échantillon correspondant. Etant donnée une famille d'estimateurs $(\hat{s}_m)_{m \in \mathcal{M}}$, le risque de $\hat{s}_m(D_n^T)$ peut être estimé par

$$\mathrm{HO}_T(m) = \frac{1}{n - |T|} \sum_{i \notin T} \gamma(\hat{s}_m(D_n^T), Z_i).$$

On sélectionne alors l'estimateur dont le risque estimé est le plus faible, ce qui donne:

$$\widehat{f}_T^{\mathrm{ho}}(\mathcal{M}, D_n) = \hat{s}_{\hat{m}_T}(D_n^T), \text{ où } \hat{m}_T \in \operatorname*{argmin}_{m \in \mathcal{M}} \mathrm{HO}_T(m).$$

Cette méthode fournit un estimateur sans biais du risque qui ne nécessite aucune connaissance à priori sur la loi P ou sur les estimateurs $\hat{s}_m$, mais au prix d'une détérioration de la performance des estimateurs $\hat{s}_m$ (car ils sont entraînés sur un échantillon de taille réduite). Comme $\mathrm{HO}_T(\cdot)$ estime le risque des $\hat{s}_m(D_n^T)$ et non des $\hat{s}_m(D_n)$, il est plus naturel de comparer $\ell(s, \widehat{f}_T^{\mathrm{ho}})$ à $\inf_{m \in \mathcal{M}} \ell(s, \hat{s}_m(D_n^T))$ plutôt qu'à $\inf_{m \in \mathcal{M}} \ell(s, \hat{s}_m(D_n))$, chose qui ne peut être faite sans des hypothèses de stabilité supplémentaire sur les estimateurs $\hat{s}_m$ [60]. Pour cette raison, les inégalités d'oracle générales pour le hold-out sont du type

$$\mathbb{E}\left[\ell(s, \widehat{f}_T^{\mathrm{ho}})\right] \leqslant C\mathbb{E}\left[\inf_{m \in \mathcal{M}} \ell(s, \hat{s}_m(D_n^T))\right] + r_n. \tag{1.2}$$

1.2.1 Théorie existante

Si la fonction de contraste γ est *bornée*, l'inégalité de Hoeffding implique que

$$\sup_{m\in\mathcal{M}} |\mathrm{HO}_T(m) - P\gamma(\hat{s}_m)| = \mathcal{O}\left(\sqrt{\frac{\log(|\mathcal{M}|)}{n-|T|}}\right),$$

on peut donc aisément montrer que le hold-out vérifie une inégalité d'oracle 1.2 avec constante $C = 1$ et terme de reste $r_n = \mathcal{O}\left(\sqrt{\frac{\log(|\mathcal{M}|)}{n-|T|}}\right)$. Cette inégalité d'oracle n'est pas optimale, parce que le terme de reste n'est pas toujours négligeable: si par exemple γ représente la perte des moindres carrés en régression, la vitesse de convergence paramétrique est de $\mathcal{O}\left(\frac{1}{n}\right)$ et non $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$.

Massart [76, Corollaire 8.8] a montré que l'on pouvait améliorer ce résultat en supposant que la variance de $\mathrm{HO}_T(m)$ est majorée par une fonction de l'excès de risque. Plus précisément, l'*hypothèse de marge* affirme qu'il existe une fonction décroissante w telle que $\frac{w(x)}{x}$ décroît (sous-linéaire) et

$$\forall m, m' \in \mathcal{M}, \mathrm{Var}_Z\left(\gamma(\hat{s}_m, Z) - \gamma(\hat{s}_{m'}, Z)\right) \leqslant \left(w(\sqrt{\ell(s, \hat{s}_m)}) + w(\sqrt{\ell(s, \hat{s}_{m'})})\right)^2. \tag{1.3}$$

Sous cette condition, Massart [76, Corollaire 8.8] montre que le hold-out vérifie une inégalité d'oracle 1.2 avec un terme de reste qui est proportionnel, à des termes logarithmiques près, à la solution de l'équation

$$w(u) = \sqrt{n-|T|}u^2.$$

1.2.2 Contributions à la théorie générale du hold-out

Dans le Théorème 3.7.2 du Chapitre 3 de cette thèse, j'étend le résultat de Massart en affaiblissant ses hypothèses de deux manières différentes. Premièrement, je montre qu'on peut remplacer la majoration uniforme de γ par une seconde hypothèse de marge, de la forme

$$\|\gamma(\hat{s}_m, \cdot) - \gamma(\hat{s}_{m'}, \cdot)\|_\infty \leqslant w_2(\sqrt{\ell(s, \hat{s}_m)}) + w_2(\sqrt{\ell(s, \hat{s}_{m'})}).$$

Cela permet à la norme uniforme $\|\gamma(\hat{s}_m, \cdot)\|_\infty$ de dépendre de $m \in \mathcal{M}$ et du cardinal n de l'échantillon. Deuxièmement, je montre que, dans l'hypothèse de marge 1.3, l'hypothèse que $\frac{w(x)}{x}$ est décroissante peut être remplacée par l'hypothèse que $\frac{w(x)}{x^2}$ est bornée, au prix de quelques complications supplémentaires. En effet, l'hypothèse $\frac{w(x)}{x^2}$ ne permet que d'obtenir une inégalité d'oracle valable avec grande probabilité $1 - \frac{1}{n^2}$ (par exemple), tandis que pour contrôler l'espérance, il faut que

l'hypothèse de Massart soit vérifiée pour une autre fonction w'. Un terme correctif d'ordre deux, dépendant de w', apparaît donc dans l'inégalité d'oracle.

L'intérêt de ces changements est qu'ils permettent d'étendre la théorie générale du hold-out à de fonctions de contraste non-bornées γ, en particulier en régression. En régression, $\gamma(\hat{s}_m,(x,y)) = \phi(y - \hat{s}_m(x))$, où ϕ est généralement convexe et non-bornée sur $\mathbb{R}$: par exemple, $\phi(x) = x^2$ (moindres carrés) ou $\phi(x) = |x|$ (régression L^1). Dans ces cas, $\gamma(\hat{s}_m,(x,y))$ et généralement non-bornée à moins que $Y \in L^\infty$ et que $\|\hat{s}_m\|_\infty \leqslant A$ pour une certaine constante A. Choisir $\hat{s}_m$ uniformément bornée n'est pas pratique car la fonction de régression s peut ne pas être bornée et, même si elle l'est, sa norme $\|s\|_\infty$ est inconnue en général. $\|\hat{s}_m\|_\infty$ doit donc dépendre de m et/ou n afin d'obtenir un estimateur consistant dans ce cas. Si $\|\hat{s}_m\|_\infty$ *n'est pas* uniformément bornée, alors l'hypothèse de marge 1.3 n'est en général pas vérifiée pour une fonction w telle que $\frac{w(x)}{x}$ décroît. En effet, $\mathrm{Var}\left(\gamma(\hat{s}_m,Z) - \gamma(\hat{s}_{m'},Z)\right)$ est une fonction quadratique de $\gamma(\hat{s}_m,Z)$, tandis que $\ell(s,\hat{s}_m) = P\gamma(\hat{s}_m) - P\gamma(s)$ est une fonction linéaire de $\gamma(\hat{s}_m,Z)$. Dans le cas non-borné, il faut donc permettre à w de croître aussi vite que $x \mapsto x^2$ pour qu'une hypothèse de marge 1.3 soit vérifiée: c'est ce que fait le Théorème 3.7.3 de cette thèse. Ce théorème s'applique notamment aux méthodes à noyaux avec fonction de perte Lipschitz (présentées en section 1.7.1) et en régression parcimonieuse avec perte Huber (cadre présenté en section 1.7.2).

1.3 Validation croisée

Bien que le hold-out se prête bien à l'étude théorique, deux inconvénients limitent son utilisation pratique. Premièrement, il y a la nécessité de soustraire une certaine quantité de données à l'échantillon d'entraînement, ce qui dégrade en général la performance des estimateurs. Deuxièmement, il faut choisir arbitrairement un sous-ensemble d'entraînement T parmis les indices $\{1,\dots,n\}$ des données. Pour un cardinal fixé $|T| = n_t$, le choix de T n'a aucune influence sur la loi de $\mathrm{HO}_T(m)$ ni sur celle de $\widehat{f}_T^{\mathrm{ho}}$, on peut donc supposer qu'il s'agit d'une variable aléatoire uniformément distribuée dans l'ensemble $\{T \subset \{1,\dots,n\} : |T| = n_t\}$: le choix de T est donc purement une source de bruit. La variance induite a des chances d'être particulièrement forte quand les estimateurs $\hat{s}_m$ sont *instables* (c'est à dire quand une légère modification de l'échantillon d'entraînement provoque un changement important de l'estimateur).

Pour réduire cette variance, une pratique courante consiste à moyenner plusieurs estimateurs de risque HOT_jm correspondant à des sous-ensembles T_j différents: cela s'appelle la *validation croisée*. Les méthodes de validation croisée modernes ont été introduites par Stone [98] et Geisser [45]. Soit $[n] = \{1,\dots,n\}$ et soit $\mathcal{T} = (T_j)_{1\leqslant j\leqslant V}$ une suite finie de sous-ensembles de $[n]$ (souvent, mais pas tou-

jours, de même cardinal n_t). Geisser [45] défini l'estimateur par validation croisée du risque de l'estimateur $\hat{s}_m$ par

$$CV_{\mathcal{T}}(m) = \frac{1}{V} \sum_{j=1}^{V} \mathrm{HO}_{T_j}(m) .$$

Cette définition comprend le *leave-one-out* ($\mathcal{T} = ([n]\backslash j)_{1 \leqslant j \leqslant n}$), la validation croisée Monte Carlo ($(T_j)_{1 \leqslant j \leqslant V}$ i.i.d uniformes parmis les sous-ensembles de cardinal n_t) ou la validation croisée V-fold ($\mathcal{T} = ([n]\backslash I_j)_{1 \leqslant j \leqslant V}$ pour une partition $(I_j)_{1 \leqslant j \leqslant V}$ de $[n]$ en sous-ensembles de cardinal n/V), parmis les méthodes les plus classiques. Le lecteur intéressé trouvera une présentation plus détaillée des différentes méthodes de validation croisée dans l'article bibliographique d'Arlot et Celisse [3]. On voit donc que la validation croisée permet d'utiliser des échantillons d'entraînement plus grands que pour le hold-out: en effet, dans le cas du hold-out, il serait absurde de n'utiliser qu'une donnée pour la validation ($|T_j| = n - 1$), mais cela se fait dans le cas de la validation croisée (*leave one out*).

Quand la validation croisée est utilisée pour sélectionner un estimateur, on choisit le paramètre (ou modèle) minimisant le risque estimé,

$$\hat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} CV_{\mathcal{T}}(m).$$

Pour construire l'estimateur final, plusieurs options existent: la plus standard, en pratique en tous cas, est d'utiliser $\hat{s}_{\hat{m}}(D_n)$, l'estimateur entraîné sur toutes les données. D'autres variantes sont parfois considérées pour les besoins de la théorie.

L'article [3] présente les principaux résultats théoriques connus sur la validation croisée, avant 2010. Pour une variante adaptée, des inégalités d'oracle générales semblables à celles de Massart pour le hold-out [76] ont été démontrées par Van der Waart et al [105]. A ma connaissance, aucun résultat comparable n'existe pour la validation croisée standard $\hat{s}_{\hat{m}}(D_n)$.

1.4 Agrégation de modèles et d'hyperparamètres

Les méthodes définies jusqu'ici cherchent toutes à identifier un seul bon estimateur parmis les $(\hat{s}_m)_{m \in \mathcal{M}}$. Comme le but est de se comparer au meilleur des $\hat{s}_m$, un tel procédé semble naturel; cependant, rien n'oblige à se restreindre à ce type d'estimateurs. A la différence des méthodes de *sélection*, les méthodes d' *agrégation* considèrent des estimateurs qui sont des sommes pondérées des $(\hat{s}_m)_{m \in \mathcal{M}}$, i.e

$$\hat{s} = \sum_{m \in \mathcal{M}} \hat{w}_m \hat{s}_m, \tag{1.4}$$

où les coefficients $(\hat{w}_m)_{m\in\mathcal{M}} \in \mathbb{R}^{\mathcal{M}}$ dépendent en général des données. On parle d'agrégation convexe quand les poids sont positifs et ont pour somme 1, et d'agrégation linéaire quand aucune contrainte n'est imposée. Pour que l'équation (1.4) ait un sens, il faut que les $\hat{s}_m$ appartiennent à un espace vectoriel (pour l'agrégation linéaire) ou à un ensemble convexe (dans le cas de l'agrégation convexe), ce qui est une première différence par rapport aux méthodes de sélection. De plus, pour que l'agrégation convexe soit pertinente, il vaut mieux que le risque soit convexe, puisque cela garantit que pour toute combinaison convexe $\sum_{m\in\mathcal{M}} \hat{w}_m \hat{s}_m$,

$$P\gamma\left(\sum_{m\in\mathcal{M}} \hat{w}_m \hat{s}_m\right) \leqslant \sum_{m\in\mathcal{M}} \hat{w}_m P\gamma(\hat{s}_m).$$

Si cette inégalité n'est pas satisfaite, cela signifie qu'il est préférable de tirer au hasard l'un des $\hat{m}$ selon la loi $(\hat{w}_m)_{m\in\mathcal{M}}$ plutôt que de former la combinaison convexe $\sum_{m\in\mathcal{M}} \hat{w}_m \hat{s}_m$. Ainsi, dans le cas non-convexe, il se peut que la sélection d'un estimateur soit préférable à l'agrégation.

L'hypothèse de convexité du risque est vérifiée en régression et en estimation de densité, pour la plupart des fonctions de perte utilisées, ainsi qu'en classification (cas des *relaxations convexes* du risque $0-1$ [10]). L'agrégation peut alors améliorer les performances de façon importante, puisque le minimum de $P\gamma$ sur l'enveloppe convexe des $\hat{s}_m$ peut être bien inférieur à $\min_{m\in\mathcal{M}} P\gamma(\hat{s}_m)$.

Même sans aucune hypothèse de ce type, l'agrégation a un intérêt. En effet, en estimation de densité [117, 28] et en régression [116] [29], Yang et Catoni ont construit des méthodes d'agrégation qui vérifient une inégalité d'oracle avec constante $C = 1$ et terme de reste $\mathcal{O}\left(\frac{\log|\mathcal{M}|}{n_v}\right)$, ce qui est optimal [102], et meilleur que ce que pourrait faire une méthode de sélection [29] [57]. Ces méthodes reposent, comme le hold-out, sur la subdivision de l'échantillon en sous-échantillon d'entraînement et sous-échantillon de validation, utilisé dans ce cas pour construire les coefficients $\hat{w}_m$. Ces coefficients $\hat{w}_m$ sont fonction décroissante du risque empirique de $\hat{s}_m$ sur l'échantillon de validation: ainsi, l'essentiel de la masse de la loi $m \mapsto \hat{w}_m$ est concentré là où le risque est petit. Un inconvénient de ces méthodes d'agrégation est qu'en dehors de l'estimation de densité, elles ont des paramètres libres qui doivent être fixés en utilisant des connaissances à priori sur la loi P, afin que les inégalités d'oracle optimales soient bien vérifiées.

Dans l'ensemble, ce type d'agrégation reste assez proche de la sélection de modèles en terme de méthodes et d'objectifs, ainsi que pour les garanties théoriques. Tous deux sont conçus pour être appliqués à de bonnes familles d'estimateurs, telles que dans toute situation, au moins l'un des estimateurs ait de bonnes performances. C'est pourquoi l'inégalité d'oracle 1.2 est un critère pertinent.

1.5 Agrégation randomisée

Le problème est différent quand l'agrégation est appliquée à de mauvais estimateurs. Dans ce cas, il n'est pas suffisant d'atteindre la performance de l'oracle - du meilleur de ces mauvais estimateurs. On espère en fait que l'agrégat sera bien meilleur que n'importe lequel des estimateurs, pris individuellement.

Une raison possible pour qu'un estimateur ait un risque élevé est qu'il soit *instable* ou, autrement dit, qu'il ait une *variance* élevée par rapport à l'échantillon d'entraînement. Breiman [22] suggère de réduire cette variance en agrégeant les estimateurs $\hat{s}(D^*_{n,1}) \dots \hat{s}(D^*_{n,B})$, obtenus par évaluation successive de $\hat{s}_m$ sur des échantillons $D^*_{n,j}$ modifiés aléatoirement.

Plus précisément, Breiman suggère de tirer uniformément, indépendemment et avec remplacement des données de l'échantillon D_n, formant ainsi le *rééchantillon* $D^*_n = (X_{I_1}, \dots, X_{I_n})$, où $(I_j)_{1 \leqslant j \leqslant n}$ sont des indices i.i.d, uniformes dans $[n]$. L'estimateur *baggé* $\hat{s}^{bag}$ est alors

$$\hat{s}^{bag} = \mathbb{E}_*[\hat{s}(D^*_n)] = \mathbb{E}[\hat{s}(D^*_n)|D_n].$$

Comme cet estimateur est difficile à calculer en général, on lui substitue souvent son approximation par la méthode de Monte Carlo,

$$\hat{s}^{bag}_B = \frac{1}{B} \sum_{j=1}^{B} \hat{s}(D^*_{n,j}),$$

où les rééchantillons $(D^*_{n,j})_{1 \leqslant j \leqslant B}$ sont indépendants conditionnellement à D_n.

Une autre variante du bagging est le subagging [25], qui consiste à former des sous-échantillons par tirage aléatoire *sans remplacement*, ce qui donne $D^{T_j}_n = \{(X_i)_{i \in T_j}\}$, où les T_j sont des sous-ensembles i.i.d de cardinal fixé $n_t < n$. L'estimateur "subaggé" est alors

$$\hat{s}^{sub}_B = \frac{1}{B} \sum_{j=1}^{B} \hat{s}(D^{T_j}_n).$$

Le (su)bagging est particulièrement intéressant quand il est appliqué à des estimateurs peu réguliers, commes les arbres ou les réseaux de neurone. Appliqué aux arbres CART, le bagging peut réduire la variance et le risque d'un facteur constant (Bühlmann et Yu [25]). Appliqué à l'estimateur du plus proche voisin ($1-$NN), il le rend consistant [13], et le fait même converger à la vitesse optimale [14].

Le bagging fonctionne en introduisant de l'aléas dans l'échantillon, par rapport auquel l'estimateur est instable, puis en agrégeant le résultat. Plus généralement, on peut envisager d'introduire de l'aléatoire dans d'autres éléments instables d'un

algorithme statistique, pour ensuite l'agréger. En particulier, plusieurs auteurs ont suggéré d'introduire de l'aléas dans la construction des arbres CART, une idée dont la réalisation la plus célèbre est la méthode des forêts aléatoires de Breiman [23]. Les forêts aléatoires fonctionnent très bien en pratique, au point qu'elles sont considérées comme une des meilleurs méthodes "boîte noire" et généraliste en apprentissage machine [12].

1.6 Agrégation d'hold-out

Les résultats théoriques et pratiques sur l'agrégation randomisée montrent qu'il est possible de rendre performants des estimateurs *instables* comme les arbres, la méthode du plus proche voisin ou les régressogrammes en introduisant de l'aléatoire dans leur construction et en les agrégeant. Dans le cadre de la sélection de modèles, on peut envisager d'appliquer cette méthode au hold-out, habituellement négligé en raison de son instabilité. L'instabilité du hold-out est due en partie au choix arbitraire d'un sous-ensemble $T \subset [n]$, utilisé pour subdiviser l'échantillon. Il est donc possible de réduire la variance du hold-out en agrégeant plusieurs estimateurs hold-out correspondant à des choix différents du paramètre T.

Plus précisément, l'agrégation d'hold-out consiste à calculer un estimateur

$$\widehat{f}_{\mathcal{T}}^{\mathrm{ag}} = \frac{1}{V} \sum_{j=1}^{V} \widehat{f}_{T_j}^{\mathrm{ho}}$$

pour une certaine famille $\mathcal{T} = (T_j)_{1 \leqslant j \leqslant V}$ de sous-ensembles de $[n]$, indépendants de l'échantillon D_n. Comme pour l'agrégation d'estimateurs plus généralement, cette définition suppose que les estimateurs appartiennent à un espace vectoriel, ce qui est le cas quand $\mathcal{S}$ est constitué de fonctions à valeur réelle (cas de la régression et de l'estimation de densité). Dans cette thèse, je ne considère que des familles $\mathcal{T}$ contenant des sous-ensembles T_j de même cardinal n_t, puisque cela garantit par l'inégalité de Jensen que le risque de $\widehat{f}_{T}^{\mathrm{ag}}$ est inférieur à celui du hold-out, pour une taille donnée $n_t = |T_j|$ de l'échantillon d'entraînement $D_n^{T_j}$ (voir plus loin). Comme pour la validation-croisée, il y a plusieurs façons de générer une famille de sous-ensembles $\mathcal{T}$. Par analogie avec la validation croisée, on parlera d' "Agghoo V-fold" quand $\mathcal{T} = ([n] \backslash I_j)_{1 \leqslant j \leqslant V}$, où les sous-ensembles de validation $(I_j)_{1 \leqslant j \leqslant V}$ sont disjoints et de même cardinal n/V. On parlera d" Agghoo Monte-Carlo" quand la famille $\mathcal{T}$ est constituée de sous-ensembles aléatoires i.i.d T_j de même taille n_t, tirés sans remise dans $[n]$.

L'agrégation d'hold-out ressemble par certains aspects à la validation croisée, à l'agrégation pour la sélection de modèles et à l'agrégation randomisée. Agghoo agrège plusieurs estimateurs hold-out en variant le sous-ensemble T utilisé pour

subdiviser l'échantillon, à l'instar de la validation croisée. Comme les méthodes d'agrégation pour la sélection de modèles, Agghoo agrège plusieurs estimateurs différents appartenant à une famille donnée $(\hat{s}_m)_{m\in\mathcal{M}}$, et accorde un poids plus grand aux estimateurs ayant un risque empirique faible sur un sous-échantillon $D_n^{T_j^c}$. Enfin, Agghoo Monte Carlo s'obtient en randomisant le paramètre T du hold-out et en agrégeant le résultat. De plus, tout come le subagging, Agghoo agrège des estimateurs qui ont été entraînés sur des sous-échantillons distincts $D_n^{T_j}$.

Des méthodes ressemblant plus ou moins à Agghoo ont déjà été proposées [59, 114, 51, 87, 58] .La méthode étudiée dans cette thèse se distingue par la généralité de sa définition, due en partie au fait que ce sont les estimateurs $\hat{s}_m$, et non les paramètres m, qui sont agrégés, contrairement à [59, 51]. Ainsi, Agghoo ne dépend que de l'ensemble d'estimateurs $\{\hat{s}_m : m \in \mathcal{M}\}$ et non de la paramétrisation $m \mapsto \hat{s}_m$, ce qui évite des ambiguïtés quand plusieurs paramétrisations sont d'utilisation courante (cas des méthodes à noyaux, notamment [95]).

L'intérêt pour Agghoo est motivé par ses bonnes performances en pratique, remarquées notamment par [107] et [54]. Les simulations menées au cours de cette thèse montrent que Agghoo peut se révéler meilleure que la validation croisée en application aux méthodes à noyaux en régression (Chapitre 1, Figure 3.1), au Lasso (Chapitre 4, Figures 4.1, 4.2, 4.3) et aux séries trigonométriques en estimation de densité (Figure 2.1). Dans plusieurs de ces simulations, Agghoo s'est même révélée capable de faire mieux que l'oracle de sélection.

Sur le plan théorique, Agghoo se distingue par sa sûreté: tant que le risque $P\gamma$ est convexe, Agghoo fait toujours mieux que le hold-out, par l'inégalité de Jensen:

$$\mathbb{E}\left[P\gamma\left(\frac{1}{V}\sum_{j=1}^{V}\widehat{f}_{T_j}^{\mathrm{ho}}\right)\right] \leqslant \frac{1}{V}\sum_{j=1}^{V}\mathbb{E}\left[P\gamma(\widehat{f}_{T_j}^{\mathrm{ho}})\right] = \mathbb{E}\left[P\gamma(\widehat{f}_{T_1}^{\mathrm{ho}})\right]. \tag{1.5}$$

Une telle garantie n'existe pas pour les méthodes fondées sur l'agrégation d'hyperparamètres: en effet, la convexité de $P\gamma$ n'implique pas celle de $\lambda \mapsto P\gamma(\hat{s}_\lambda)$, qui est en général beaucoup plus compliquée à établir, car elle dépend à la fois de la famille d'estimateur $\hat{s}_\lambda$ et (potentiellement) des données.

Une conséquence de l'inégalité (1.5) est que Agghoo vérifie les mêmes inégalités d'oracle que le hold-out, tant que le risque $P\gamma$ est convexe. En particulier, l' inégalité d'oracle générale du Chapitre 3, décrite en section 1.2.2, implique directement un résultat similaire pour Agghoo. Le Théorème 3.7.3 énonce l'inégalité d'oracle résultante. Ces Théorèmes suggèrent que Agghoo, comme le hold-out, se comporte bien dans des situations variées.

1.7 Contributions: inégalités d'oracle pour le hold-out et Agghoo dans des cadres spécifiques

Comme une théorie générale existe, on peut penser qu'Agghoo et le hold-out font aussi bien que l'oracle en général. Cependant, les Théorèmes 3.7.2 et 3.7.3 ne permettent pas une telle interprétation. En effet,ces théorèmes généraux ne nous disent pas si une condition de marge 1.3 est vérifiée dans une situation donnée et si oui, pour quelle fonction w. Sans connaître la fonction w, il est difficile de savoir si le terme de reste r_n dans l'inégalité d'oracle est oui ou non négligeable par rapport au risque de l'oracle. Afin d'identifier des hypothèses plus claires et plus intuitives que l'inégalité de marge générale 1.3, et pour calculer explicitement le terme de reste r_n, il est nécessaire de considérer des cas particuliers. C'est ce qui est fait dans les chapitres 3 et 4 de cette thèse. Comme expliqué en section 1.2.2, la théorie classique fondée sur des hypothèses de marge avec $x \mapsto \frac{w(x)}{x}$ décroissante ne s'applique pas en général aux fonctions de perte non-bornées: c'est donc pour de tels problèmes que les techniques développées dans cette thèse trouvent leur champ d'application le plus intéressant. Ainsi, dans les chapitres 3 et 4, on s'intérèsse à des familles d'estimateurs non-bornés en régression et classification, pour lesquelles la théorie classique ne s'applique pas. Plus précisément, le chapitre 3 a pour sujet les méthodes à noyaux avec fonction de perte Lipschitz, tandis que le chapitre 4 se situe dans le cadre de la régression linéaire parcimonieuse, avec fonction de perte Huber.

1.7.1 Contributions: hold-out appliqué aux méthodes à noyaux

Les *méthodes à noyaux* sont des méthodes de minimisation du risque empirique pénalisé pour lesquelles la pénalité $\Omega(t) = \|t\|_{\mathcal{H}}^2$, où $\|\cdot\|_{\mathcal{H}}^2$ désigne la norme d'un *espace de Hilbert à noyau reproduisant* (RKHS) $\mathcal{H}$. Un RKHS $\mathcal{H}$ est un espace de Hilbert de fonctions réelles sur un ensemble $\mathcal{X}$ associé de façon unique à toute fonction symétrique définie positive K sur $\mathcal{X}$ (le *noyau*). Si $\mathcal{S}$ est un ensemble de fonctions mesurables $\mathcal{X} \to \mathbb{R}$ (cas de la régression ou des *relaxations convexes* de la classification), on peut définir, pour tout noyau K (ou RKHS $\mathcal{H}$), la *méthode à noyau* associée:

$$\hat{s}_\lambda = \operatorname*{argmin}_{t \in \mathcal{H}} \left\{ P_n \gamma(t) + \lambda \|t\|_{\mathcal{H}}^2 \right\}.$$

La méthode à noyau a donc un *paramètre de régularisation* λ. Je renvoie le lecteur au livre de Scholkopf et Smola [95] pour plus d'information sur les méthodes à noyau.

Dans le chapitre 3, on considère des fonctions de perte γ qui s'expriment sous

la forme $\gamma(t,(x,y)) = c(t(x),y)$ pour une fonction c qui est Lipschitz et convexe en son premier argument. Cela inclut la régression ($c(t(x),y) = \phi(y - t(x))$) ainsi que les *relaxations convexes* du problème de classification ($c(t(x),y) = \phi(yt(x))$), du moment que ϕ est Lipschitz et convexe. En régression, on peut notamment citer la perte L^1, $\phi(x) = |x|$ et en classification, la perte charnière $\phi(u) = (1-u)_+$, utilisée dans les *machines à vecteur de support*.

Le chapitre 3, écrit en collaboration avec mes directeurs de thèse Sylvain Arlot et Matthieu Lerasle, énonce et démontre une inégalité d'oracle pour le hold-out appliqué au choix du paramètre de régularisation λ des méthodes à noyaux (Théorème 3.4.3). Nos hypothèses sont de deux types: une hypothèse de marge (l'hypothèse $SC_{\rho,\nu}$) utile en général, et des hypothèses spécifiques qui permettent de gérer l'absence de borne uniforme sur le risque. Dans le cas de la perte L^1, nous montrons que l'hypothèse $SC_{\rho,\nu}$ peut être déduite d'un cas particulier des conditions énoncés par Steinwart [97] dans son analyse de l'hypothèse de marge pour la perte L^1. Ces conditions ne dépendent que de la loi conditionnelle de Y sachant X au voisinage de $s(X)$, elles sont donc sans rapport avec la question des bornes uniformes sur le risque $\gamma(t,(x,y))$, sur s ou sur les estimateurs $\hat{s}_\lambda$. Les hypothèses spécifiques aux méthodes à noyaux sont que le noyau est uniformément borné ($\|K\|_\infty < +\infty$) et que le paramètre de régularisation est borné inférieurement par une valeur $\lambda_m(n) > 0$. Le terme de reste de l'inégalité d'oracle , r_n, fait intervenir les constantes ρ, ν de l'hypothèse de marge, ainsi que $\lambda_m(n)$ et $\|K\|_\infty$. Le Théorème 3.4.3 permet de démontrer que le hold-out converge à la "bonne" vitesse pour certaines des vitesses de convergence obtenues dans [40] pour le noyau gaussien ($K(x,x') = e^{-\nu\|x-x'\|^2}$), à ceci près que nous ne considérons que le choix du paramètre de régularisation λ et pas celui du paramètre ν du noyau RBF gaussien.

A ma connaissance, il s'agit de la première inégalité d'oracle pour le hold-out appliqué aux méthodes à noyaux. Le choix du paramètre λ de ces méthodes a bien été traité par Eberts et Steinwart [40] , mais pour utiliser la théorie classique du hold-out, ils ont en fait modifié les estimateurs à noyaux afin de les rendre bornés, en les tronquant, c'est à dire en considérant à la place de $\hat{s}_\lambda$,

$$\mathrm{Trunc}_M(\hat{s}_\lambda) = \max\left(\min\left(\hat{s}_\lambda, M\right), -M\right)$$

pour une constante $M > 0$. Cette opération peut se justifier si un majorant de $\|s\|_\infty$ est connu, mais si jamais $M < \|s\|_\infty$, l'opération de troncature risque fort d'empêcher les estimateurs $\hat{s}_\lambda$ de converger vers l'optimum s. En tous cas, elle ne semble pas être utilisée en pratique.

1.7.2 Régression parcimonieuse

Dans le chapitre 4, j'étudie le hold-out (et Agghoo) appliqués à des familles d'estimateurs affines $x \mapsto \hat{q}_k + \langle \hat{\theta}_k, x \rangle$ qui ne dépendent que d'un petit nombre de variables x_j (parsimonie). Plus précisément, on suppose que

$$\left\| \hat{\theta}_k \right\|_0 = \left| \left\{ j : \hat{\theta}_{k,j} \neq 0 \right\} \right| \leqslant k, \tag{1.6}$$

i.e que les vecteurs $\hat{\theta}_k$ ont moins de k composantes non-nulles. Le risque des estimateurs est évalué en utilisant la fonction de perte Huber, une fonction de perte Lipschitz classique en *régression robuste* [55].

L'entier k mesure donc la *complexité* des estimateurs $\hat{\theta}_k$. Des méthodes de régression parcimonieuse telles que la régression l^0 (ou sélection de modèle complète), forward stepwise [52, Section 3.3] et LARS [42] donnent naturellement lieu à des estimateurs qui vérifient l'équation (1.6), car ces méthodes ajoutent les variables une à une ($\left\| \hat{\theta}_m \right\|_0$ est monotone par rapport au paramètre m de ces méthodes). Il est aussi possible d'extraire des estimateurs vérifiant (1.6) du "chemin de régularisation" $(\hat{s}_\lambda)_{\lambda>0}$ du Lasso [99], comme l'ont suggéré Zou et al [120].

Dans ce cadre, je démontre que le hold-out vérifie une inégalité d'oracle de type (1.2) avec constante $C > 1$ et terme de reste $r_n = \mathcal{O}\left(\frac{\log(K)}{n_v}\right)$, où K désigne le nombre d'estimateurs et n_v désigne le cardinal de l'échantillon de validation ($n_v = |T^c|$). Si n_v est d'ordre n, ce terme de reste est plus petit que les vitesses de convergence théoriques en régression parcimonieuse [15, 90]. Le risque du hold-out est donc du bon ordre de grandeur: il *s'adapte* aux vitesses de convergence théoriques.

Les estimateurs θ_k et les variables prédictives X sont supposés bornés, mais pas uniformément: il est seulement supposé que $\|X\|_{\infty,L^\infty}$ et $\|\theta_k\|$ sont majorées par une fonction polynômiale de n, la taille de l'échantillon. Cette hypothèse est donc peu contraignante. L'absence de borne uniforme est gérée grâce à une hypothèse d'équivalence des normes L^∞ et L^2 de la forme

$$\forall \theta, \|\theta\|_0 \leqslant 2K \implies \|\langle \theta, X \rangle\|_\infty \leqslant \kappa(n, n_v) \|\langle \theta, X \rangle\|_{L^2}, \tag{1.7}$$

pour une constante $\kappa(n, n_v)$ telle que $\kappa(n, n_v) = \mathcal{O}\left(\sqrt{\frac{n_v}{\log n}}\right)$ quand $n_v \to +\infty$. Le chapitre 4 contient deux exemples qui montrent que l'équation (1.7) est vérifiée sous des hypothèses raisonnables.

L'inégalité d'oracle (Théorème 4.3.2) est obtenue sous des hypothèses particulièrement faibles, en particulier en ce qui concerne les bornes sur $\|\theta_k\|$ et X. Concernant le Lasso, des résultats ont été obtenus sous des hypothèses faibles par Chetverikhov, Liao et Chernozhukov [33] (en particulier, sans supposer de bornes

sur $\hat{\theta}_k$), mais il ne s'agit pas d'inégalités d'oracle: [33] montre "seulement" que le hold-out converge à la bonne vitesse théorique (vitesse minimax) sous des hypothèses de parsimonie. Cela ne garantit pas que le hold-out s'adapte à d'autres types d'hypothèse et à d'autres vitesses de convergence.

Pour ce qui est des inégalités d'oracle, Lecué et Mitchell [69] en démontrent pour le hold-out appliqué au Lasso, mais ils supposent que les vecteurs $\hat{\theta}_k$ sont uniformément bornés dans ℓ^2 (et ils doivent modifier l'algorithme du Lasso pour cela). Wegkamp [114] démontre une inégalité d'oracle générale en régression L^2, sans supposer que les estimateurs sont bornés. Néanmoins, il suppose que les estimateurs $\hat{s}_k$ vérifient une inégalité $L^4 - L^2$ de la forme

$$\int (\hat{s}_k - s)^4 dP_X \leqslant R \int (\hat{s}_k - s)^2 dP_x \tag{1.8}$$

pour une constante $R > 0$, ce qui implique en fait une borne sur les estimateurs $\hat{s}_k$. En effet, l'inégalité (1.8) n'est pas homogène, contrairement à l'inégalité (1.7), elle ne peut donc pas être vérifiée pour des éléments arbitrairement grands d'un espace vectoriel de fonctions. L'inégalité (1.8) appliquée à $\hat{s}_k : x \mapsto \hat{q}_k + \langle \hat{\theta}_k, x \rangle$ implique donc une borne sur $\left\| \hat{\theta}_k \right\|$.

En dehors de l'étude théorique du hold-out, des hypothèses d'équivalence des normes ont été utilisées récemment pour étudier la minimisation du risque empirique en régression L^2 [7, 81] ainsi que ses concurrentes "robustes" [7, 66]. Signalons en particulier l'article [7], qui utilise une hypothèse de la forme

$$\|t\|_\infty \leqslant \kappa \|t\|_{L^2(X)} ,$$

pour toute fonction t appartenant au *modèle* m sur lequel le risque doit être minimisé. Ces hypothèses sont en général plus contraignantes que (1.7), car le rapport des deux normes est supposé uniformément borné, tandis que l'équation (1.7) permet à la constante $\kappa(n, n_v)$ de croître en fonction de n_v à la vitesse $\sqrt{\frac{n_v}{\log n}}$.

1.8 Etude détaillée du hold-out et de Agghoo

Des inégalités d'oracle montrent que Agghoo et le hold-out sont de bonnes méthodes de sélection de modèle *en théorie*, dans la mesure où leur performance est proche de celle du meilleur estimateur donné (l'oracle). Cependant, d'autres méthodes, comme la validation croisée, vérifient aussi des inégalités d'oracle. Ces résultats ne peuvent donc pas nous dire lesquelles de ces méthodes fonctionnent le mieux dans une situation donnée. De plus, les résultats des simulations menées au cours de cette thèse, qui montrent que Agghoo peut parfois faire mieux que l'oracle de sélection de modèle, ne peuvent pas être expliqués par une inégalité d'oracle

(1.2) avec constante $C > 1$ et terme de reste $r_n > 0$. L'intérêt pour Agghoo est motivé en grande partie par de telles simultations, où Agghoo se montre clairement supérieure aux méthodes de sélection de modèles telles que la validation croisée. Pour des raisons pratiques aussi bien que théoriques, il serait intéressant de savoir dans quels cas ce phénomène se produit, et pour quelles valeurs des paramètres d'Agghoo. Cela permettrait de faire le bon choix entre Agghoo et ses concurrentes dans les applications, et de bien calibrer les paramètres d'Agghoo lorsque cette méthode est utilisée. D'un point de vue plus théorique, une réponse à de telles questions nous renseignerait sur le comportement du hold-out et de l'agrégation d'hyperparamètres.

Dans les Chapitres 5 et 6 de cette thèse, je mène à bien une analyse précise du hold-out et de Agghoo, afin de répondre à ces questions dans un cas particulier, celui de l'estimation de densité L^2 par des séries de Fourier empiriques. Plus précisément, j'étudie l'estimation de densité L^2 d'une fonction de densité symmétrique $s \in L^2([0;1])$ à l'aide des estimateurs

$$\hat{s}_k = 1 + \sum_{j=1}^{k} P_n(\varphi_j)\varphi_j, \tag{1.9}$$

où $\varphi_j = \sqrt{2}\cos(2\pi j \cdot)$ est la base des fonctions cosinus. Ce cadre a été choisi afin de réaliser un compromis entre la difficulté technique de son étude et son intérêt théorique et pratique. L'estimateur par séries de Fourier est certainement utilisable en pratique, et d'un point de vue théorique, il s'adapte à toutes les classes de régularité Hölder/Sobolev, du moins pour les fonctions de densité périodiques. Par ailleurs, la structure algébrique des polynômes trigonométriques et la simplicité de la formule (1.9) définissant les estimateurs $\hat{s}_k$ rendent l'analyse théorique plus facile, ce qui permet d'être plus précis dans les résultats.

Pour faciliter l'analyse théorique, je fais quelques hypothèses sur les coefficients de Fourier θ_j de s sur la base des cosinus. La première de ces hypothèses est que la suite θ_j^2 est décroissante: cela garantit la convexité du risque moyen $\mathbb{E}\left[\|\hat{s}_k - s\|^2\right]$ en fonction du paramètre k. En particulier, l'unicité du minimum k_* est garantie si la suite θ_j^2 est strictement décroissante. L'interprétation qualitative des autres hypothèses est que la suite θ_j^2 décroît polynômialement et ne "saute" pas de façon trop brutale - il n'y a pas de chute soudaine où la suite décroît soudain très rapidement. Ces hypothèses sont faites à la fois dans le chapitre 5 et dans le chapitre 6.

Pour comprendre le comportement du hold-out et de Agghoo, la première étape est d'analyser l'estimateur hold-out du risque $\mathrm{HO}_T(m)$ dont le hold-out calcule le minimum. Le chapitre 5 est consacré à la construction d'une approximation en loi

de cet estimateur empirique du risque au voisinage du "vrai" minimum,

$$k_*(n_t) = \operatorname*{argmin}_k \mathbb{E}\left[\|\hat{s}_k(D_{n_t}) - s\|^2\right].$$

Les outils habituels de l'analyse asymptotique sont inadaptés ici car le processus dont il s'agit (une fois centré et mis à l'échelle) ne converge pas, et à défaut de limite, il faut avoir recours à une approximation du processus pour chaque valeur de n, la taille de l'échantillon. Je construis cette suite d'approximations en utilisant le théorème de Komlos-Major-Tusnady pour se ramener à un processus Gaussien, que j'approxime ensuite par un autre processus Gaussien ayant une fonction de variance-covariance proche. La bonne approximation se trouve être la somme d'une fonction convexe f_n et d'un mouvement brownien bilatère changé de temps W_{g_n}. Comme il ne semble pas y avoir de formule simple pour f_n et g_n, je démontre des minorations et majorations des incréments de ces fonctions qui se révèlent utiles dans le chapitre 6.

Dans le chapitre 6, je mène à bien une analyse précise du risque du hold-out et de Agghoo. D'abord, en utilisant le Théorème 3.7.2 du chapitre 3, je démontre que le hold-out vérifie une inégalité d'oracle préliminaire, ce qui implique que le paramètre $\hat{k}$ sélectionné par le hold-out se trouve avec grande probabilité dans une région de l'espace des paramètres $\mathbb{N}$ où l'estimateur du risque hold-out est bien approximé par le processus construit au chapitre 5 précédent. Deuxièmement, en utilisant des techniques introduites par Leandro R. Pimentel [88] , je montre que l'approximation de l'estimateur hold-out du risque construite au Chapitre 5 conduit à une approximation (en loi) $\hat{k}^\infty$ de son argmin $\hat{k}$, le paramètre sélectionné par le hold-out. Au premier ordre, les risques d'Agghoo et du hold-out s'expriment à l'aide de la fonction de répartition de $\hat{k}^\infty$. En utilisant ces formules, je calcule une approximation au premier ordre du risque du hold-out, de la forme

$$\mathbb{E}\left[\|\hat{s}_{\hat{k}}(D_{n_t}) - s\|^2\right] = or(n_t) + r_n + o(r_n),$$

où $or(n_t)$ désigne le risque de l'oracle entraîné sur un échantillon de taille $n_t = n - n_v$ (Théorème 6.4.3). Je montre aussi que le risque de Agghoo vérifie une inégalité de la forme

$$\mathbb{E}\left[\left\|\widehat{f}^{\mathrm{ag}}_{\mathcal{T}} - s\right\|^2\right] \leqslant or(n_t) + r_n - d_n + o(d_n).$$

Le risque de Agghoo est majoré par la somme de deux termes: le risque du hold-out, $or(n_t) + r_n$, et un terme *négatif* d_n provenant de l'agrégation. En général, d_n est toujours supérieur à cr_n, pour une constante c indépendante de n, de sorte qu' Agghoo réduit au moins le terme de reste r_n d'un facteur constant.

Sous des hypothèses supplémentaires sur s, il est possible d'aller plus loin. Dans une dernière partie (Section 6.4.3), je montre que si θ_j^2 décroît à une vitesse

polynômiale fixe, $\theta_j^2 \sim cj^{-\alpha}$, alors le risque d'Agghoo peut être inférieur à celui de l'oracle, le rapport entre les deux pouvant descendre jusqu'à une constante $\rho < 1$ dans l'asymptotique. Je définis des intervalles de valeurs du paramètre $\tau_n = \frac{n_t}{n}$ de Agghoo sur lesquels ce phénomène se produit.

1.9 Conclusion

Cette thèse démontre que le hold-out et sa version agrégée vérifient une inégalité d'oracle générale (le Théorème 3.7.3). Ce Théorème général a été appliqué dans deux contextes, les méthodes à noyaux et la régression parcimonieuse, où il donne de nouvelles majorations du risque du hold-out (et d'Agghoo). Enfin, une étude détaillée de l'agrégation d'hold-out a été menée dans le cadre de l'estimation de densité L^2 par des séries trigonométriques. Cette étude montre qu'en fonction du choix de ses paramètres, Agghoo peut avoir de meilleures performances que l'oracle entraîné sur un échantillon de taille n_t, ou même que l'oracle entraîné sur toutes les données, avec un gain allant jusqu'à un facteur constant par rapport à l'oracle. Dans tous les cas, Agghoo améliore au moins le terme de reste dans l'inégalité d'oracle, par rapport au hold-out.

Cette thèse ouvre un certain nombre de perspectives qui sont présentées plus en détails dans le chapitre 7.

Inégalités d'oracles pour le hold-out Concernant les inégalités d'oracles sur le hold-out, on remarque que la démonstration des résultats du chapitre 4 repose essentiellement sur trois hypothèses:

- Une inégalité entre les normes L^∞ et L^2
- Une majoration de $\|\gamma(\hat{s}_m, \cdot)\|_\infty$ polynômiale en n
- Une condition de forte convexité locale sur l'excès de risque, du type $\ell(s,t) \geqslant \mu \|t-s\|^2$ pour t tel que $\|t-s\|_\infty \leqslant \delta$ $(\mu, \delta > 0)$.

Cela suggère que des inégalités d'oracle semblables à celles du chapitre 4 peuvent être démontrées dans tous les cas où ces trois hypothèses sont vérifiées.

Minimisation du risque empirique Conditionnellement à l'échantillon d'entraînement D_n^T, le hold-out minimise le risque empirique sur l'ensemble $\hat{s}_m(D_n^T)$. Il est donc vraisemblable que les techniques développées dans la thèse pour étudier le hold-out puissent aussi s'appliquer plus généralement à la minimisation du risque empirique. Cela nécessite de remplacer $|\mathcal{M}|$, le nombres d'estimateurs considérés,

par des quantités permettant de maîtriser les processus empiriques sur des ensembles infinis, telles que l'entropie à crochet, l'entropie de Rademacher ou encore la dimension de Vapnik-Chervonenkis.

Agghoo Dans cette thèse, j'ai réussi à mettre en évidence les effets de l'agrégation dans le cas particulier d'estimateurs par séries trigonométriques, en densité L^2. Il est naturel de se demander si il est possible de dégager de cette étude des principes généraux régissant le comportement de l'agrégation d'hold-out, du moins en estimation de densité L^2.

Dans le chapitre 6, on constate que plus le risque est "plat" au voisinage de son optimum, plus l'agrégation améliore le hold-out. Un objectif pour de futures recherches serait de formaliser cette condition à l'aide de la géométrie de l'espace $L^2(P)$ auquel appartiennent les estimateurs.

Alternatives à Agghoo Dans le chapitre 6, pour garantir un gain d'un facteur constant par rapport à l'oracle, il est nécessaire de choisir les paramètres d'Agghoo de façon optimale. Il semble que si ses paramètres sont mals choisis, Agghoo n'agrège pas toujours suffisamment d'estimateurs différents: les estimateurs hold-out restent trop proches de l'oracle, ce qui limite l'effet de l'agrégation. On aimerait disposer d'une méthode capable de donner par elle même une taille optimale à l'ensemble d'estimateurs qu'elle agrège. Cet objectif peut se formaliser en introduisant une famille croissante d'ensembles d'estimateurs et en se comparant, non au meilleur estimateur individuel, mais au meilleur agrégat parmis la famille d'ensembles. Le problème est alors de définir une méthode statistique dont le risque est semblable à celui du meilleur agrégat.

Chapter 2

Introduction

Elementary parametric statistics typically deals with situations in which there is a single known, finite-dimensional model and the amount of data is large enough to make an asymptotic analysis possible.

Developments in computer technology and the proliferation of data has increased the ambitions and the capacities of statistics and led statisticians to consider more general problems. Examples include regression problems where there is no single "true model" and the regression function must instead be sought in some large class of smooth functions (nonparametric statistics), or linear regression problems where the number of covariates is not small compared to the sample size (high-dimensional statistics), so that inference over the whole model is inconsistent.

These situations confront statisticians with a tradeoff between generality of the model and feasibility of estimation within it. The hope is that some simple enough model will turn out to be adequate to describe reality, so that accurate estimation is possible. As the required complexity is typically unknown, a standard practice is to introduce collections of models of various sizes and complexities. The statistician is then left with the problem of correctly choosing a model, or estimator within these collections, so as to adapt to the intrinsic complexity of the problem.

For specific model collections, it is possible to develop ad-hoc methods using theoretical calculations. However, there are many cases where this approach is impractical, either because the requisite calculations are intractable or because it requires some distribution-dependent quantities (such as the noise level) to be known a priori. It is therefore important to have "black box" procedures that do not require any information about the distribution of the data or the model collection. Such procedures are generally based on validation, i.e, witholding part of the data from the estimators to provide an independent assessment of their performance. These risk estimates can be used either to *select* a single estimator from the collection — usually by minimizing a risk estimate — or they can be

used to construct *weights* in order to form a convex combination (*aggregate*) of the estimators.

In this thesis, I study a procedure (Agghoo) which mixes elements of both approaches. Data splitting is performed several times. Each time, the validation sample is used to estimate the risk and select an estimator trained on the rest of the data. At the end, the various estimators corresponding to the different subdivisions are aggregated.

2.1 Statistical learning setting

In statistics, it is very frequent to define risk-functionals to measure the quality of estimators. As a result, many statistical problems are equivalent to minimizing a risk functional over some set $\mathcal{S}$.

Agghoo, like cross-validation, applies to such risk-minimization problems where the *risk* $R(\hat{s})$ of an estimator $\hat{s}$ can be expressed as an expectation $P\gamma(\hat{s}) := \mathbb{E}_Z[\gamma(\hat{s}, Z)]$, where γ is a known *contrast function*, Z follows the unknown distribution P and is independent from $\hat{s}$.

When the risk arises from an estimation problem, where it measures how well $\hat{s}$ estimates a distribution-dependent parameter s, the risk R should be non-negative and reach a minimum of 0 at the target value s. However, sometimes it is impossible to find a fixed contrast function γ such that $P\gamma(t) = R(t)$, but only such that $P\gamma(t) = R(t) + P\gamma(s)$, where $P\gamma(s)$ is a non-zero constant depending only on s and P, so that minimizing R is entirely equivalent to minimizing $P\gamma(t)$. R can then be expressed as the *excess risk* relative to the optimum s:

$$\ell(s,t) = P\gamma(t) - P\gamma(s) = P\gamma(t) - \operatorname*{argmin}_{t' \in \mathcal{S}} P\gamma(t').$$

A statistical problem which can be formulated in this way is *least-squares density estimation*. In this problem, we are given data Z_i distributed according to an unknown density s with respect to a known measure μ (usually the Lebesgue measure), and the goal is to construct a good approximation $\hat{s}$ of s in terms of the (squared) L^2 distance, $R(\hat{s}) = \|\hat{s} - s\|^2_{L^2(\mu)}$. Setting $\gamma(t, Z_i) = \|t\|^2_{L^2(\mu)} - 2t(Z_i)$ yields $\mathbb{E}[\gamma(t, Z_i)] = \|t - s\|^2_{L^2(\mu)} - \|s\|^2_{L^2(\mu)}$ and $\ell(s,t) = \|t - s\|^2_{L^2(\mu)}$.

In other cases, the statistical problem is directly one of risk minimization, with a given contrast function γ. This notably includes the *prediction* problem of statistical learning, in which the data consist of pairs $Z_i = (X_i, Y_i)$ and the goal is to predict the variable of interest Y using a function t of X, on an independent copy $Z = (X, Y)$. The discrepancy between the prediction $t(X)$ and the observed value Y is then measured using some function $\gamma(t, (X, Y)) = d(t(X), Y)$, and the goal is to minimize $\mathbb{E}[d(t(X), Y)]$. Here, $s \in \operatorname{argmin}_{t \in \mathcal{S}} P\gamma(t)$ denotes the optimal

predictor that could be used if P were known, called the *Bayes predictor* and $\ell(s,t) = P\gamma(t) - P\gamma(s)$ measures the performance of a given predictor t relative to this benchmark.

In both examples, the same risk-minimization formalism can be used. This formalism is interesting because it naturally gives rise to a large class of estimators. Assume that an i.i.d sample $Z_1, \ldots, Z_n$ with distribution P is available. Since the risk of t can be expressed as an expectation over the data $\mathbb{E}[\gamma(t, Z)]$, it can be estimated in a natural way by

$$P_n\gamma(t) = \frac{1}{n}\sum_{i=1}^{n}\gamma(t, Z_i).$$

If instead of a parameter t we have an estimator $\hat{t}$, the same method applies so long as the data Z_i used to estimate the risk is independent from the data used to compute the estimator $\hat{t}$.

Since the goal of estimation is to minimize the risk, a natural idea is then to compute

$$\hat{s} = \operatorname*{argmin}_{t\in\mathcal{S}} P_n\gamma(t)$$

as an estimator for $s \in \operatorname{argmin}_{t\in\mathcal{S}} P\gamma(t)$, a strategy known as *empirical risk minimization* (ERM). Empirical risk minimization typically works well in a parametric setting, when $\mathcal{S}$ is finite-dimensional (finite linear dimension in regression, or finite "Vapnik dimension" in classification [38, Chapter 12]). However, in a statistical learning setting, since generally no a priori information is available about the distribution of (X, Y), $\mathcal{S}$ is generally a very large class of functions. Some functions in this class reproduce the data exactly and hence have an empirical risk of zero, so they cannot be distinguished using the empirical-risk minimization rule. As a result, ERM over the whole set $\mathcal{S}$ is inconsistent. Statisticians have used two main approaches to turn ERM into a consistent procedure. The first is to restrict ERM to a smaller subset m of $\mathcal{S}$, called a *model*, yielding:

$$\hat{s}_m = \operatorname*{argmin}_{t\in m} P_n\gamma(t).$$

In least-squares regression for example, m is typically a vector-space of functions, such as piecewise constant functions (yielding regressograms), piecewise polynomial functions, wavelets or other classical spaces used in function approximation.

The other way to adapt the ERM to the non-parametric setting is to penalize excessively large or complex predictors t, using a *penalty* function $\Omega(t)$. For any $\lambda > 0$, this yields the estimator

$$\hat{s}_\lambda = \operatorname*{argmin}_{t\in\mathcal{S}} \left\{P_n\gamma(t) + \lambda\Omega(t)\right\},$$

where λ is called a regularization parameter. An important class of penalized empirical-risk minimizers are built from positive definite *kernels*. Suppose that the parameter s of interest is a function $s : \Xi \to \mathbb{R}$, which is the case in regression and density estimation, as well as with convex relaxations of the classification problem. Then, given a *positive definite function* $K : \Xi \times \Xi \to \mathbb{R}$ called the *kernel*, a *reproducing kernel Hilbert space* $\mathcal{H}$ of functions $\Xi \to \mathbb{R}$ can be defined (I refer the reader to the book by Scholkopf and Smola [95] for more details). The kernel estimators corresponding to the loss function γ and kernel K are the penalized empirical risk minimizers:

$$\hat{s}_\lambda = \operatorname*{argmin}_{t \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \gamma(t, Z_i) + \lambda \|t\|_{\mathcal{H}}^2 \right\}. \tag{2.1}$$

This paradigm includes support vector machines (SVM) [95, Chapter 7], support vector regression (SVR) [95, Chapter 9], kernel ridge regression and in particular, smoothing splines [108]. Another example of a penalty is $\Omega(\langle \theta, \cdot \rangle) = \|\theta\|_{\ell^1}$ in linear regression (the Lasso, introduced by Tibshirani [99]).

Both approaches raise similar questions. How large should the regularization parameter or the model be? In the model-based approach, if m is too small, the minimum of $P\gamma$ over m will be far from the global optimum; there will be a finite, positive gap

$$\operatorname*{argmin}_{t \in m} \{P\gamma(t) - P\gamma(s)\}.$$

On the other hand, an excessively large model m may make the ERM inconsistent: $\hat{s}_m$ may be far from the true minimum of the risk over m, i.e

$$\mathbb{E}\left[P\gamma(\hat{s}_m) - \operatorname*{argmin}_{t \in m} P\gamma(t) \right]$$

may be large. For penalized ERM, choosing λ too large will lead to an estimator that is excessively biased towards a "simple" and "small" estimator. For the RKHS and Lasso penalties, $\hat{s}_\lambda \to 0$ as $\lambda \to +\infty$. On the other hand, taking $\lambda = 0$ results in the unpenalized ERM, which is inconsistent. Thus, choosing correctly the model or regularization parameter is of crucial importance. Consistency in the nonparametric setting requires the model to grow, or the regularization parameter to decrease with the sample size, but not too fast. For this reason, statisticians generally consider families of penalties $(\lambda\Omega)_{\lambda>0}$ and collections of models $m \in \mathcal{M}$ instead of single penalties or models. For example, in least-squares regression, given an ordered orthonormal basis $(\varphi_j)_{j \in \mathbb{N}}$ of $L^2(\mu)$, such as the trigonometric basis, one can form the models $m_k = \langle (\varphi_j)_{1 \leqslant j \leqslant k} \rangle$ generated by the k first elements of the basis. Given such a collection, the optimal choice of regularization parameter

or model will depend on the "complexity" of the unknown parameter s: how well can s be approximated by the models under consideration? How small is $\Omega(s)$?

Theory can shed some light on the issue. For example, in least-squares density estimation on the torus, given models m_k formed from the trigonometric basis as above, it is classical that if the true density s has r square-integrable derivatives, choosing k of order $n^{\frac{1}{2r+1}}$, where n is the sample size, leads to a convergence rate of $n^{-\frac{2r}{2r+1}}$ in the L^2 norm [109] [49], which is optimal in the minimax sense. However, as s is unknown, so are its smoothness properties. The question is then to achieve the correct minimax convergence rate $n^{-\frac{2r}{2r+1}}$ without knowing r, a property known as *adaptivity*. Given a known collection of convergence rates, such as the above, it may be possible to estimate the relevant parameter (here, r) and to use a "plug-in" estimator (replacing r by its estimated value in the theoretical formula $n^{\frac{1}{2r+1}}$). In general however, the performance of the best estimator in the collection may depend on more subtle properties of s than its smoothness, properties which may not be fully understood theoretically. In that case, a general way to guarantee adaptivity with respect to s is to construct an estimator $\hat{s}$ and show that it performs as well as any of those in the collection. There are many ways to formalize this. From an asymptotic viewpoint, one may require that

$$\frac{\ell(s,\hat{s})}{\inf_{m\in\mathcal{M}} \ell(s,\hat{s}_m)} \to 1,$$

where $\to$ denotes a classical notion of convergence (L^p, convergence in probability, almost surely, etc), a property known as *asymptotic optimality*. From a non-asymptotic viewpoint, a typical goal is to prove *oracle inequalities* of the general form:

$$\mathbb{E}[\ell(s,\hat{s})] \leqslant C\mathbb{E}\left[\inf_{m\in\mathcal{M}} \ell(s,\hat{s}_m)\right] + r_n, \tag{2.2}$$

where C is a constant and r_n is a "small" remainder term. Such an inequality automatically guarantees adaptation to any convergence rate slower than r_n, the remainder term. The estimator $\hat{s}_{\hat{m}_*}$ realizing the infimum in equation (2.2), if it exists, is called the *oracle*: it is the best estimator in the collection. If $C = 1$ and r_n is negligible with respect to the risk of the oracle, the oracle inequality is said to be optimal: it shows that asymptotically, the expected risk of $\hat{s}$ is the same as that of the best estimator in the collection. The oracle inequality can be weakened somewhat by placing the expectation inside the $\inf_{m\in\mathcal{M}}$, or it can be strengthened by replacing the expectations with deviations bounds: many variants are possible.

2.2 Model selection

In order to construct an estimator $\hat{s}$ that satisfies an oracle inequality, a first class of methods proceed by estimating the risk of the estimators $(\hat{s}_m)_{m\in\mathcal{M}}$ (or $(\hat{s}_\lambda)_{\lambda>0}$), and selecting one of them with low estimated risk: this is called *model selection*. If a *validation sample* $Z_1, \ldots, Z_{n_v}$ is available, and it is independent from the estimators $\hat{s}_m$, then the risk of the $\hat{s}_m$ can simply be evaluated using the empirical mean, as in the definition of the ERM:

$$P\gamma(\hat{s}_m) \approx \frac{1}{n_v}\sum_{i=1}^{n_v}\gamma(\hat{s}_m, Z_i).$$

This estimator is unbiased and consistent (at least in the asymptotic where $n_v \to +\infty$ while holding $\hat{s}_m$ fixed). It is then possible to select a model $\hat{m}$ by *empirical risk minimization* on the collection $(\hat{s}_m)_{m\in\mathcal{M}}$. What if there is no validation sample? The empirical risk $P_n\gamma(\hat{s}_m)$ cannot be used, as it is heavily biased; indeed, in the penalty case, it is easy to see that minimizing the empirical risk over λ simply results in selecting the smallest allowed value of λ. In some cases, it is possible to calculate theoretically a deterministic correction $pen(m)$ (usually a function of the model dimension), so that $P_n\gamma(\hat{s}_m) + pen(m)$ becomes a good risk estimator. Two famous examples are Akaike's AIC [1] and Mallows' C_p [74] . An overview of the classical methods can be found in [16] and more involved theory in Massart [76]. A weakness of this approach is that it lacks generality: Mallows' theory is specific to linear estimators in least-squares regression, whereas AIC relies on the concept of linear dimension, which may be inappropriate for penalty-based methods or in classification. For example, Kearns et al [61] showed that no penalty based on the dimension alone could lead to a universally consistent procedure in classification. Moreover, the theoretical penalties generally involve nuisance parameters that have to be estimated, such as the variance. This means that often, a second estimator selection procedure is used to calibrate this unknown constant.

2.3 Hold-out

A simpler approach is to form a validation sample out of the available data, sacrificing some performance (by training the estimators on a reduced dataset) in order to gain a distribution-free, unbiased risk estimator. More precisely, given a subset $T \subset \{1 \ldots n\}$, let $D_n^T = (Z_i)_{i\in T}$ denote the corresponding subsample. Given a collection $(\hat{s}_m)_{m\in\mathcal{M}}$ of estimators, the risk of $\hat{s}_m(D_n^T)$ can be estimated empirically by

$$\mathrm{HO}_T(m) = \frac{1}{n-|T|}\sum_{i\notin T}\gamma(\hat{s}_m(D_n^T), Z_i).$$

Then, selecting the estimator with the lowest estimated risk yields:

$$\widehat{f}_T^{\text{ho}}(\mathcal{M}, D_n) = \hat{s}_{\hat{m}_T}(D_n^T), \text{ where } \hat{m}_T \in \operatorname*{argmin}_{m\in\mathcal{M}} \text{HO}_T(m).$$

Considering $\text{HO}_T(m)$ as a biased estimator of $P\gamma(\hat{s}_m(D_n))$ instead of an unbiased estimator of $P\gamma(\hat{s}_m(D_n^T))$ leads to the slightly different result $\hat{s}_{\hat{m}_T}(D_n)$: the standard terminology does not distinguish between these two procedures (for example, the survey article [3] defines the hold-out estimator of the risk, but does not explicitly specify how the final estimator is constructed). While in practice, $\hat{s}_{\hat{m}_T}(D_n)$ might be better, as it is evaluated on more data, very little can be said about its risk in general, when the estimators $\hat{s}_m$ are arbitrary. On the other hand, the risk of $\widehat{f}_T^{\text{ho}}$ is well-estimated by $\text{HO}_T(\hat{m}_T)$, with only a moderate bias if $\mathcal{M}$ is not too big. For this reason, the theoretical literature on the hold-out, discussed below, has mostly focused on $\widehat{f}_T^{\text{ho}}$, and so will I. For similar reasons, $\widehat{f}_T^{\text{ho}}$ can only be compared to $\inf_{m\in\mathcal{M}} \ell(s, \hat{s}_m(D_n^T))$ in general, not to $\inf_{m\in\mathcal{M}} \ell(s, \hat{s}_m(D_n))$. The two can only be related under some stability conditions on the estimators $\hat{s}_m$ [60]. Thus, for the hold-out, potential general oracle inequalities are of the form:

$$\mathbb{E}\left[\ell(s, \widehat{f}_T^{\text{ho}})\right] \leqslant C\mathbb{E}\left[\inf_{m\in\mathcal{M}} \ell(s, \hat{s}_m(D_n^T))\right] + r_n. \tag{2.3}$$

2.3.1 Existing theory

Because $\text{HO}_T(m)$ is an empirical average, if the data is i.i.d and γ is uniformly bounded, Hoeffding's inequality implies that, with high probability,

$$\sup_{m\in\mathcal{M}} |\text{HO}_T(m) - P\gamma(\hat{s}_m)| = \mathcal{O}\left(\sqrt{\frac{\log(|\mathcal{M}|)}{n - |T|}}\right). \tag{2.4}$$

As a result, one can easily show that

$$\ell(s, \widehat{f}_T^{\text{ho}}) \leqslant \min_{m\in\mathcal{M}} \ell(s, \hat{s}_m(D_n^T)) + \mathcal{O}\left(\sqrt{\frac{\log(|\mathcal{M}|)}{n - |T|}}\right). \tag{2.5}$$

This oracle inequality is not always optimal, because the remainder term is not always negligible: for example, if γ is the square loss of regression, then the parametric convergence rate is $\mathcal{O}(\frac{1}{n})$, not $\mathcal{O}(\frac{1}{\sqrt{n}})$ and many non-parametric classes also have worst-case convergence rates faster than $\frac{1}{\sqrt{n}}$ (for example, $\alpha-$ Hölder functions with $\alpha > \frac{1}{2}$).

Massart showed that one can improve the oracle inequality (2.5) by assuming in addition that the variance of $\text{HO}_T(m)$ is bounded by a function of the excess

risk of $\hat{s}_m$ [76, Corollary 8.8]. More precisely, this *margin assumption* states that there exists a non-decreasing, sublinear function w such that

$$\forall m, m' \in \mathcal{M}, \mathrm{Var}_Z\left(\gamma(\hat{s}_m, Z) - \gamma(\hat{s}_{m'}, Z)\right) \leqslant \left(w(\sqrt{\ell(s, \hat{s}_m)}) + w(\sqrt{\ell(s, \hat{s}_{m'})})\right)^2. \tag{2.6}$$

As a result, minimizing the risk $\ell(s,t)$, as the hold-out attempts to do, simultaneously reduces the variance: this results in a smaller remainder term than suggested by the uniform bound (2.4). Thus, Massart [76] proved that under the margin assumption, the hold-out satisfies an oracle inequality (2.3) with a remainder term r_n which can be bounded, up to log terms, by the solution of the equation

$$w(u) = \sqrt{n - |T|}u^2.$$

2.3.2 Contributions to the general theory of the hold-out

In Chapter 3, Theorem 3.7.2 of this thesis, I extend Massart's result in two ways. First, I show that the uniform boundedness assumption on γ can be relaxed to a second margin assumption, of the form:

$$\|\gamma(\hat{s}_m, \cdot) - \gamma(\hat{s}_{m'}, \cdot)\|_\infty \leqslant w_2(\sqrt{\ell(s, \hat{s}_m)}) + w_2(\sqrt{\ell(s, \hat{s}_{m'})}).$$

This allows the uniform norm $\|\gamma(\hat{s}_m, \cdot)\|_\infty$ to depend on $m \in \mathcal{M}$ and on the sample size n. Secondly, I show that, in the margin assumption (2.6), the sublinear function w can be replaced with a subquadratic function w, with some caveats. These caveats are that only assuming w subquadratic leads only to an oracle inequality valid with high probability $1 - \frac{1}{n^2}$ (for example), while controlling the expectation requires Massart's original assumption to hold for some other function w', leading to a second order correction.

The significance of these changes is to provide better theoretical support for the application of the hold-out to unbounded loss-functions γ, most particularly in regression. In regression, $\gamma(\hat{s}_m, (x,y)) = \phi(y - \hat{s}_m(x))$ where ϕ is typically convex and unbounded on $\mathbb{R}$: for example, $\phi(x) = x^2$ (least-squares) or $\phi(x) = |x|$ (least-absolute deviations). In that case, $\gamma(\hat{s}_m, (x,y))$ is generally unbounded unless $Y \in L^\infty$ and $\|\hat{s}_m\|_\infty < A$ for some constant A. Choosing $\hat{s}_m$ uniformly bounded a priori is impractical since the regression function s might be unbounded, and even if it is bounded, $\|s\|_\infty$ is generally unknown. Having $\|\hat{s}_m\|_\infty$ depend on m and n is necessary to obtain a consistent estimator in this case. Now, if the norms $\|\gamma(\hat{s}_m, \cdot)\|_\infty$ are *not* uniformly bounded, then the margin assumption (2.6) generally cannot hold for a fixed sublinear function w. This is because $\mathrm{Var}\left(\gamma(\hat{s}_m, Z) - \gamma(\hat{s}_{m'}, Z)\right)$ is a quadratic function of $\gamma(\hat{s}_m, Z)$, whereas $\ell(s, \hat{s}_m) = P\gamma(\hat{s}_m) - P\gamma(s)$ is a linear function of $\gamma(\hat{s}_m, Z)$. Extending the margin assumption to subquadratic w allows

to resolve this difficulty. Applications include kernel methods with Lipschitz loss functions (discussed in section 2.9.2) and sparse linear regression with the Huber loss (discussed in section 2.9.2).

2.4 Cross-validation

Though the hold-out lends itself well to theoretical investigation, two drawbacks have limited its use in practice. Firstly, there is the need to remove sufficient data from the training set, which typically reduces the performance of the estimators. Secondly, the hold-out requires the choice of a training subset T out of the full dataset $\{1 \dots n\}$. For a given cardinality $|T| = n_t$, the choice of T does not affect the distribution of $\mathrm{HO}_T(m)$ or $\widehat{f}_T^{\mathrm{ho}}$, so we can equivalently assume that it is random and uniformly distributed on the set $\{T \subset \{1 \dots n\} : |T| = n_t\}$. This shows that T is a source of variability for $\widehat{f}_T^{\mathrm{ho}}$ which affects simultaneously the risk estimator $\mathrm{HO}_T(m)$ and the $\hat{s}_m(D_n^T)$. This variability is likely to be particularly strong when the estimators $\hat{s}_m$ are *unstable* (i.e, when a small change in the input sample results in a big change in their output) [37] [60].

To reduce this variability, practitioners often resort to averaging $\mathrm{HO}_T(m)$ over several splits of the data in order to obtain a more stable risk estimator, a strategy known as *cross-validation*. Modern cross-validation procedures were first introduced by Stone [98] and Geisser [45]. Let $[n]$ denote the set $\{1, \dots, n\}$. Let $\mathcal{T} = (T_j)_{1 \leqslant j \leqslant V}$ be a finite sequence of subsets of $[n]$ (usually, but not necessarily, of the same size n_t). Geisser [45] defines the general CV risk estimator

$$CV_{\mathcal{T}}(m) = \frac{1}{V} \sum_{j=1}^{V} \mathrm{HO}_{T_j}(m) .$$

This includes the leave-one out procedure ($\mathcal{T} = ([n] \backslash j)_{1 \leqslant j \leqslant n}$), Monte-Carlo CV ($(T_j)_{1 \leqslant j \leqslant V}$ i.i.d uniform among subsets of size n_t) or V-fold CV ($\mathcal{T} = ([n] \backslash I_j)_{1 \leqslant j \leqslant V}$ for a partition $(I_j)_{1 \leqslant j \leqslant V}$ of $[n]$ into subsets of cardinality n/V), among the most classical CV procedures. A more detailed description of existing CV procedures can be found in the survey by Arlot and Celisse [3]. When used for model selection, the cross-validation procedure is to select $\hat{m}_{\mathcal{T}}^{cv} \in \operatorname{argmin}_{m \in \mathcal{M}} CV_{\mathcal{T}}(m)$, as with most methods that use risk estimation. For building a final estimator from $\hat{m}_{\mathcal{T}}^{cv}$, there are several options. Unlike with the hold-out, there is no distinguished training set (all sets in $\mathcal{T}$ are "equal"), so the final estimator usually respects this symmetry. By far the most standard choice, certainly in practice, is to take $\widehat{f}_{\mathcal{T}}^{\mathrm{cv}} := \hat{s}_{\hat{m}_{\mathcal{T}}^{cv}}(D_n)$, which has the advantage of yielding an estimator trained on the whole sample. Because CV estimates the risk of estimators trained on subsamples $T \in \mathcal{T}$, the theoretical

literature sometimes discusses other estimators built from the $(\hat{s}_{\hat{m}^{cv}_{\mathcal{T}}}(D_n^T))_{T\in\mathcal{T}}$, such as the bagged variant $\frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}}\hat{s}_{\hat{m}^{cv}_{\mathcal{T}}}(D_n^T)$ [69].

Like the hold-out, CV procedures must remove some data from the training set in order to perform risk estimation. However, CV procedures typically require far less data to be set aside for validation compared to the hold out. As a result, estimators can almost be trained with the full dataset. A striking example of this difference is that while the "leave-1-out hold-out" ($|T| = n-1$) is obviously absurd, *leave-1-out* cross-validation ($|T_j| = n-1$) is a classical, respected procedure which works quite well in practice [3]. Because of this, one might expect cross-validation to satisfy oracle inequalities (2.2) with respect to the full data oracle instead of equations of the type (2.3) in which the oracle is trained on a sample of size $n_t < n$. However, no result of this type has been stated in full generality, presumably because relating $\hat{s}_m(D_n^T)$ to $\hat{s}_m(D_n)$ requires non-trivial assumptions of "algorithmic stability" on $\hat{s}_m$, even when $|T| = n-1$.

Instead, van der Vaart, Dudoit and van der Laan [105] have proved a general oracle inequality very similar to that of Massart for the hold-out ([76, Corollary 8.8]), discussed above. Their upper bound is an inequality of the form (2.3) with an adjustable constant $C > 1$ and a remainder term r_n that is finite if and only if a margin hypothesis (2.6) holds with $w(u) = cu^\theta$, for some $\theta \in [0;1]$. Because of the generality of the setting they consider, their upper bound applies, not to the risk of the more common CV procedure $\widehat{f}^{\mathrm{cv}}_{\mathcal{T}} = \hat{s}_{\hat{m}^{cv}_{\mathcal{T}}}$, but to the quantity

$$\frac{1}{V}\sum_{i=1}^{V} P\gamma\left(\hat{s}_{\hat{m}^{cv}_{\mathcal{T}}}(D_n^{T_i})\right),$$

which can be seen as the risk of the estimator $\hat{s}_{\hat{m}^{cv}_{\mathcal{T}}}(D_n^{T_J})$, where J is an auxiliary random variable distributed uniformly among $1,\ldots,V$. More recently, Lecué and Mitchell [69] have proved a similar oracle inequality to [105] under slightly different margin and moment assumptions. Like van der Vaart et al, Lecué and Mitchell build their alternative CV procedure from the $\hat{s}(D_n^{T_j})$. However, instead of selecting a random subset T_J among the given collection $(T_j)_{1\leqslant j\leqslant V}$, they consider the bagged variant:

$$\frac{1}{V}\sum_{i=1}^{V}\hat{s}_{\hat{m}^{cv}_{\mathcal{T}}}(D_n^{T_j}),$$

under the assumption that the risk $P\gamma$ is convex, and the asymmetric choice $\hat{s}_{\hat{m}^{cv}_{\mathcal{T}}}(D_n^{T_1})$, $T_1 = \{1,\ldots,\frac{V-1}{V}n\}$ for $V-$fold CV. Lecué and Mitchell also discuss a general hypothesis under which the normal CV procedure $\hat{s}_{\hat{m}^{cv}_{\mathcal{T}}}(D_n)$ satisfies the same oracle inequality as their variants.

As in Massart's oracle inequality for the hold-out [76], these oracle inequalities have remainder terms r_n of order $\frac{\log|\mathcal{M}|}{n_v}$ when the margin hypothesis holds for a

linear function $w(u) = cu$, where $n_v = n - |T|$ denotes the size of the validation sample $D_n^{T^c}$. Because r_n depends on n_v, not on n, these results are uninformative when applied to leave-one-out or leave-p-out cross-validation, for which $n_v = 1$ and $n_v = p$, respectively. According to Kearns and Ron [60], similarly general results on the leave-one-out would require an additional assumption of "algorithmic stability" on the estimators $\hat{s}_m$.

2.5 Model and hyperparameter aggregation

The methods discussed thus far all seek to identify one good estimator among the collection $(\hat{s}_m)_{m\in\mathcal{M}}$. Since the aim is to perform as well as the best estimator among the given collection, this is quite a natural idea; however, there is no obligation to restrict oneself to such estimators. Instead, *aggregation* procedures consider estimators that are weighted sums of the $(\hat{s}_m)_{m\in\mathcal{M}}$, i.e

$$\hat{s} = \sum_{m\in\mathcal{M}} \hat{w}_m \hat{s}_m \tag{2.7}$$

with data-dependent weights $(\hat{w}_m)_{m\in\mathcal{M}} \in \mathbb{R}^{\mathcal{M}}$. One speaks of convex aggregation when the weights are required to be non-negative and sum to 1, and of linear aggregation when no constraint is imposed on the weights [6, Chapter 3].

In order for formula (2.7) to make sense, the $\hat{s}_m$ should belong to a vector space (for linear aggregation) or to a convex set (for convex aggregation). Moreover, for convex aggregation to be appropriate, the risk should be convex, since this guarantees that for any convex combination $\sum_{m\in\mathcal{M}} \hat{w}_m \hat{s}_m$,

$$P\gamma\left(\sum_{m\in\mathcal{M}} \hat{w}_m \hat{s}_m\right) \leqslant \sum_{m\in\mathcal{M}} \hat{w}_m P\gamma(\hat{s}_m).$$

If this inequality is not satisfied, this means that drawing one $\hat{m}$ at random according to the distribution $(\hat{w}_m)_{m\in\mathcal{M}}$ is preferable to forming the convex combination $\sum_{m\in\mathcal{M}} \hat{w}_m \hat{s}_m$. Hence, in the non-convex case, model selection may be preferable to aggregation. Settings where the risk $P\gamma$ is convex and estimators $\hat{s}_m$ belong to a vector space include regression and density estimation, for most commonly used loss functions, and also convex relaxations of the classification problem, which are commonly used in practice. Aggregation can potentially lead to large improvements in such cases because the convex hull or linear span of $(\hat{s}_m)_{m\in\mathcal{M}}$ is a much larger set than $(\hat{s}_m)_{m\in\mathcal{M}}$, hence the optimal aggregate may perform much better than the best $\hat{s}_m$. However, this improvement does not come for free in general: with the larger space comes a correspondingly larger difficulty of estimating the optimal convex aggregate, compared to the optimal $\hat{s}_m$. Tsybakov [102] showed that

if the best $\hat{s}_m$ is replaced by the best convex combination in the oracle inequality (2.3), then the remainder term r_n is necessarily of order $\sqrt{\frac{\log|\mathcal{M}|}{n_v}}$ in the worst case, where n_v denotes the size of the validation sample. As faster convergence rates are achieved over many smooth non-parametric classes, trying to perform optimal aggregation may be unnecessary or even detrimental, especially if the estimators in the collection are known to perform well and converge fast.

However, aggregation may be of interest even when the aim is to match the performance of the best $\hat{s}_m$ (model selection oracle), instead of the best convex combination. In density estimation, Catoni [28] and Yang [117] independently proposed the "progressive mixture rule" and showed that it satisfies an oracle inequality (2.3) with optimal constant $C = 1$ and remainder term $\mathcal{O}\left(\frac{\log|\mathcal{M}|}{n_v}\right)$. By Tsybakov [102], this remainder term is optimal in the worst case. The progressive mixture rule was extended to least-squares regression by Yang [116] and Catoni [29], and a generalized version called "mirror averaging" was later defined by Juditsky, Rigollet and Tsybakov [57], who gave sufficient conditions on a general loss function γ to obtain an optimal oracle inequality. Given a validation sample $(Z_1 \dots Z_{n_v})$, this general procedure builds weights $\hat{w}_m$ according to the formula:

$$\hat{w}_m = \frac{1}{n_v} \sum_{k=1}^{n_v} \frac{\exp\left(-\beta P_k \gamma(\hat{s}_m)\right)}{\sum_{m \in \mathcal{M}} \exp\left(-\beta P_k \gamma(\hat{s}_m)\right)} \tag{2.8}$$

$$\text{where} \qquad P_k \gamma(t) = \frac{1}{k} \sum_{j=1}^{k} \gamma(t, Z_j)\,. \tag{2.9}$$

The weights $\hat{w}_m$ are exponentially decreasing functions of the empirical risks $P_k\gamma(\hat{s}_m)$, so as expected, most of the weight is located where the risk is small.

Compared to model selection methods which use the same validation sample, mirror averaging and progressive mixture rules satisfy sharper oracle inequalities: indeed, Catoni [29] and Juditsky et al [57] remark that no selection algorithm can attain the optimal $\frac{\log|\mathcal{M}|}{n_v}$ convergence rate for the remainder term r_n in an oracle inequality (2.3) with optimal constant $C = 1$. However, these procedures rely, like the hold-out, on splitting the data between a training sample and a validation sample, with the same disadvantages compared to model selection methods which use the whole data for training. Moreover, outside density estimation, these aggregation procedures have free parameters which must be set based on a priori knowledge about the data distribution in order to obtain the desired optimality properties. For example, in least-squares regression, Juditsky et al [57] assume that there are known upper bounds on the exponential moments of the noise $Y - s(X)$ and on the uniform norm of the regression function s.

Overall, this type of aggregation remains quite similar to model selection in its methods and goals, as well as its theoretical guarantees. Both are designed

for ensembles where one of the estimators will perform well in any given situation — in particular, collections that are *adaptive* to classical non-parametric classes. This is why the oracle is the benchmark. In both cases, the idea is to identify one or a few very good estimators among a given collection — either by choosing one or by weighting estimators according to their empirical risk.

2.6 Randomized aggregation

A different problem occurs when aggregation is applied to weak estimators. In that case, matching the performance of the oracle is insufficient. Instead, the hope is that combining an ensemble of individually poor estimators may yield an aggregate that is (much) better than all of them.

One cause of poor performance of an estimator $\hat{s}(D_n)$ is instability or, in other words, high *variance* with respect to the training data D_n. Breiman [22] suggested that this variance might be reduced by aggregating an ensemble $\hat{s}(D^*_{n,1}) \dots \hat{s}(D^*_{n,B})$ formed by evaluating $\hat{s}$ over different perturbations of the initial sample D_n. More precisely, Breiman suggested resampling uniformly and with replacement from D_n, forming the resamples $D^*_n = (X_{I_1}, \dots, X_{I_n})$, where $(I_j)_{1 \leqslant j \leqslant n}$ are i.i.d indices, uniform over $[n]$. The bagged estimator $\hat{s}^{bag}$ is then the uniform aggregate over these resamples, i.e

$$\hat{s}^{bag} = \mathbb{E}_*[\hat{s}(D^*_n)] = \mathbb{E}[\hat{s}(D^*_n)|D_n].$$

As this estimator is typically difficult to compute, it is often replaced with the Monte-Carlo approximation

$$\hat{s}^{bag}_B = \frac{1}{B} \sum_{j=1}^{B} \hat{s}(D^*_{n,j}),$$

where the $(D^*_{n,j})_{1 \leqslant j \leqslant B}$ are resamples drawn independently from D_n, conditionally on D_n. An alternative is subagging [25], i.e drawing subsamples without replacement, yielding $D^{T_j}_n = \{(X_i)_{i \in T_j}\}$, where T_j are i.i.d subsets of some size $n_t < n$. The subagged estimator is then

$$\hat{s}^{sub}_B = \frac{1}{B} \sum_{j=1}^{B} \hat{s}(D^{T_j}_n).$$

Initial theoretical work on bagging (Friedman and Hall [44], Buja and Stuetzle [26]) focused on cases where the estimator $\hat{s}$ has a Taylor expansion with respect to the empirical distribution P_n, making exact computations possible. These authors showed that under this assumption, bagging has second-order effects only. It can

decrease variance and mean-squared error under some conditions, while generally increasing squared bias. However, the main interest in bagging lies in its application to less regular estimators, such as trees or neural networks, where greater improvements may be possible. Bühlmann and Yu [25] showed that bagging can reduce variance and mean-squared error by a constant factor when applied to discontinuous estimators, namely thresholded least-squares estimators and CART trees of bounded depth (a type of regressogram). More surprisingly, bagging can turn inconsistent estimators into consistent ones: Biau and Devroye [13] showed the consistency of the subagged nearest-neighbour estimator in regression, when the size of the subsamples is chosen appropriately. Biau and Guyader [14] went on to prove that this estimator in fact converges at the optimal rate, at least when the size of the subsamples is optimally chosen. Thus, there is theoretical support for Breiman's heuristic that bagging leads to improvements for unstable estimators. Bagging works by randomizing the sample with respect to which $\hat{s}$ is unstable, and aggregating the results.

More generally, one might consider randomizing and aggregating other unstable components of a statistical algorithm. In particular, several authors suggested randomizing the construction of CART trees, culminating in the seminal paper by Breiman [23], which introduced the random forest method. Random forests, which combine bagging with randomized tree construction, have been very successful in practice, to the point that they are considered one of the best "black box", generalist machine learning methods [12]. Theoretical results show that aggregation of trees can bring substantial improvements over single trees. Arlot and Genuer [4] showed that for a variety of randomized trees, aggregation increases the rate of convergence compared to single trees.

2.7 Aggregated hold-out

2.7.1 Description of the procedure

Basic idea The literature on randomized aggregation dramatically illustrates how basic estimators such as trees, nearest neighbour or regular regressograms can be turned into high-performance estimators, by being randomized and aggregated. In the context of model selection, this suggests that the simple hold-out, neglected in practice on account of its instability, might in fact be turned into an efficient aggregation method for model selection. The instability of the hold-out is partly due to its dependence on a free parameter, the subset $T \subset [n]$ which is used to split the data into a training sample (D_n^T) and a validation sample. Randomizing this parameter and aggregating may serve to reduce the variance and expected risk of the hold-out.

Reducing the risk If the data are i.i.d and T is fixed (or independent from the sample D_n), the distribution of the hold-out estimator $\widehat{f}_T^{\text{ho}}$ only depends on T through its cardinality. As a result, *aggregating* hold-out estimators $\widehat{f}_{T_j}^{\text{ho}}$, for some collection of subsets $(T_1 \dots T_V)$ with the same cardinality $|T_j| = n_t$, is bound to reduce the risk compared to a single estimator, as long as the risk $P\gamma(\cdot)$ is convex. Assuming that it is, Jensen's inequality implies that for a collection $(T_1 \dots T_V)$ such that $|T_1| = \dots = |T_n| = n_t$,

$$\mathbb{E}\left[P\gamma\left(\frac{1}{V}\sum_{j=1}^{V}\widehat{f}_{T_j}^{\text{ho}}\right)\right] \leqslant \frac{1}{V}\sum_{j=1}^{V}\mathbb{E}\left[P\gamma(\widehat{f}_{T_j}^{\text{ho}})\right] = \mathbb{E}\left[P\gamma(\widehat{f}_{T_1}^{\text{ho}})\right]. \tag{2.10}$$

Definitions The above considerations lead to the definition of aggregated hold-out (Agghoo) as a procedure which computes an estimator of the form

$$\widehat{f}_{\mathcal{T}}^{\text{ag}} = \frac{1}{V}\sum_{j=1}^{V}\widehat{f}_{T_j}^{\text{ho}}$$

for some collection $\mathcal{T} = (T_j)_{1\leqslant j\leqslant V}$ of subsets of $[n]$, independent from the sample D_n. In this thesis, I only consider collections $\mathcal{T}$ consisting of subsets T_j of equal cardinality n_t, as this ensures that $\widehat{f}_T^{\text{ag}}$ improves on the hold-out for a given size $n_t = |T_j|$ of the training sample $D_n^{T_j}$. As for cross-validation, there are several ways to generate a collection of subsets $\mathcal{T}$. By analogy with cross-validation, the term "V-fold Agghoo" is used when $\mathcal{T} = ([n]\backslash I_j)_{1\leqslant j\leqslant V}$, where the validation subsets $(I_j)_{1\leqslant j\leqslant V}$ are disjoint and of equal cardinality $n - n_t$. "Monte-Carlo Agghoo" refers to collections $\mathcal{T}$ made of independent random subsets T_j of the same size n_t, drawn uniformly from $[n]$.

Links between Agghoo, aggregation and cross-validation Aggregated hold-out combines elements of cross-validation, model selection aggregation, and bagging / randomized aggregation. Agghoo aggregates hold-out estimators and does so by varying the subset T used to split the sample, just like cross-validation. Like model-selection aggregation, Agghoo aggregates several different estimators from a given collection $(\hat{s}_m)_{m\in\mathcal{M}}$, and puts greater weight on estimators with low empirical risk on some validation sample $D_n^{T_j^c}$. Finally, Agghoo can be obtained by *randomizing* the parameter T of the hold-out, and aggregating the resulting ensemble. Moreover, just like subagging, Agghoo aggregates estimators that have been trained on different subsamples $D_n^{T_j}$.

2.7.2 Agghoo in context

Agghoo and similar methods in the litterature

Agghoo and related methods have already been proposed in specific contexts and in applications. In an article on the hold-out in least-squares regression, Wegkamp [114] suggested aggregating the hold-out as a sure way to decrease its risk, by Jensen's inequality. Jung and Hu [59] proposed the use of "K-fold averaging cross-validation" (AKCV) in three settings: model selection for least-squares linear regression, selection of the Lasso regularization parameter and selection of the regularization parameter for cubic splines. In model selection, AKCV averages the regression coefficients $\hat{\beta}_m$ obtained by least-squares regression on the model m, while Agghoo averages the predictors $X\hat{\beta}_m$: by linearity, these two methods are equivalent. On the other hand, for the penalized estimators $(\hat{s}_\lambda)_{\lambda>0}$, Jung and Hu advocate for averaging the hyperparameter λ rather than the estimator $\hat{s}_\lambda$. Apart from this difference, the structure of their algorithm is identical to K-fold Agghoo, involving multiple uses of the hold-out over different folds T_j. A similar "parameter aggregation" idea was proposed by Hall and Robinson [51] in the context of kernel density estimation. They proposed (su)bagging the kernel bandwith $\hat{h}_{CV}$ obtained by minimization of the CV criterion, and multiplying the result by a deterministic factor to correct for the bias. Since CV is bagged, rather than the hold-out, their method is more computationally intensive than the hold-out; it also bags a more stable estimator (CV), which may be less advantageous according to the "randomized aggregation" philosophy. Petersen et al [87] also studied bagged cross-validation, applied in this case to CART trees parametrized by their depth. Bagged cross-validation was considered as one of four alternatives. Unlike Hall and Robinson [51], Petersen et al [87] applied bagging to the final estimators, not to the hyperparameters chosen by CV, which makes their version of "bagged cross-validation" much more similar to Agghoo. A fourth combination of aggregation and cross-validation, "Efficient K-fold cross-validation" (EKCV) was proposed by Jung [58]. EKCV is a weighted aggregation scheme similar to those discussed in Section 2.5, except that it uses K-fold cross-validation instead of a validation sample to estimate the risk of the given estimators.

Agghoo has also been discussed in an applied setting, by Varoquaux et al [107] and Hoyos-Idrobo et al [54], both for neuroimaging. Hoyos-Idrobo et al proposed an algorithm for sparse logistic regression which combines clustering, subsampling and the Lasso penalty. They use the hold-out to select the Lasso parameter within each fold and aggregate the results, in the same spirit as Agghoo, though the intermediate use of clustering makes it hard to see whether their algorithm is an exact application of Agghoo. Varoquaux et al [107] investigate Agghoo in a methodological paper which compares different parameter tuning strategies for SVM or

logistic regression with an ℓ^1 or ℓ^2 penalty. They consider "CV + averaging" as an alternative to cross-validation, which they define as follows: "we select for each split the model that minimizes the corresponding test error and average the models across splits". Depending on what is meant by "averaging the models", this could refer to Agghoo — in any case, the concept is very similar.

Comparison between Agghoo and similar procedures

As shown above, there are many ways to combine cross-validation with aggregation. Agghoo stands out because of the simplicity and generality of its definition. Compared to AKCV, EKCV and other methods which average hyperparameters instead of estimators, Agghoo only requires specification of a risk function γ and a collection of estimators $(\hat{s}_m)_{m\in\mathcal{M}}$. In contrast, aggregating hyperparameters only makes sense if the set $\mathcal{M}$ is convex — which is not the case when the parameter m is a set (as in model selection) or an integer (as in $k-$nearest neighbours, for example). Moreover, hyperparameter aggregation depends on the specific indexation of the set $\{\hat{s}_m : m \in \mathcal{M}\}$, whereas Agghoo does not. This may be a source of ambiguity when several parametrizations are possible, as for penalty-based methods $\hat{s}_\lambda = \operatorname{argmin}_{t\in E}\{P_n\gamma(t) + \lambda\Omega(t)\}$, which often have an alternative "constrained formulation"

$$\hat{s}_C = \operatorname*{argmin}_{t\in E:\Omega(t)\leqslant C} P_n\gamma(t),$$

as well as Lagrangian dual problems. For example, support-vector machines can be parametrized either by the regularization parameter λ or by a constraint parameter C (see [95]). Agghoo will yield the same result whatever the parametrization — at least if it is performed over the whole regularization path $(\hat{s}_\lambda)_{\lambda>0}$ — whereas hyperparameter aggregation requires the user to choose whether to average λ or C. Compared to bagged cross-validation defined by Petersen et al [87], Agghoo is less computationally expensive, since it performs only a single fold of cross-validation on each bagging subsample. Moreover, as it aggregates the hold-out, which is less stable than cross-validation, Breiman's heuristic for bagging [23] suggests that it may yield greater improvements.

In practice, the performance of Agghoo and similar methods has generally been good. Jung and Hu [59] found lower risks for AKCV compared to K-fold CV (with the same K) in all the settings which they considered (model selection for linear regression, the Lasso and smoothing splines). Varoquaux et al [107] found that "CV + averaging" performs well in terms of prediction error, especially in sparse models and when overall prediction accuracy is poor. They also noted that averaging provides a noticeable improvement in stability compared to ordinary cross-validation. My simulations in the sparse-regression setting (Chapter 4) confirm these findings: Agghoo can perform significantly better than cross-validation,

and its advantage is greatest when overall prediction accuracy is low. Petersen et al [87] provide a more negative assessment: they conclude that it is better to first bag CART trees, then select their depth by cross-validation, rather than to bag the cross-validation procedure itself. Their explanation is that bagging changes the bias-variance tradeoff between different depths, leading to a suboptimal choice of depth if cross-validation is used on non-bagged trees.

This conclusion assumes that bagging is the main factor driving performance. While this seems to be the case for CART trees (at least in the simulations of [87]), in general, Agghoo does not just aggregate single estimators $\hat{s}_m(D_n^{T_j})$ trained on different datasets, but different estimators $\hat{s}_{\hat{m}_j}(D_n^{T_j})$. Hence, bagging is not the only factor affecting its performance, there is also aggregation over different hyperparameters $\hat{m}_j$. If the estimators $\hat{s}_m(D_n)$ are *stable* as functions of the data D_n, but *unstable* with respect to their parameter m, then the gains due to aggregating different estimators $\hat{s}_m$ may significantly outweigh the benefits of bagging. This aggregation of different estimators $\hat{s}_{\hat{\lambda}_j}$ is also a potential advantage of Agghoo compared to methods which aggregate different hyperparameters $\hat{\lambda}_j$. By definition, the output of such a method belongs to the original collection $(\hat{s}_\lambda)_{\lambda>0}$, whereas the output of Agghoo belongs to its convex envelope. Thus, a method based on hyperparameter averaging can never outperform the model selection oracle $\operatorname{argmin}_{t\in\{\hat{s}_\lambda:\lambda>0\}} P\gamma(t)$, whereas this is possible for Agghoo.

For example, consider the problem of estimating a density $s \in L^2([0;1])$ that is symmetric ($s(\frac{1}{2}+x) = s(\frac{1}{2}-x)$) using empirical orthogonal projections on the cosine basis $\varphi_j(x) = \sqrt{2}\cos(2\pi jx)$. The estimators,

$$\hat{s}_k(D_n) = 1 + \sum_{j=1}^{k} (P_n\varphi_j)\varphi_j\,, \qquad 1 \leqslant k \leqslant n\,,$$

are linear with respect to the empirical distribution P_n, so bagging has no effect: at best, it recovers the original estimator trained on the full dataset. Thus, Petersen et al's proposal of cross-validating bagged estimators [87] is equivalent in this case to ordinary cross-validation. Moreover, as the parameter is an integer, averaging it is impossible, and if an ad-hoc aggregation method is chosen (for example, taking the median), it will necessarily choose one element of the original collection $(\hat{s}_k)_{1\leqslant k\leqslant n}$.

On the other hand, Agghoo constructs an average of the $\hat{s}_k$, which can potentially perform better than any of them. In least-squares density estimation, the following simulation shows that this can indeed happen. An i.i.d sample of size $n = 1000$ was generated according to the density $s(x) = \frac{5}{2}\mathbb{I}_{|x-\frac{1}{2}|\leqslant\frac{1}{5}}$ and Monte-Carlo Agghoo and CV were applied to the collection $(\hat{s}_k)_{1\leqslant k\leqslant n}$ with different parameters $\tau = \frac{n_t}{n}$ and V (the number of splits used).

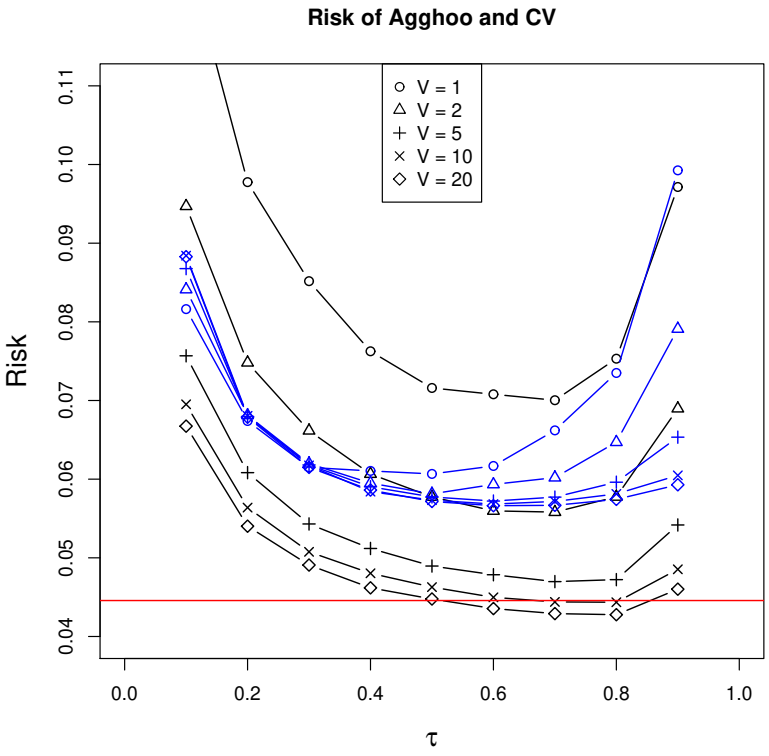

Figure 2.1: Performance of Agghoo and CV in density estimation using trigonometric series. Agghoo in black, CV in blue, oracle in red.

Figure 2.1 shows the average (squared) L^2 risks obtained over 1000 repetitions. The red line represents the average risk of the oracle. For large enough V and a range of values of τ, Agghoo performs significantly better than the oracle (at $\tau = 0.8$, the gap is about 4.5 standard deviations). The simulations I conducted during this thesis show that this happens also with the Lasso (Chapter 4).

This property is not universal, however. In Chapter 4, I show through a simulation that when there is a low-dimensional true model in sparse regression with the Lasso, Agghoo may be inferior to cross-validation. In general, what can be said about Agghoo is that it is *safe* in the sense that by equation (2.10), it always improves on the hold-out when the risk measure $P\gamma$ is convex, a property which holds for many classical statistical problems, including least-squares regression, quantile regression, least-squares density estimation, and many others. There is no equivalent guarantee for cross-validation or for hyperparameter averaging, to the best of my knowledge. For the same convexity argument to apply to hyperparameter averaging, the function $\lambda \mapsto P\gamma(\hat{s}_\lambda)$ would have to be convex, a property which depends on the collection $\hat{s}_\lambda$ and potentially also on the distribution P and the sample D_n.

2.7.3 Contributions: oracle inequalities for Agghoo

General theory Because Agghoo's performance is always better than that of the hold-out, in the convex setting, any oracle inequality satisfied by the hold-out is also satisfied by its aggregated version. In particular, the general oracle inequalities described in paragraph 2.3.2 also apply to Agghoo, which suggests that Agghoo, like the hold-out, behaves well in a wide variety of contexts. Theorem 3.7.3 of Chapter 3 states such a general result, following from the corresponding Theorem 3.7.2 for the hold-out.

From general theory to its applications Since a general theory exists, one might hope to show that Agghoo and the hold-out perform almost as well as the oracle in general. However, the general theorems do not tell us whether the margin assumption (2.6) holds in a given setting and if so, for which function w. Without knowing the function w, it is also hard to know whether the remainder term r_n in the oracle inequality is large or small. In order to work out more intuitive and explicit conditions under which a margin hypothesis holds, and in order to explicitly compute the remainder term r_n and verify that it is negligible, it is necessary to consider more particular settings in which more is known about the risk function γ and the collection of estimators. In this thesis, I focused on two such settings: kernel methods (as defined in equation (2.1)) with a Lipschitz-continuous loss function (Chapter 3), and selection of the "sparsity" of sparse estimators in regression

with the Huber loss (Chapter 4). These are two settings in which the application of Theorem 3.7.3 of Chapter 3 allows a significant improvement compared to previously known results on the hold-out and cross-validation. In the following three sections, I discuss how the results of Chapters 3 and 4 of this thesis compare to the literature in three main problems of statistical learning: classification, regression and density estimation. As argued in paragraph 2.3.2, methods based on traditional margin hypotheses have trouble dealing with unbounded losses: it is therefore in such settings that the techniques developed in this thesis may yield improvements. Thus, a major factor in this discussion will be the boundedness or unboundedness of the risk in various settings.

2.8 Hold-out and Agghoo in classification

2.8.1 Setting

In classification, the aim is typically to minimize the probability of misclassification of the label $Y \in \{1 \dots M\}$ by a classifier $t : \Xi \to \{0, \dots, M-1\}$ given input $X \in \Xi$, that is, to minimize $\mathbb{P}(t(X) \neq Y)$ on a new, independent observation (X, Y). If $t = \hat{s}_m$ is an estimator, this measures the generalization ability of the classification rule $\hat{s}_m$. This risk corresponds to the loss function $\gamma(t, (x, y)) = \mathbb{I}_{t(x) \neq y}$, which is uniformly bounded. I will focus here on the case $M = 2$ (*binary classification*), which is simpler.

2.8.2 Hold-out

For the loss function $\gamma(t, (x, y)) = \mathbb{I}_{t(x) \neq y}$ of binary classification, the general margin hypothesis (2.6) turns out to be related to similar assumptions introduced to study empirical risk minimization, and also called "margin hypotheses". The most well-known among those are Tsybakov's margin assumption [103]:

$$\forall h > 0, \qquad \mathbb{P}\left(\left|\mathbb{E}[Y|X] - \tfrac{1}{2}\right| \leqslant h\right) \leqslant Ch^{\beta} \tag{2.11}$$

for some constants C, $\beta > 0$, and Massart's margin assumption (Massart and Nédélec [77]):

$$\mathbb{P}\left(\left|\mathbb{E}[Y|X] - \tfrac{1}{2}\right| \leqslant h\right) = 0,$$

for some $h > 0$.

Under one of these two assumptions, equation (2.6) holds for a function $w : x \mapsto Cx^{\theta}$, for some constants $C > 0$ and $\theta \in (0; 1]$. Hence, Massart's Theorem [76, Corollary 8.8] applies, as $x \mapsto \frac{w(x)}{x}$ is non-increasing and γ is bounded. The optimality of the remainder term r_n given by the application of [76, Corollary

8.8] is less obvious. However, because the margin assumption plays a similar role in the study of empirical risk minimization as it does in the study of the hold-out, Blanchard and Massart [18] argue that r_n is typically much smaller than the risk of an empirical risk minimizer $\hat{s}_m$ over a finite-dimensional model m, hence smaller than the risk of the oracle in model selection. Moreover, results of Lecué [67] show that the hold-out satisfies an oracle inequality (2.3) with minimax-optimal remainder term r_n under the Tsybakov margin assumption (2.11). Thus, in classification, the hold-out is nearly optimal, at least up to the difference between estimators $\hat{s}_m$ trained on samples of size n and $n_t = \tau n < n$.

2.8.3 Contributions: aggregating the hold-out

Majhoo

Aggregated hold-out does not apply directly to classification, as averaging makes no sense over a finite set of labels. However, it is common practice to use *majority voting* between classifiers instead of averaging to perform aggregation in classification (this is used, in particular, by random forests). In Chapter 3, I consider the possibility of aggregating the hold-out by majority voting among classifiers. Together with Sylvain Arlot and Matthieu Lerasle, we show that when hold-out classifiers are aggregated using a majority vote (a variant of Agghoo that is called "Majhoo"), the excess risk $\ell(s, \cdot)$ increases at most by a constant factor, equal to the number of classes (Proposition 3.10.1). This means that Majhoo, the analogue of Agghoo, shares the good properties of the hold-out — at least in terms of rate of convergence — despite the risk not being convex. More precisely, we prove that Majhoo satisfies an oracle inequality (Theorem 3.4.5) under the Tsybakov margin assumption (2.11), with leading constant 3. The factor $C \geqslant 2$ which appears in the oracle inequality is the price to pay for the lack of convexity of the $0-1$ loss.

How Agghoo applies to classification

Majhoo is not necessarily the only way to aggregate the hold-out in classification. Because the $0-1$ loss is very difficult to optimize in practice, it is frequent to introduce a *surrogate loss* $\gamma_\phi(t,(x,y)) = \phi(yt(x))$, where ϕ is a convex function, and t is a real-valued function. Note that for the surrogate problem, the labels are $\{-1;1\}$ rather than $\{0;1\}$, and the sign of t predicts the label $y \in \{-1;1\}$. Provided the surrogate loss is *calibrated for classification*, the minimizer of the risk $\mathbb{E}[\phi(Yt(X))]$ leads to a Bayes optimal classifier, and there is a quantitative relationship between the excess risk $\ell_\phi(t)$ with respect to γ_ϕ and the excess risk with respect to the $0-1$ loss [10].

Because the functions t used in the surrogate problem are real-valued, they

can be averaged. Thus, given real-valued estimators $\hat{s}_m$, one can apply Agghoo to the $\hat{s}_m$, measuring the risk with the surrogate loss γ_ϕ and aggregating by averaging. Because the surrogate loss γ_ϕ is convex, Jensen's inequality (equation (2.10)) applies and Agghoo improves the performance of the hold-out (as measured by the surrogate loss). One could also consider using the hold-out with $0-1$ loss to select m, and aggregating the $\hat{s}_m$ by averaging. However it is better for theoretical purposes at least to consistently apply Agghoo: if the hold-out uses the $0-1$ loss, the surrogate risk of $\widehat{f}_T^{\text{ho}}$ is not controlled by the oracle inequalities described in paragraph 2.3.2, which assume that the same contrast function used by the hold-out to estimate risk is used to assess its performance theoretically.

2.8.4 Contributions: Agghoo and the hold-out with unbounded contrasts

Unbounded contrasts in classification

Assume that the loss function γ_ϕ is used for risk estimation. A difference with the original classification problem is that the contrast γ_ϕ is unbounded. A bounded contrast can be recovered by restricting the predictors t to taking values in some bounded interval $[-1;1]$, for example by *truncation*, that is, replacing any predictor t with $\tilde{t} = \max(-1, \min(t, 1))$. Moreover, when the minimizer of $\mathbb{E}[\phi(Yt(X))]$ is a function $s : \mathcal{X} \to [-1;1]$, as in the case of the hinge loss [95, Table 12.1], truncation always improves performance: $\mathbb{E}[\phi(Y\tilde{t}(X))] \leqslant \mathbb{E}[\phi(Yt(X))]$. However, such truncation may be undesirable for computational reasons, notably because it is non-linear. There are also contrasts, such as the logistic loss, for which the optimal predictor s is *not* bounded independently of the distribution P [95, Table 12.1], in which case truncation may degrade performance according to the surrogate loss (but not the $0-1$ loss). If truncation is not performed, then the contrast γ_ϕ is unbounded. Paragraph 2.3.2 argues that for unbounded contrasts, no margin hypothesis (2.6) can hold for a function w satisfying Massart's assumption that $\frac{w(x)}{x}$ is non-increasing.

SVMs

An important class of methods using surrogate losses in classification are the SVM classifiers and, more generally, kernel methods in the sense of equation (2.1) using as loss function a surrogate classification loss γ_ϕ (the classical SVMs correspond to the hinge loss $\phi(u) = (1-u)_+$). Most functions ϕ used for this purpose are Lipschitz: this includes the hinge loss, logistic loss and also the Huber loss of classification [95, Table 12.1]. For such loss functions, Chapter 3 of this thesis states and proves oracle inequalities that do not require the predictors $\hat{s}_\lambda$ to be uniformly

bounded: thus, truncation is not necessary. Instead of uniform boundedness, the results of Chapter 3 assume that the regularization parameter λ has a lower bound $\lambda_m(n)$, which is allowed to tend to 0. Chapter 3 in fact studies kernel methods in general, not just in the classification setting. Regression is its main focus, since it is in regression that unboundedness is most important. Hence, I will discuss the results of Chapter 3 in some more detail in Section 2.9.2 below.

2.9 Hold-out and Agghoo in regression

In regression, the aim is to predict a real variable $Y \in \mathbb{R}$ using *covariates* $X \in \mathcal{X}$ and predictive functions, or predictors $t : \mathcal{X} \to \mathbb{R}$. The *residual* $y - t(x)$ is the gap between the prediction $t(x)$ and the observation y. Risk is usually measured using a non-negative, convex function of the residual $\gamma(t, (x, y)) = \phi(y - t(x))$. This includes the square loss $\phi(x) = x^2$ of least-squares regression, as well as Lipschitz losses used in robust regression, such as the L^1 loss or absolute value $\phi(x) = |x|$ of least-absolute deviations and the *Huber loss* $\phi_c : x \mapsto \frac{x^2}{2}\mathbb{I}_{|x|\leqslant c} + c\left(x - \frac{c}{2}\right)$.

2.9.1 State-of-the-art: hold-out in bounded regression

Though in regression the risk is generally unbounded, there are situations in which the variable Y is known to be bounded: for example, the statistician may choose to replace an unbounded variable Z with $Y = g(Z)$ for some function $g : \mathbb{R} \to [0; 1]$. In this case, it makes sense to only consider predictors t that also range in $[0; 1]$, since any other predictor t can be improved by replacing it with $\max(0, \min(t, 1))$.

Bounded least-squares regression

In *bounded*, *least-squares* regression where it is assumed that the variable Y and the predictors t all take value in the interval $[0; 1]$, as noted by Massart [76], the margin assumption (2.6) holds with linear w, because of the inequality:

$$\mathbb{E}\left[\left((Y - t_1(X))^2 - (Y - t_2(X))^2\right)^2\right] \leqslant 2\left\|t_2 - t_1\right\|^2_{L^2(X)} \leqslant 2\left(\left\|t_1 - s\right\|_{L^2(X)} + \left\|t_2 - s\right\|_{L^2(X)}\right)^2,$$

which implies that (2.6) holds with $w(u) = \sqrt{2}u$. As a result, Massart's oracle inequality for the hold-out [76, Corollary 8.8] applies and yields a remainder term $r_n = \mathcal{O}\left(\frac{\log|\mathcal{M}|}{n_v}\right)$ in the oracle inequality (2.3), where $n_v < n$ denotes the size of the validation sample. A very similar result was proved directly by Györfi et al in 2002 [48, Chapter 7]. Gyorfi et al [48, Chapter 7] show that this oracle inequality can be used to construct adaptive estimators with respect to non-parametric classes: in this non-parametric context, the remainder term $r_n = \frac{\log|\mathcal{M}|}{n_v}$ is negligible. Thus,

the hold-out yields an optimal oracle inequality in a non-parametric setting. The oracle inequality of van der Vaart et al [105] yields a similar result for cross-validation, albeit for a non-standard form of CV where the final estimator is not retrained on the whole sample (see section 2.4 for more details). For the leave-one out, Györfi et al [48, Chapter 8] prove that CV satisfies oracle inequalities when applied to kernel and nearest-neighbour estimators. These two classes of estimators share the property of being bounded by $\max_{1\leqslant i\leqslant n}|Y_i|$, so they are indeed uniformly bounded when $Y\in L^\infty$. For both types of estimators, [48, Chapter 8] states oracle inequalities with remainder term $r_n=\mathcal{O}\left(\sqrt{\frac{\log|\mathcal{M}|}{n}}\right)$, which is larger than for the hold-out when n_v is of order n. In the case of kernel estimators, [48, Chapter 8] also proves an oracle inequality with remainder term $\frac{|\mathcal{M}|}{n}$. Compared to previous oracle inequalities, n appears in the remainder term instead of n_v and estimators are trained on samples of size $n-1$, which is an improvement, but the dependency on $|\mathcal{M}|$ is significantly worse.

Bounded regression with Lipschitz loss functions

For $L-$Lipschitz functions ϕ, the boundedness of Y is not necessary to the application of [76, Corollary 8.8], since for any predictors t_1,t_2,

$$|\phi(y-t_1(x))-\phi(y-t_2(x))|\leqslant L|t_1(x)-t_2(x)|,$$

which does not depend on y. Hence, if the predictors t are bounded (for example if they take value in $[-1;1]$), the loss function $\gamma:(t,(x,y))\mapsto\phi(y-t(x))$ can be replaced by the bounded loss function $(t,(x,y))\mapsto\phi(y-t(x))-\phi(y-t_0(x))$ for the purpose of theoretical analysis: this does not change s or $\ell(s,t)$, nor does it affect any algorithm based on empirical risk minimization, such as the hold-out. Considering now the margin assumption, the Lipschitz property of ϕ implies that for any $r\in[0;2]$

$$\operatorname{Var}\left(\phi(Y-t_1(X))-\phi(Y-t_2(X))\right)\leqslant L^2\left\|t_1-t_2\right\|_{L^2(X)}^2\leqslant L^2\left\|t_1-t_2\right\|_\infty^{2-r}\left\|t_1-t_2\right\|_{L^r}^r,$$

so that to prove a margin hypothesis (2.6), it is enough to show that

$$\ell(s,t)\geqslant\kappa\left\|t-s\right\|_{L^r}^{r+\theta},\tag{2.12}$$

for some $r,\theta,\kappa>0$, a *self-calibration inequality* in the terminology of Steinwart [97]. Steinwart [97] gives conditions on the distribution of X under which a *self-calibration inequality* (2.12) holds for the "pinball loss" of quantile regression — in particular, for the absolute value loss of median regression. Eberts and Steinwart [40] later used this analysis to derive convergence rates for the family $\hat{s}_{\nu,\lambda}$ of kernel

methods using the pinball loss and a Gaussian kernel with parameter ν. To enforce boundedness, they "truncate" the estimators $\hat{s}_{\nu,\lambda}$ at some level $M > 0$, where the "truncation" operation replaces a given function t with

$$Trunc_M(t) : x \mapsto \min(\max(t(x), -M), M). \tag{2.13}$$

They showed that the hold-out can be used to obtain an adaptive estimator with respect to their rates. As in classification, the margin assumption plays a similar role in the analysis of the (penalized) empirical risk minimizer and in that of the hold-out, which explains why the hold-out yields an adaptive estimator.

2.9.2 Unbounded regression

If the *estimators* $(\hat{s}_m)_{m\in\mathcal{M}}$ are unbounded, then the contrast $\gamma(\hat{s}_m, (x, y))$ will also be unbounded in general, for both the least-squares loss and Lipschitz losses, even if the variable $Y \in L^\infty$. One solution to recover a bounded problem is to truncate the estimators as in equation (2.13). Truncation makes sense when the target regression function s is known to be bounded by some fixed constant $M > 0$. However, this is typically not the case. If $\|s\|_\infty$ is unknown and a constant $M < \|s\|_\infty$ is chosen, truncating the estimators $\hat{s}_\lambda$ at level M will in general make them inconsistent. Moreover, truncation does not appear to be used in practice. In practice, CV is often applied to collections of estimators that are *not* uniformly bounded in general, even if Y is, such as least-squares estimators on linear models. Thus, there is reason to consider unbounded regression as a subject of theoretical research. However, the considerations of paragraph 2.3.2, as well as the inconsistency of some least-squares estimators [48, Chapter 10, Problem 10.3] suggest that oracle inequalities for the hold-out may not straightforwardly generalize to the unbounded setting. Accordingly, efforts to relax the constraints of the bounded regression setting have focused on specific collections of estimators of practical interest.

Contributions in kernel regression

The standard kernel methods (equation (2.1)) a priori lead to unbounded estimators, which is why Eberts and Steinwart [40] used truncation. However, the hold-out and cross-validation still seem to be the standard approach to selecting their parameter λ in practice [52, Section 12.3.8]. This suggests that truncation and the corresponding boundedness might be unnecessary to obtain oracle inequalities in this case.

However, if the estimators $\hat{s}_\lambda$ are *unbounded*, assuming that ϕ is Lipschitz, a margin hypothesis of the form (2.6) cannot hold with a sublinear function w, since $\mathrm{Var}\left(\phi(Y - t_1(X)) - \phi(Y - t_2(X))\right)$ is of order $\|t_1 - t_2\|_{L^2}^2$ (quadratic), whereas

$\ell(s,t_1)+\ell(s,t_2)$ is of order $\|t_1-s\|_{L^1}+\|t_2-s\|_{L^1}$ (linear). In particular, [76, Corollary 8.8] cannot be applied. The general oracle inequality proved in this thesis (Chapter 3, Theorem 3.7.3) is potentially interesting in this situation, since it allows to make use of margin conditions with super-linear functions w.

In joint work with Sylvain Arlot and Matthieu Lerasle, we applied our general Theorem 3.7.3 on Agghoo to the case of kernel estimators with a Lipschitz loss function, and proved an oracle inequality of the form (2.3) (Theorem 3.4.3 of Chapter 3). These are the first results relaxing the boundedness assumption for kernel methods, to the best of my knowledge. Our result applies to the standard kernel estimator parametrized by λ, as in equation (2.1), unlike Eberts and Steinwart [40] who considered truncated estimators $Trunc_M(\hat{s}_\lambda)$. Our hypotheses can be divided conceptually into two parts: a generalized margin hypothesis (called hypothesis $SC_{\rho,\nu}$), and extra assumptions needed to handle the unbounded case. In the case of the L^1 loss $\phi(u)=|u|$, we show that hypothesis $SC_{\rho,\nu}$ follows from a special case of the conditions considered by Steinwart [97] in his analysis of margin hypotheses for the L^1 loss. These conditions depend only on the behaviour of the conditional distribution of Y given X in the vicinity of $s(X)$, hence they are completely orthogonal to the question of boundedness of s, or of the estimators $\hat{s}_\lambda$. The extra assumptions which we need in the unbounded case, relative to the bounded one, state that there is a lower bound on the regularization parameter λ, of the form $\lambda>\lambda_m(n)$, and that the kernel K is bounded: $\|K\|_\infty<+\infty$. The remainder term r_n in our oracle inequality depends on λ_m. When the size of the validation sample n_v is of order n, that is, $n_v=(1-\tau)n$, taking λ_m as small as $\frac{1}{n}$ yields $r_n=\mathcal{O}(\frac{1}{\sqrt{n}})$, up to $\log n$ terms. This is sufficient to yield adaptation to some range of convergence rates among those obtained by [40] for the Gaussian kernel, with the caveat that we only consider selection of λ, not of the kernel bandwidth ν.

State-of-the-art on cross-validation: the Lasso

In contrast to kernel methods, the (unmodified) Lasso has been a focus of recent research on cross-validation. The Lasso denotes a penalized empirical risk minimizer over a linear model, with square loss and ℓ^1 penalty, or in other words,

$$\hat{\theta}_\lambda \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d}\left\{\frac{1}{n}\sum_{i=1}^n(Y_i-\langle\theta,X_i\rangle)^2+\lambda\sum_{j=1}^d|\theta_j|\right\}.$$

This estimator is *à priori* unbounded, since no constraint is imposed on θ. The Lasso is used in situations where it is infeasible to directly estimate the regression coefficient $\theta_*=\operatorname{argmin}_{\theta\in\mathbb{R}^d}\mathbb{E}[(Y-\langle\theta,X\rangle)^2]$ by linear least-squares, typically because the dimension d is too big (high-dimensional setting). Interest in the Lasso comes from the fact that if θ_* is *sparse*, i.e if it has only $s_*\ll n$ non-zero entries,

then the risk of the Lasso may be much less than that of the least-squares estimator over $\mathbb{R}^d$ — provided λ is well chosen. It can be as small as $\frac{s_* \log d}{n}$ under some conditions on X [15], which is the minimax rate of convergence under the sparsity assumption $\|\theta_*\|_0 \leqslant s_*$ [90]. Attaining these rates requires choosing λ as a function of the noise variance [15], which is unknown. In practice, cross-validation is typically used in order to choose λ as well as possible. For example, the R implementations of the Lasso (glmnet, lars) all propose cross-validation as a subroutine to automatically choose λ. Theoretical work on the cross-validated Lasso has focused on proving that it adapts to the theoretical risk estimates, i.e that the same bounds apply to the Lasso as to the theoretically chosen λ, based on knowledge of the noise variance. In 2013, Homrighausen and MacDonald [53] proved that the cross-validated Lasso converges at the fast rate $\frac{s_* \log d}{n}$ up to log terms, albeit under some rather restrictive assumptions on λ ($\lambda > c\sqrt{\frac{\log d}{n}}$ for some constant c) and on the distribution: truth of the model (i.e $\mathbb{E}[Y|X] = \langle \theta_*, X \rangle$), Gaussian noise, and a diagonal-dominant covariance matrix $\mathbb{E}[XX^T]$. In 2015, Chetverikhov, Liao and Chernozhukov [33] proved a similar result under less restrictive assumptions. They required in particular a lower bound on the eigenvalues of submatrices of the variance-covariance matrix $\mathbb{E}[XX^T]$. These results have the common property that they require very little in the way of boundedness assumptions on the predictors $x \mapsto \langle \hat{\theta}_\lambda, x \rangle$: in the first case, [53] only requires coordinatewise boundedness of X whereas [33] only requires boundedness of $\|X\|_{\ell^\infty} = \max_{1 \leqslant j \leqslant d} |X_j|$ in some L^q space, not in L^∞. On the other hand, they show adaptation only to a fixed theoretical rate of convergence, whereas an oracle inequality implies adaptation to whatever rate the optimal Lasso estimator happens to converge at in any given situation.

State-of-the-art on cross-validation for selecting among linear models

Linear model selection in regression deals with collections of estimators composed of *empirical risk minimizers* $\hat{s}_m$ over vector spaces of functions $m \in \mathcal{M}$ called *models*. Unlike kernel and nearest-neigbour regressors, discussed earlier, least-squares estimators over linear models may be unbounded even when the data Y is bounded [48, Chapter 10]. Therefore, even the case of bounded Y data is a non-trivial extension of the bounded regression setting of paragraph 2.9.1.

In the least-squares setting, Navarro and Saumard [83] proved oracle inequalities for model selection of a certain kind of model, which possesses a "strongly localized basis". Their results apply to the standard form of V-fold cross-validation, as well as to the bias-corrected "V-fold penalization" method introduced by Arlot [2]. The oracle inequalities hold with high probability, instead of in expectation, and have a negligible remainder term $r_n = \mathcal{O}\left(\frac{\log^3 n}{n}\right)$. Like [105] and results for

the hold-out, their oracle involves estimators trained on a sample of size $n_t = \frac{V-1}{V}n$ instead of the whole sample. They assume uniform boundedness of Y and of the "true" projections $\hat{s}_m$ of s on the model m in $L^2(P)$, but not uniform boundedness of the estimators $\hat{s}_m$ themselve. As a result, they can use the standard least-squares estimators $\hat{s}_m$ and do not have to modify them to make them bounded.

In the case of the hold-out, boundedness assumptions were further relaxed by Wegkamp [114] for selection of general estimators in least-squares regression, with a focus on model selection. He considers the additive regression model, in which $Y_i = s(X_i) + \varepsilon_i$, for i.i.d noise variables ε_i, independent from X_i. Wegkamp makes weak moment assumptions on the noise $\|\varepsilon_i\|_{L^p} \leqslant \tau_p$ (for $p > 2$), and obtains a remainder term r_n which depends polynomially on $|\mathcal{M}|$, instead of the $\log|\mathcal{M}|$ dependency in results which make stronger, exponential moment assumptions on ε. Wegkamp assumes that $\|s - \hat{s}_m\|_\infty \leqslant B$ but notes that his results hold under the weaker assumption that

$$\int (\hat{s}_m - s)^4 \, dP_X \leqslant R \int (\hat{s}_m - s)^2 dP_X, \tag{2.14}$$

where P_X denotes the distribution of X. This assumption is closely related to *margin hypotheses*, by the following argument. For any point (x, y) and any $(m, m') \in \mathcal{M}$,

$$\big((y - \hat{s}_m(x))^2 - (y - \hat{s}_{m'}(x))^2\big)^2 = (\hat{s}_m(x) - \hat{s}_{m'}(x))^2 (2y - \hat{s}_m(x) - \hat{s}_{m'}(x))^2.$$

Let $(X, Y) \sim P$ be a new observation independent from the estimators $\hat{s}_m$. Suppose that $Y = s(X) + \varepsilon$ for a centred variable ε independent from X, then assumption (2.15) implies that, for some constant κ,

$$\begin{aligned} P\left(\gamma(\hat{s}_m) - \gamma(\hat{s}_{m'})\right)^2 &\leqslant 4\mathbb{E}[\varepsilon^2] \, \|t_1 - t_2\|^2_{L^2(X)} + P\left[(\hat{s}_m - \hat{s}_{m'})^2 (2s - \hat{s}_m - \hat{s}_{m'})^2\right] \\ &\leqslant 4\mathbb{E}[\varepsilon^2] \, \|t_1 - t_2\|^2_{L^2(X)} + \kappa \left(\|\hat{s}_m - s\|^4_{L^4(X)} + \|\hat{s}_{m'} - s\|^4_{L^4} \right) \\ &\leqslant \left(8\mathbb{E}[\varepsilon^2] + \kappa R\right) \left(\|t_1 - s\|^2_{L^2(X)} + \|t_1 - s\|^2_{L^2(X)} \right), \end{aligned}$$

thus equation (2.6) holds with $w(u) = \sqrt{8\mathbb{E}[\varepsilon^2] + \kappa R}u$. Equation (2.14) suffers from two problems. First, it depends on the unknown regression function s, which makes the assumption hard to verify. Secondly, it is *inhomogeneous*: the left-hand side grows like a fourth power of $\hat{s}_m - s$, while the right-hand side grows quadratically. This means that equation (2.14) cannot hold with $\hat{s}_m$ an arbitrary large element of a vector space m: like other margin assumptions used in the literature, it implicitly imposes a boundedness constraint, albeit around s rather than 0. A homogeneous version of (2.14) would take the form

$$\|\hat{s}_m - s\|_{L^4(X)} \leqslant \kappa \, \|\hat{s}_m - s\|_{L^2} \tag{2.15}$$

for some constant κ.

Such an $L^4 - L^2$ norm inequality was used by Audibert and Catoni [7] to prove risk bounds for a robust risk-minimization procedure. Other norm inequalities have been used to study risk-minimization methods in the unbounded setting. Such inequalities have the general form

$$\forall t \in m, \|t\|_{L^q(X)} \leqslant \kappa \|t\|_{L^p(X)}, \tag{2.16}$$

where $q > p$, m is the model on which the risk is to be minimized, and κ is treated as a constant (at least in the examples below). $L^2 - L^1$ ($q = 2$, $p = 1$) inequalities appear in the work of Lecué and Lerasle [68] on median of means estimators. For the empirical risk minimizer (i.e least-squares), Mendelson [81] used the "small ball assumption", which is equivalent to an $L^2 - L^1$ norm inequality [68], while Audibert and Catoni [7] used an $L^\infty - L^2$ inequality ($q = +\infty$, $p = 2$).

This raises the question of whether similar hypotheses might yield oracle inequalities for the hold-out, which is also an empirical risk minimizer (on the collection of functions $\hat{s}_m(D_n^T)_{m \in \mathcal{M}}$).

Contributions: Agghoo and the hold-out applied to sparse linear predictors in robust regression

In chapter 4, I investigate the application of the hold-out and Agghoo to collections of linear regression estimators $x \mapsto \hat{q}_k + \langle \hat{\theta}_k, x \rangle$ which are *sparse* in the sense that

$$\left\| \hat{\theta}_k \right\|_0 = \left| \left\{ j : \hat{\theta}_{k,j} \neq 0 \right\} \right| \leqslant k, \tag{2.17}$$

i.e the coefficients $\hat{\theta}_k$ have less than k non-zero components. The risk of the estimators is assessed using the Huber loss, a classic Lipschitz loss function used for robust regression [55].

The integer k parametrizes the *complexity* of the estimators $\hat{\theta}_k$. Estimators satisfying equation (2.17) arise naturally from sparse regression methods such as best-subset, forward stepwise [52, Section 3.3] and LARS [42] (which add variables one by one). They can also be extracted from the regularization path of the Lasso, as suggested by Zou et al [120]. Compared to linear model selection, the estimators are allowed to range over the whole, high-dimensional space $\mathbb{R}^d$ (instead of a subspace m); compared with the standard Lasso, the difference is that the $\hat{\theta}_k$ are assumed to have a fixed degree of sparsity (less than k non-zero coefficients).

In this setting, I prove that the hold-out satisfies an oracle inequality of type (2.3), with a leading constant $C > 1$ and a remainder term $r_n = \mathcal{O}\left(\frac{\log n}{n_v}\right)$, where n_v denotes the size of the validation sample. If n_v is of order n, this remainder term is smaller than the convergence rates of sparse regression, discussed in the

section 2.9.2. This implies that the hold-out is adaptive with respect to these rates. Only a weak, non-uniform boundedness assumption is made: the estimators $\hat{\theta}_k$ and the covariates X are assumed to be bounded by some polynomial in the sample size n. Similarly to the RKHS case (section 2.9.2), a distributional assumption is also needed, which depends only on the conditional distribution of the noise $Y - s(X)$ given X, and not on the estimators $\hat{s}_m$. Finally, the unboundedness of the estimators is dealt with through a norm inequality of the form

$$\forall \theta, \|\theta\|_0 \leqslant 2K \implies \|\langle \theta, X\rangle\|_\infty \leqslant \kappa(n, n_v) \|\langle \theta, X\rangle\|_{L^2}, \tag{2.18}$$

for some constant $\kappa(n, n_v)$ such that $\kappa(n, n_v) = \mathcal{O}\left(\sqrt{\frac{n_v}{\log n}}\right)$ as $n_v \to +\infty$. Compared to the norm hypothesis made by [114] in model selection, this inequality does not depend on the unknown regression function s and it is *homogeneous*, meaning that it may hold over entire vector spaces (which is what is required here). Compared to the homogeneous norm-inequalities (2.16) discussed in section 2.9.2, equation (2.18) corresponds to the case $q = +\infty$, $p = 2$, like the hypothesis made in Audibert and Catoni [7]. An important difference with the norm inequalities of section 2.9.2 is that equation (2.18) allows the constant $\kappa = \kappa(n)$ to grow with n at rate $\sqrt{\frac{n}{\log n}}$ (assuming that n_v is of order n), instead of treating κ like a constant. In Chapter 4, I give two examples where equation (2.18) holds under reasonable conditions.

Conclusion: contributions to the theory of the hold-out in regression The theoretical contributions of Chapters 3 and 4 of this thesis improve previously known oracle inequalities for the hold-out in regression by relaxing boundedness assumptions on the estimators. In the RKHS case, this does away with truncation of the estimators, so that the new oracle inequality holds for the standard RKHS method (as used in practice), rather than a modified version. For sparse estimators, in comparison to previously known oracle inequalities, my results combine weaker assumptions on the covariate X and weaker bounds on the estimators $\hat{\theta}_k$ for a similar conclusion: an oracle inequality of type (2.3) with a negligible remainder term. This provides better theoretical justification for the use of the hold-out and of Agghoo in regression.

2.10 Hold-out, cross-validation and Agghoo in L^2 density estimation

Given a sample $Z_1, \ldots, Z_n$ drawn from an unknown measure P, L^2-density estimation is concerned with estimation of the density s of P with respect to a reference

measure μ, where the quality of an estimate t is measured by $\|t-s\|^2_{L^2(\mu)}$. This can be formulated as a risk-minimization problem, since for $\gamma(t,z) = \|t\|^2_{L^2(\mu)} - 2t(z)$ and $Z \sim P$,

$$\|t-s\|^2_{L^2(\mu)} = E_Z[\gamma(t,Z)] - E_Z[\gamma(s,Z)] = \ell(s,t),$$

2.10.1 State of the art: cross-validation in least-squares density estimation

The least-squares setting has the particularity that the function t need not be bounded for a margin hypothesis to hold. If it is assumed that the underlying density s is uniformly bounded ($\|s\|_\infty < +\infty$), then for all functions (t_1, t_2),

$$\begin{aligned}\operatorname{Var}\left(t_1(Z) - t_2(Z)\right) &\leqslant \int (t_1(z) - t_2(z))^2 s(z) d\mu(z) \\ &\leqslant \|s\|_\infty \|t_1 - t_2\|^2_{L^2(\mu)} \\ &\leqslant \|s\|_\infty \left(\|t_1 - s\|_{L^2(\mu)} + \|t_2 - s\|_{L^2(\mu)}\right)^2.\end{aligned}$$

This is a margin hypothesis of the form (2.6) with $w(u) = \sqrt{\|s\|_\infty}u$, and it holds without any assumption on the functions t_1, t_2. However, the contrast $\gamma(t,z) = \|t\|^2_{L^2(\mu)} - 2t(z)$ remains unbounded whenever t is unbounded. Nevertheless, for important classes of (unbounded) estimators, cross-validation has been proved to satisfy oracle inequalities. Dalelane [35] showed that leave-one-out cross-validation satisfies an oracle inequality when used to select the bandwidth of a kernel density estimator. For selection of empirical risk minimizers on linear models, Celisse [30] proved that leave-p-out cross-validation satisfies an oracle inequality. Arlot and Lerasle [5] proved similar results for other cross-validation procedures, including V-fold and Monte-Carlo cross-validation. Their oracle inequalities are of type (2.2), with an oracle trained on the whole sample and a remainder term $r_n = \mathcal{O}\left(\frac{\log n}{n}\right)$. Both make an assumption similar to equation (2.18), namely that for all models m with orthogonal basis $(\varphi_j)_{j\in\Lambda(m)}$,

$$\sup_{t\in m : \|t\|_{L^2} \leqslant 1} \|t\|_\infty = \left\| \sum_{j\in\Lambda(m)} \varphi_j^2 \right\|_\infty \leqslant \sqrt{n}.$$

2.10.2 Contributions: applications of Theorem 3.7.2 to least-squares density estimation

Chapters 5 and 6 study hold-out and Agghoo applied to least-squares density estimation in $L^2([0;1])$ using the empirical Fourier series estimators

$$\hat{s}_k(D_n) = 1 + \sum_{j=1}^{k} P_n(\psi_j)\psi_j,$$

where $\psi_j : x \mapsto \sqrt{2}\cos(2\pi jx)$. The methods developed in Chapter 3 of this thesis (Theorem 3.7.2) also apply in this setting. In chapter 6, Theorem 3.7.2 is used to derive a precise oracle inequality for the hold-out (Theorem 6.6.5) which plays an important role in the proof of the other results of that chapter. This result cannot be derived from the more general results of section 2.10.1. That it nonetheless can be proved using Theorem 3.7.2 illustrates the flexibility of the general result.

2.11 Contributions on the hold-out and Agghoo: beyond oracle inequalities

Oracle inequalities show that Agghoo and the hold-out are good model selection procedures *in theory*, in that they perform almost as well as the best estimator in the given collection. However, other procedures, such as cross-validation, satisfy similar bounds, so these results cannot tell us which of the two methods performs better in a given situation. Moreover, the results of my simulations (Figure 2.1, sections 3.5 and 4.4.3), in which Agghoo sometimes performs better than the model selection oracle, cannot be explained by an oracle inequality (2.3) with constant $C > 1$ and remainder term $r_n > 0$. The interest in Agghoo is driven in large part because of such simulation results, where Agghoo clearly outperforms model selection methods such as cross-validation. For practical as well as for theoretical reasons, it would be interesting to know when such situations occur, and for which values of Agghoo's parameter. This would allow practitioners to make the right choice between Agghoo and its alternatives, and to correctly calibrate Agghoo's parameters. From a more theoretical perspective, an answer to such questions would shed light on the behaviour of the hold-out and of hyperparameter aggregation.

In Chapters 5 and 6 of this thesis, I develop a very precise analysis of the hold-out and of Agghoo in order to answer such questions in a specific setting, namely least-squares density estimation using empirical Fourier series. More precisely, I study L^2 density estimation 2.10 of a symmetric density function $s \in L^2([0;1])$

using the estimators

$$\hat{s}_k(D_n) = 1 + \sum_{j=1}^{k} P_n(\varphi_j)\varphi_j,$$

where $\varphi_j = \sqrt{2}\cos(2\pi j \cdot)$ is the cosine basis. The choice of this setting is motivated by a trade-off between its interest to theoreticians and practitioners and the technical difficulty of its study. The Fourier series estimator is certainly usable in practice, and from a theoretical viewpoint, it adapts to the whole scale of Hölder/Sobolev spaces on the torus. On the other hand, the algebraic structure of the trigonometric polynomials and the simple formula available for the estimators makes theoretical analysis easier, which allows for more accurate results. Moreover, Figure 2.1 suggests that Agghoo can indeed outperform the oracle in this setting.

To facilitate theoretical analysis, I make some assumptions on the Fourier coefficients θ_j of s on the cosine basis. The first hypothesis is that θ_j^2 should be a non-increasing sequence: this guarantees that the (expected) risk $\mathbb{E}[\|\hat{s}_k - s\|^2]$ is a convex function of the parameter k, and that it has a unique minimum k_*. The other assumptions state roughly that the sequence θ_j^2 decreases polynomially and that it does not "jump" — there are no sudden drops where it suddenly decreases very fast. Chapters 5 and 6 share these assumptions.

To understand the behaviour of the hold-out and of Agghoo, the first step is to analyze the hold-out estimator of the risk, which the hold-out minimizes. Chapter 5 is dedicated to constructing a distributional approximation of this empirical risk estimator in the vicinity of the "true" optimum,

$$k_*(n_t) = \operatorname*{argmin}_k \mathbb{E}\left[\|\hat{s}_k(D_{n_t}) - s\|^2\right].$$

The usual tools of asymptotic analysis turn out not to be adapted here, because the relevant process (suitably centered and scaled) does not converge to a limit, and what is needed instead is an approximation for each sample size n. I construct the approximating sequence using the Komlos-Major-Tusnady theorem of strong approximation [62, 63] to obtain a Gaussian process, which I then approximate by another Gaussian process with a similar variance-covariance kernel. The correct approximation turns out to be the sum of a convex function f_n and a two-sided Brownian motion changed in time W_{g_n}. Since there does not seem to be a simple formula for f_n and g_n, I instead prove some lower bounds and upper bounds on their increments, which are useful for Chapter 6.

In Chapter 6, I carry out a precise analysis of the risk of Agghoo and the hold-out. First, using Theorem 3.7.2 of Chapter 3, I prove a preliminary oracle inequality for the hold-out, which shows that the parameter $\hat{k}$ selected by the hold-out is located in a region of the parameter space $\mathbb{N}$ where the hold-out risk

estimator is well-approximated by the process constructed in Chapter 5. Secondly, using techniques developed by Leandro R. Pimentel [88], I show that the approximation of the hold-out risk estimator constructed in Chapter 5 leads to a (distributional) approximation $\hat{k}^\infty$ of its argmin, the parameter $\hat{k}$ selected by the hold-out. The risks of Agghoo and of the hold-out can be expressed using the distribution function of $\hat{k}^\infty$. Using this, I prove a first-order approximation for the risk of the hold-out,

$$\mathbb{E}\left[\|\hat{s}_{\hat{k}}(D_{n_t}) - s\|^2\right] = or(n_t) + r_n + o(r_n),$$

where $or(n_t)$ denotes the risk of the oracle trained on a sample of size $n_t = n - n_v$ (Theorem 6.4.3). I also show that Agghoo's risk satisfies an inequality of the form

$$\mathbb{E}\left[\left\|\widehat{f}^{\mathrm{ag}}_{\mathcal{T}} - s\right\|^2\right] \leqslant or(n_t) + r_n - d_n + o(d_n).$$

The risk of Agghoo is bounded by two terms: the risk of the hold-out, $or(n_t) + r_n$, and a *negative* term d_n resulting from aggregation. In general, d_n is always bigger than cr_n, for some constant c independent of n, so Agghoo at least reduces the remainder term r_n by a constant factor. However, more is true if more assumptions are made on s. In a last part (Section 6.4.3), I show that if θ_j^2 decreases at a fixed polynomial rate, $\theta_j^2 \sim cj^{-\alpha}$, then Agghoo's risk can be smaller than the oracle, by as much as a constant factor. I give bounds on the values of Agghoo's parameter $\tau_n = \frac{n_t}{n}$ for which this occurs.

Chapter 3

Aggregated Hold-out

Keywords: cross-validation, aggregation, bagging, hyperparameter selection, regularized kernel regression

3.1 Introduction

The problem of choosing from data among a family of learning rules is central to machine learning. There is typically a variety of rules which can be applied to a given problem —for instance, support vector machines, neural networks or random forests. Moreover, most machine learning rules depend on hyperparameters which have a strong impact on the final performance of the algorithm. For instance, k-nearest-neighbors rules [11] depend on the number k of neighbors. A second example, among many others, is given by regularized empirical risk minimization rules, such as support vector machines [96] or the Lasso [99, 24], which all depend on some regularization parameter. A related problem is model selection [27, 76], where one has to choose among a family of candidate models.

In supervised learning, cross-validation (CV) is a general, efficient and classical answer to the problem of selecting a learning rule [3]. It relies on the idea of splitting data into a training sample —used for training a predictor with each rule in competition— and a validation sample —used for assessing the performance of each predictor. This leads to an estimator of the risk —the hold-out estimator when data are split once, the CV estimator when an average is taken over several data splits—, which can be minimized for selecting among a family of competing rules.

A completely different strategy, called aggregation, is to *combine* the predictors obtained with all candidates [84, 118, 103]. Aggregation is the key step of ensemble methods [39], among which we can mention bagging [22], AdaBoost [43] and random forests [23, 12]. A major interest of aggregation is that it builds a

learning rule that may not belong to the family of rules in competition. Therefore, it sometimes has a smaller risk than the best of all rules [93, Table 1]. In contrast, cross-validation, which selects only one candidate, cannot outperform the best rule in the family.

Aggregated hold-out (Agghoo) This paper studies a procedure mixing cross-validation and aggregation ideas, that we call *aggregated hold-out* (Agghoo). Data are split several times; for each split, the hold-out selects one predictor; then, the predictors obtained with the different splits are aggregated. A formal definition is provided in Section 3.3. This procedure is as general as cross-validation and it has roughly the same computational cost (see Section 3.3.3). Agghoo is already popular among practicioners, and has appeared in the neuro-imaging literature [54, 107] under the name "CV + averaging". Yet, to the best of our knowledge, existing experimental studies do not give any indication on how to choose Agghoo's parameters. No general mathematical definition has been provided, so it is unclear how to generalize Agghoo beyond a given article's setting. Theoretical guarantees on Agghoo have not been established yet, to the best of our knowledge. The closest results we found study other procedures, called ACV [59], EKCV [58], or "bagged cross-validation" [51], and they do not prove oracle inequalities. We explain in Section 3.3.2 why Agghoo should be preferred to these procedures in the general prediction setting.

Because of the aggregation step, Agghoo is an ensemble method, and like bagging, it combines resampling with aggregation. The application of bagging to the hold-out was first suggested by Breiman [22] as a way to combine pruning and bagging of CART trees. The combination of bagging and cross-validation has been studied numerically by [87]. A major difference with Agghoo is that the training and validation samples are not independent with bagging, which uses sampling *with replacement*. If the bootstrap is replaced by subsampling, bagging becomes subagging [25], and its combination with cross-validation yields a procedure much closer to Agghoo, but still different, see Section 3.3.2. Overall, previous results on bagging or subagging do not apply to Agghoo; new developments are required.

Contributions In this article, Agghoo's performance is studied both theoretically and experimentally. We consider Agghoo from a prediction point of view. Performance is measured by a risk functional. On the theoretical side, the aim is to show that the risk of Agghoo's final predictor is as low as the risk of the optimal rule among the given collection. This is known as an oracle inequality. By a convexity argument, Agghoo always improves on the hold-out, provided that the risk is convex. Hence, Agghoo can safely replace the hold-out in any application where this hypothesis holds true. Another consequence is that oracle inequalities

for Agghoo can be deduced from oracle inequalities for the hold-out.

This kind of result on the hold-out has already appeared in the literature: for example, Massart [76, Corollary 8.8] proves a general theorem under an abstract noise assumption; more explicit results have been obtained in specific settings such as least-squares regression [48, Theorem 7.1] or maximum-likelihood density estimation [76, Theorem 8.9]. A review on cross-validation —which includes the hold-out— can be found in [3].

Most existing theoretical guarantees on the hold-out have a limitation: they assume that the loss function is uniformly bounded. In regression, the variable Y and the regressors are also usually assumed to be bounded, which excludes some standard least-squares estimators. Even when the boundedness assumption holds true, constants arising from general bounds may be of the wrong order of magnitude, leading to vacuous results. By replacing uniform supremum bounds by local ones, we are able to relax these hypotheses in a general setting (Theorem 3.7.3). This enables us to prove an oracle inequality for the hold-out and Agghoo in regularized kernel regression with a general Lipschitz loss (Theorem 3.4.3). This oracle inequality allows for instance to recover state-of-the-art convergence rates in median regression without knowing the regularity of the regression function (adaptivity), both in the general case and, for small enough regularity, also in the specific setting of [40]. To illustrate the implications of Theorem 3.4.3, we also apply it to ε-regression (Corollary 3.4.4). To the best of our knowledge, all these oracle inequalities are new, even for the hold-out.

A limitation of Agghoo is that it does not cover settings where averaging does not make sense, such as classification. In classification with the 0–1 loss, the natural way to aggregate classifiers is to take a majority vote among them. This yields a procedure which we call Majhoo. Using existing theory for the hold-out in classification, we prove that Majhoo satisfies a general, margin-adaptive oracle inequality (Theorem 3.4.5) under Tsybakov's margin assumption [75].

All our oracle inequalities are valid for any number of training and test subsets, provided that they have the same size and that the splits are made independently of the data. Qualitatively, since bagging and subagging are well-known for their stabilizing effects [22, 25], we can expect Agghoo to behave similarly. In particular, aggregating over a large number of splits should improve much the prediction performance of CV when the hold-out selected predictor is unstable.

For further insights into Agghoo and Majhoo, we conduct in Section 3.5 a numerical study on simulated datasets. Its results confirm our intuition: in all settings considered, Agghoo and Majhoo actually perform much better than the hold-out, and even better than CV, provided their parameters are well-chosen. When choosing the number of neighbors for k-nearest neighbors, the prediction performance of Majhoo is much better than the one of CV, which illustrates the

strong interest of using Agghoo/Majhoo when learning rules are "unstable". In support vector regression, Agghoo can sometimes perform better than the oracle, while matching its performance on average. This improvement is made possible by aggregation. Based upon our experiments, we also give in Section 3.5 some guidelines for choosing Agghoo's parameters: the training set size and the number of data splits.

The remaining of the article is structured as follows. In Section 2, we introduce the general statistical setting. In Section 3, we give a formal definition of Agghoo. In Section 4, we state the main theoretical results. In Section 5, we present our numerical experiments and discuss the results. Finally, in Section 6, we draw some qualitative conclusions about Agghoo. The proofs are postponed to the Appendix.

3.2 Setting and Definitions

We consider a general statistical learning setting, following the book by Massart [76].

3.2.1 Risk minimization

The goal is to minimize over a set $\mathbb{S}$ a risk functional $\mathcal{L} : \mathbb{S} \to \mathbb{R} \cup \{+\infty\}$. The set $\mathbb{S}$ may be infinite dimensional for non-parametric problems. Assume that $\mathcal{L}$ attains its minimum over $\mathbb{S}$ at a point s, called a Bayes element. Then the *excess risk* of any $t \in \mathbb{S}$ is the nonnegative quantity

$$\ell(s,t) = \mathcal{L}(t) - \mathcal{L}(s) \ .$$

Suppose that the risk can be written as an expectation over an unknown probability distribution:

$$\mathcal{L}(t) = \mathbb{E}\big[\gamma(t,\xi)\big] \ ,$$

for a *contrast function* $\gamma : \mathbb{S} \times \Xi \to \mathbb{R}$ and a random variable ξ with values in some set Ξ and unknown distribution P, such that

$$\forall t \in \mathbb{S}, \qquad \widetilde{\xi} \in \Xi \mapsto \gamma(t,\widetilde{\xi}) \text{ is } P\text{-measurable} \ .$$

The statistical learning problem is to use data $D_n = \{\xi_1, ..., \xi_n\}$, where $\xi_1, ..., \xi_n$ are independent and identically distributed (i.i.d.), with common distribution P, to find an approximate minimizer for $\mathcal{L}$. The quality of this approximation is measured by the excess risk.

3.2.2 Examples

Supervised learning aims at predicting a quantity of interest $Y \in \mathcal{Y}$ using explanatory variables $X \in \mathcal{X}$. The statistician observes pairs $(X_1, Y_1), \dots (X_n, Y_n)$, so that $\Xi = \mathcal{X} \times \mathcal{Y}$, and seeks a predictor in $\mathbb{S} = \{t : \mathcal{X} \to \mathcal{Y} : \text{t measurable}\}$. The contrast function is defined by $\gamma(t, (x, y)) = g(t(x), y)$ for some *loss function* $g : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Here, $g(y', y)$ measures the loss incurred by predicting y' instead of the observed value y. Two classical supervised learning problems are classification and regression, which we detail below.

Example 3.2.1 (Classification) *In classification Y belongs to a finite set of labels $\mathcal{Y} = \{0, \dots, M\}$. We wish to correctly label any new data point X, and the risk is the probability of error*:

$$\forall t \in \mathbb{S}, \qquad \mathcal{L}(t) = \mathbb{P}\big(t(X) \neq Y\big) \ ,$$

which corresponds to the loss function $g(y', y) = \mathbb{I}\{y' \neq y\}$. Classification with convex losses (such as the hinge loss or logistic loss) can also be described using the formalism of Section 3.2.1.

Example 3.2.2 (Regression) *In regression we wish to predict a continuous variable $Y \in \mathcal{Y} = \mathbb{R}^d$. The error made by predicting y' instead of y is measured by the loss function defined by $g(y', y) = \phi(\|y' - y\|)$ where $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is nondecreasing and convex. Some typical choices are $\phi(x) = x^2$ (least squares), $\phi(x) = x$ (median regression) or $\phi(x) = (|x| - \varepsilon)_+$ (Vapnik's ε-insensitive loss, leading to ε-regression). The risk is given by*

$$\mathcal{L}(t) = \mathbb{E}\Big[\phi\big(\|Y - t(X)\|\big)\Big] \ .$$

If ϕ is strictly convex, the minimizer of $\mathcal{L}$ over $\mathbb{S}$ is a unique function, up to modification on a set of probability 0 under the distribution of X.

In some applications, such as robust regression, it is of interest to define s and $\ell(s, t)$ even when $\phi(\|Y\|) \notin L^1$. This is possible for Lipschitz contrasts, by the following remark.

Remark 3.2.1 *When ϕ is convex and increasing (as in Example 3.2.2), and also Lipschitz-continuous, it is always possible to define*

$$s : x \mapsto \operatorname*{argmin}_{u \in \mathbb{R}} \mathbb{E}\big[\phi(\|Y - u\|) - \phi(\|Y\|) \,\big|\, X = x\big] \ .$$

When $s \in L^1(X)$, it is a Bayes element for the loss function $g(y', y) = \phi(\|y' - y\|) - \phi(\|y\|)$. Whenever $\phi(\|Y\|) \in L^1$, this loss yields the same Bayes element and excess risk as in Example 2.2.

This small adjustment to the general definition allows to consider Example 3.2.2 when $\phi(\|Y - s(X)\|)$ is not integrable, for example when $Y = s(X) + \eta$, where η is independent from X and follows a multivariate Cauchy distribution with location parameter 0.

Some density estimation problems, such as maximum likelihood or least-squares density estimation, also fit the formalism of Section 3.2.1, see [76].

3.2.3 Learning rules and estimator ensembles

Statistical procedures use data to compute an element of $\mathbb{S}$ which approximately minimizes $\mathcal{L}$. Since Agghoo uses subsampling, we require learning rules to accept as input datasets of any size. Therefore, we define a learning rule to be a function which maps any dataset to an element of $\mathbb{S}$.

Definition 3.2.1 *A* dataset D_n *of length* n *is a finite i.i.d sequence* $(\xi_i)_{1 \leqslant i \leqslant n}$ *of* Ξ*-valued random variables with common distribution* P*.*

A learning rule $\mathcal{A}$ *is a measurable function*[1]

$$\mathcal{A} : \bigcup_{n=1}^{\infty} \Xi^n \to \mathbb{S} \ .$$

In the risk minimization setting, $\mathcal{A}$ should be chosen so as to minimize $\mathcal{L}(\mathcal{A}(D_n))$.

A generic situation is when a family $(\mathcal{A}_m)_{m \in \mathcal{M}}$ of learning rules is given, so that we have to select one of them (estimator selection), or to combine their outputs (estimator aggregation). For instance, when $\mathcal{X}$ is a metric space, we can consider the family $(\mathcal{A}_k^{\mathrm{NN}})_{k \geqslant 1}$ of nearest-neighbors classifiers —where k is the number of neighbors—, or, for a given kernel on $\mathcal{X}$, the family $(\mathcal{A}_\lambda^{\mathrm{SVM}})_{\lambda \in [0, +\infty)}$ of support vector machine classifiers —where λ is the regularization parameter. Not all rules in such families perform well on a given dataset. Bad rules should be avoided when selecting the hyperparameter, or be given small weights if the outputs are combined in a weighted average. This requires a data-adaptive procedure, as the right choice of rule in general depends on the unknown distribution P.

Aggregation and parameter selection methods aim to resolve this problem, as described in the next section.

---

[1] For any n,

$$\begin{cases} \Xi^n \times \Xi & \to \mathbb{R} \\ (\xi_{1:n}, \xi) & \mapsto \gamma(\mathcal{A}(\xi_{1:n}), \xi) \end{cases}$$

is assumed to be measurable (with respect to the product σ-algebra on Ξ^{n+1}).

3.3 Cross-Validation and Aggregated Hold-Out (Agghoo)

This section recalls the definition of cross-validation for estimator selection, and introduces a new procedure called aggregated hold-out (Agghoo). For more details and references on cross-validation, we refer the reader to the survey by Arlot and Celisse [3].

3.3.1 Background: cross-validation

Cross-validation uses subsampling and the empirical risk. We introduce first some notation.

Definition 3.3.1 (Empirical risk) *For any dataset $D_n = (\xi_i)_{1 \leqslant i \leqslant n}$ and any $t \in \mathbb{S}$, the empirical risk of t over D_n is defined by*

$$P_n \gamma(t, \cdot) = \frac{1}{n} \sum_{i=1}^{n} \gamma(t, \xi_i) \ .$$

For any nonempty subset $T \subset \{1, \dots, n\}$, let also

$$D_n^T = (\xi_i)_{i \in T}$$

be the subsample of D_n indexed by T, and define the associated empirical risk by

$$\forall t \in \mathbb{S}, \qquad P_n^T \gamma(t, \cdot) = \frac{1}{|T|} \sum_{i \in T} \gamma(t, \xi_i) \ .$$

The most classical estimator selection procedure is to *hold out* some data to calculate the empirical risk of each estimator, and then select the estimator with the lowest empirical risk. This ensures that the data used to evaluate the risk are independent from the training data used to compute the learning rules.

Definition 3.3.2 (Hold-out) *For any dataset D_n and any subset $T \subset \{1, \dots, n\}$, the associated hold-out risk estimator of a learning rule $\mathcal{A}$ is defined by*

$$HO_T(\mathcal{A}, D_n) = P_n^{T^c} \gamma\left(\mathcal{A}(D_n^T), \cdot\right) \ .$$

Given a collection of learning rules $(\mathcal{A}_m)_{m \in \mathcal{M}}$, the hold-out procedure selects

$$\widehat{m}_T^{ho}(D_n) \in \operatorname*{argmin}_{m \in \mathcal{M}} HO_T(\mathcal{A}_m, D_n) \ ,$$

measurably with respect to D_n. The overall learning rule is then given by

$$\widehat{f}_T^{\mathrm{ho}}\big((\mathcal{A}_m)_{m \in \mathcal{M}}, D_n\big) = \mathcal{A}_{\widehat{m}_T^{ho}(D_n)}(D_n^T) \ .$$

Hold-out depends on the arbitrary choice of a training set T, and is known to be quite unstable, despite its good theoretical properties [76, Section 8.5.1]. Therefore, practicioners often prefer to use cross-validation instead, which considers several training sets.

Definition 3.3.3 (Cross-validation) *Let D_n denote a dataset. Let $\mathcal{T}$ denote a collection of nonempty subsets of $\{1,\dots,n\}$. The associated cross-validation risk estimator of a learning rule $\mathcal{A}$ is defined by*

$$CV_{\mathcal{T}}(\mathcal{A}, D_n) = \frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}} HO_T\left(\mathcal{A}, D_n\right).$$

The cross-validation procedure then selects

$$\widehat{m}^{cv}_{\mathcal{T}}(D_n) \in \operatorname*{argmin}_{m\in\mathcal{M}} CV_{\mathcal{T}}(\mathcal{A}_m, D_n) \ .$$

The final predictor obtained through this procedure is

$$\widehat{f}^{\mathrm{cv}}_{\mathcal{T}}\big((\mathcal{A}_m)_{m\in\mathcal{M}}, D_n\big) = \mathcal{A}_{\widehat{m}^{cv}_{\mathcal{T}}(D_n)}(D_n) \ .$$

Depending on how $\mathcal{T}$ is chosen, this can lead to leave-one-out, leave-p-out, V-fold cross-validation or Monte-Carlo cross-validation, among others [3]. In the following, we omit some of the arguments $\mathcal{A}, D_n$ which appear in Definitions 3.3.2 and 3.3.3, when they are clear from context. For example, we often write $\mathrm{HO}_T\left(\mathcal{A}\right), \widehat{m}^{ho}_T, \widehat{f}^{\mathrm{ho}}_T$ instead of $\mathrm{HO}_T\left(\mathcal{A}, D_n\right), \widehat{m}^{ho}_T(D_n), \widehat{f}^{\mathrm{ho}}_T\big((\mathcal{A}_m)_{m\in\mathcal{M}}, D_n\big)$ (respectively).

3.3.2 Aggregated hold-out (Agghoo) estimators

In this paper, we study another way to improve on the stability of hold-out selection, by *aggregating* the predictors $\widehat{f}^{\mathrm{ho}}_T$ obtained by the hold-out procedure applied repeatedly with different training sets $T \in \mathcal{T}$. When $\mathbb{S}$ is convex (e.g., regression), *aggregated hold-out* (Agghoo) consists in averaging them.

Definition 3.3.4 (Agghoo) *Assume that $\mathbb{S}$ is a convex set. Let $(\mathcal{A}_m)_{m\in\mathcal{M}}$ denote a collection of learning rules, D_n a dataset, and $\mathcal{T}$ a collection of subsets of $\{1,\dots,n\}$. Using the notation of Definition 3.3.2, the associated Agghoo estimator is defined by*

$$\widehat{f}^{\mathrm{ag}}_{\mathcal{T}}\big((\mathcal{A}_m)_{m\in\mathcal{M}}, D_n\big) = \frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}} \widehat{f}^{\mathrm{ho}}_T\big((\mathcal{A}_m)_{m\in\mathcal{M}}, D_n\big) \ .$$

In the classification framework, as seen in Example 3.2.1, $\mathbb{S} = \{f : \mathcal{X} \to \{0, \dots, M\}\}$ which is not convex. However, there is still a natural way to aggregate several classifiers, by taking a majority vote.

Definition 3.3.5 (Majhoo) *Let $\mathcal{Y} = \{0, \dots, M\}$ be the set of labels. Given a collection of learning rules $(\mathcal{A}_m)_{m\in\mathcal{M}}$, a dataset D_n and a collection $\mathcal{T}$ of subsets of $\{1, \dots, n\}$, the majority hold-out (Majhoo) classifier is any measurable $\widehat{f}^{\mathrm{mv}}_{\mathcal{T}}\big((\mathcal{A}_m)_{m\in\mathcal{M}}, D_n\big) : \mathcal{X} \to \mathcal{Y}$ such that, using the notation $\widehat{f}^{\mathrm{ho}}_T$ introduced in Definition 3.3.2, for all $x \in \mathcal{X}$,*

$$\widehat{f}^{\mathrm{mv}}_{\mathcal{T}}\big((\mathcal{A}_m)_{m\in\mathcal{M}}, D_n\big)(x) \in \operatorname*{argmax}_{j\in\mathcal{Y}} \left| \left\{ T \in \mathcal{T} \;\middle|\; \widehat{f}^{\mathrm{ho}}_T\big((\mathcal{A}_m)_{m\in\mathcal{M}}, D_n\big)(x) = j \right\} \right| \ .$$

In most situations, it is clear how hold-out rules should be aggregated and there is no ambiguity in discussing hold-out aggregation. However, there is an important exception where both Agghoo and Majhoo can be used.

Remark 3.3.1 (Two options for binary classification) *In binary classification (Example 3.2.1 with $M = 2$), it is classical to consider classifiers of the form $\mathbb{I}_{f\geqslant 0}$ where $f \in \mathbb{S}_{conv} = \{f : \mathcal{X} \to \mathbb{R}\}$ aims at minimizing a surrogate convex risk associated with the loss $g_{conv} : (y', y) \mapsto \phi[(2y' - 1)(2y - 1)]$ with $\phi : \mathbb{R} \to \mathbb{R}$ convex [20]. Then, given a family of $\mathbb{S}_{conv}$-valued learning rules $\big(\mathcal{A}_m\big)_{m\in\mathcal{M}}$, one can either apply Agghoo to the surrogate problem and get*

$$\mathbb{I}_{\widehat{f}^{\mathrm{ag}}_{\mathcal{T}}\left((\mathcal{A}_m)_{m\in\mathcal{M}}, D_n\right)\geqslant 0} \ ,$$

or apply Majhoo to the binary classification problem and get

$$\widehat{f}^{\mathrm{mv}}_{\mathcal{T}}\left(\left(\mathbb{I}_{\mathcal{A}_m(\cdot)\geqslant 0}\right)_{m\in\mathcal{M}}, D_n\right) \ .$$

In the rest of this section, we focus on Agghoo, though much of the following discussion applies also to Majhoo.

Compared to cross-validation rules (Definition 3.3.3), Agghoo reverses the order between aggregation (majority vote or averaging) and minimization of the risk estimator: instead of averaging hold-out risk estimators before selecting the hyperparameter, the selection step is made first to produce hold-out predictors $\big(\widehat{f}^{\mathrm{ho}}_T\big)_{T\in\mathcal{T}}$ (given by Definition 3.3.2) and then an average is taken.

Related procedures To the best of our knowledge, Agghoo has not been studied theoretically before, though it is used in applications [54, 107], under the name "CV + averaging" in [107]. According to [107], Agghoo is commonly used by the machine learning community thanks to the Scikit-learn library [86].

A closely related procedure is "K-fold averaging cross-validation" (ACV), proposed by [59] for linear regression. With our general notation, ACV corresponds to averaging the $\mathcal{A}_{\widehat{m}^T_{ho}}(D_n)$, which are "retrained" on the whole dataset, while Agghoo averages the $\mathcal{A}_{\widehat{m}^T_{ho}}(D^T_n)$. An advantage of averaging the rules $\mathcal{A}_{\widehat{m}^T_{ho}}(D^T_n)$ is that they have been selected for their good performance on the validation set T^c, unlike the $\mathcal{A}_{\widehat{m}^T_{ho}}(D_n)$ whose performance has not been assessed on independent data. Furthermore, similarly to bagging, using several distinct training sets may result in improvements for unstable methods through a reduction in variance. Note finally that the theoretical results of [59] on ACV are limited to a specific setting, and much weaker than an oracle inequality.

A second family of related procedures is averaging the chosen *parameters* $\left(\widehat{m}^{ho}_T\right)_{T\in\mathcal{T}}$, contrary to Agghoo which averages the chosen *prediction rules*. This leads to different procedures for learning rules that are not linear functions of their parameters. This idea has been put forward under the name "bagged cross-validation" (BCV) [51] —with numerical and theoretical results in the case of bandwidth choice in kernel density estimation—, and under the name "efficient K-fold cross-validation" (EKCV) [58] for the choice of a regularization parameter in high-dimensional regression —with numerical results only. Unlike Agghoo, which only depends on the set $\{\mathcal{A}_m \,|\, m \in \mathcal{M}\}$ of learning rules, EKCV and BCV depend on the parametrization $m \mapsto \mathcal{A}_m$. Sometimes, the most natural parametrization does not allow the use of such procedures: for example, model dimensions are integers, and averaging them does not make sense. In contrast, in regression, it is always possible to average the real-valued functions $\mathcal{A}_m(D_{n_t}) \in \mathbb{S}$.

Even when all procedures are applicable, averaging rules is generally safer than averaging hyperparameters. Often in regression, the risk $\mathcal{L}$ is known to be convex over $\mathbb{S}$, so given $t_1, \ldots, t_V \in \mathbb{S}$,

$$\mathcal{L}\left(\frac{1}{V}\sum_{i=1} t_i\right) \leqslant \frac{1}{V}\sum_{i=1}^{V} \mathcal{L}(t_i) \ .$$

Hence, averaging regressors (Agghoo) always improves performance compared to selecting a single t_i at random (hold-out). On the other hand, if $(t_\theta)_{\theta\in\Theta}$ is a family of elements of $\mathbb{S}$ parametrized by a convex set Θ, there is no guarantee in general that the function $\theta \mapsto \mathcal{L}(t_\theta)$ is convex over Θ. So, for some $\theta_1, \ldots, \theta_V \in \Theta$, it may happen that

$$\mathcal{L}\left(t_{\frac{1}{V}\sum_{i=1}^{V}\theta_i}\right) \geqslant \frac{1}{V}\sum_{i=1}^{V} \mathcal{L}(t_{\theta_i}) \ .$$

In such a case, it is better to choose one parameter at random (hold-out) than to average them (EKCV or BCV).

A third family of related procedures is bagging or subagging applied to hold-out selection $D_n \mapsto \widehat{f}_T^{\text{ho}}((\mathcal{A}_m)_{m\in\mathcal{M}}, D_n)$. The bagging case has been studied numerically by [87], but clearly differs from Agghoo since it relies on bootstrap resamples, in which the original data can appear several times. Subagging —which is not explicitly studied in the literature, to the best of our knowledge— is closer to Agghoo, but there is still a slight difference. When applying subagging to the hold-out, the sample is divided into three parts: the training part of the bagging subsample, the validation part of the bagging subsample, and the data not in the bagging subsample. With Agghoo, the sample is only divided into two parts.

3.3.3 Computational complexity

In general, for a given value of $V = |\mathcal{T}|$, both Agghoo ($\widehat{f}_{\mathcal{T}}^{\text{ag}}$) and CV ($\widehat{f}_{\mathcal{T}}^{\text{cv}}$) must compute V hold-out risk estimators over all values of $m \in \mathcal{M}$. Let $C_{ho}(\mathcal{M}, n_t, n_v)$ be the average computational complexity of the hold-out, with a training dataset of size n_t and validation dataset of size n_v. Then the overall complexity of risk estimation is of order $V \times C_{ho}(\mathcal{M}, n_t, n_v)$ for both Agghoo and CV. Next, CV must average V risk vectors of length $|\mathcal{M}|$ and find a single minimum, while Agghoo computes V minima over $m \in \mathcal{M}$; these operations have similar complexity, of order $V \times |\mathcal{M}|$. Thus, computing the ensemble aggregated by Agghoo takes about as much time as selecting a learning rule using cross-validation.

A potential difference occurs when evaluating Agghoo and CV on new data. If there is no fast way to perform aggregation at training time, it is always possible to evaluate each predictor in the ensemble on the new data, and to average the results; then, Agghoo is slower than CV by a factor of order V at test time.

3.4 Theoretical results

The purpose of Agghoo is to construct an estimator whose risk is as small as possible, compared to the (unknown) best rule in the class $(\mathcal{A}_m)_{m\in\mathcal{M}}$. This is guaranteed theoretically by proving "oracle inequalities" of the form

$$\mathbb{E}\big[\ell(s, \widehat{f}_{\mathcal{T}}^{\text{ag}})\big] \leqslant C\mathbb{E}\Big[\inf_{m\in\mathcal{M}} \ell\big(s, \mathcal{A}_m(D_n)\big)\Big] + \varepsilon_n \ , \tag{3.1}$$

with ε_n negligible compared to the oracle excess risk $\mathbb{E}[\inf_{m\in\mathcal{M}} \ell(s, \mathcal{A}_m(D_{n_t}))]$ and C close to 1. Equation (3.1) then implies that Agghoo performs as well as the best choice of $m \in \mathcal{M}$, up to the constant C. In the following, we actually prove slightly weaker inequalities that are more natural in our setting.

By definition, Agghoo is an average of predictors chosen by hold-out over the collection $(\mathcal{A}_m)_{m\in\mathcal{M}}$. Therefore, when the risk is convex, an oracle inequality

(3.1) can be deduced from an oracle inequality for the hold-out, provided that there exists an integer $n_t \in \{1, \dots, n-1\}$ such that

$$\mathcal{T} \text{ is independent from } D_n \qquad \text{and} \qquad \forall T \in \mathcal{T}, \quad |T| = n_t \ . \tag{3.2}$$

We make this assumption in the rest of the article. Most cross-validation methods satisfy hypothesis (3.2), including leave-p-out, V-fold cross-validation (with $n - n_t = n_v = n/V$) and Monte-Carlo cross-validation [3].

In the remainder of this section, we introduce the RKHS setting of interest, and prove an oracle inequality for Agghoo without changing the standard estimators or requiring Y to be bounded.

3.4.1 Agghoo in regularized kernel regression

Kernel methods such as support vector machines, kernel least squares or ε-regression use a kernel function to map the data X_i into an infinite-dimensional function space, more specifically a reproducing kernel Hilbert space (RKHS) [95, 96]. We consider in this section regularized empirical risk minimization using a training loss function c, with a penalty proportional to the square norm of the RKHS, to solve the supervised learning problem (defined in Section 2.2) with loss function g. Hence, the contrast γ can be written $\gamma(t,(x,y)) = g(t(x),y) := (g \circ t)(x,y)$. We assume that g and c are convex in their first argument.

Definition 3.4.1 (Regularized kernel estimator) *Let $c : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be convex in its first argument, and let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive-definite kernel function. Given $\lambda > 0$ and training data $(X_i, Y_i)_{1 \leqslant i \leqslant n_t}$, define the regularized kernel estimator as*

$$\mathcal{A}_\lambda(D_{n_t}) = \operatorname*{argmin}_{t \in \mathcal{H}} \left\{ P_{n_t}(c \circ t) + \lambda \left\| t \right\|_{\mathcal{H}}^2 \right\} \ ,$$

where $\mathcal{H}$ is the reproducing kernel Hilbert space induced by K. By the representer theorem, $\mathcal{A}_\lambda$ can be computed explicitly:

$$\mathcal{A}_\lambda(D_{n_t})(x) = \sum_{j=1}^{n_t} \widehat{\theta}_{\lambda,j} K(X_j, x) \qquad \textit{where}$$

$$\widehat{\theta}_\lambda = \operatorname*{argmin}_{\theta \in \mathbb{R}^{n_t}} \left\{ \frac{1}{n_t} \sum_{i=1}^{n_t} c\left(\sum_{j=1}^{n_t} \theta_j K(X_j, X_i), Y_i \right) + \lambda \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} \theta_i \theta_j K(X_i, X_j) \right\} \ . \tag{3.3}$$

The loss function c is used to measure the accuracy of the fit on the training data: for example, taking $c : (u,y) \mapsto (1-uy)_+$ (the hinge loss) in Definition 3.4.1

corresponds to SVM. The loss function g used for risk evaluation may or may not be equal to c. For example, in classification, the 0–1 loss often cannot be used for training for computational reasons, hence a surrogate convex loss, such as the hinge loss, is used instead (see Remark 3.3.1), but there is no reason to use the hinge loss for risk estimation and hyperparameter selection.

In Definition 3.4.1, the hyperparameter of interest is λ (we assume that K is fixed). We show below some guarantees on Agghoo's performance when it is applied to a finite subfamily $(\mathcal{A}_\lambda)_{\lambda\in\Lambda}$ of the one defined by Definition 3.4.1. We first state some useful assumptions.

Hypothesis $Comp_C(g,c)$: $\mathcal{L}_c : t \mapsto P(c\circ t)$ and $\mathcal{L}_g$ have a common minimum $s \in \operatorname{argmin}_{t\in\mathbb{S}} \mathcal{L}_c(t) \cap \operatorname{argmin}_{t\in\mathbb{S}} \mathcal{L}_g(t)$ and for any $t \in \mathbb{S}$, $\mathcal{L}_c(t) - \mathcal{L}_c(s) \leqslant C\,[\mathcal{L}_g(t) - \mathcal{L}_g(s)]$.

Note that $Comp_1(g,c)$ is always satisfied when $g = c$. When $g \neq c$, some hypothesis relating c and g is necessary anyway for Definition 3.4.1 to be of interest, if only to ensure consistency (asymptotic minimization of the risk) for some sequence of hyperparameters $(\lambda_n)_{n\in\mathbb{N}}$.

In addition, some information about the evaluation loss g helps to obtain an oracle inequality (3.1) with a smaller remainder term ε_n.

Hypothesis $SC_{\rho,\nu}$: Let $\ell_X(u) = \mathbb{E}[g(u,Y)|X] - \inf_{v\in\mathbb{R}} \mathbb{E}[g(v,Y)|X]$. The triple (g,X,Y) satisfies $SC_{\rho,\nu}$ if and only if, for any $u, v \in \mathbb{R}$,

$$\mathbb{E}[(g(u,Y) - g(v,Y))^2|X] \leqslant \big[\rho \vee (\nu|u-v|)\big][\ell_X(u) + \ell_X(v)]. \tag{3.4}$$

For example, in the case of median regression, that is, $g(u,y) = |u-y|$, hypothesis $SC_{\rho,\nu}$ holds whenever there is a uniform lower bound on the concentration of Y around $s(X)$, as shown by the following proposition.

Proposition 3.4.2 *Let $g(u,y) = |u-y|$ for all $u, y \in \mathbb{R}$. For any $x \in \mathcal{X}$, let F_x be the conditional cumulative distribution function of Y knowing $X = x$. Assume that, for any $x \in \mathcal{X}$, F_x is continuous with a unique median $s(x)$ and that there exists $a(x) > 0, b(x) > 0$ such that*

$$\forall u \in \mathbb{R}, \qquad \Big|F_x(u) - F_x\big(s(x)\big)\Big| \geqslant a(x)\Big[\big|u - s(x)\big| \wedge b(x)\Big] \ . \tag{3.5}$$

For instance, this holds true if $\frac{dF_x}{du} \geqslant a(x)\mathbb{I}_{|u-s(x)|\leqslant b(x)}$ for every $x \in \mathcal{X}$. Let

$$a_m = \inf_{x\in\mathcal{X}}\big\{a(x)\big\} \qquad \text{and} \qquad \mu_m = \inf_{x\in\mathcal{X}}\big\{a(x)b(x)\big\} \ .$$

If $a_m > 0$ and $\mu_m > 0$, then (g,X,Y) satisfies $SC_{\frac{4}{a_m},\frac{2}{\mu_m}}$.

Proposition 3.4.2 is proved in Appendix 3.9.1. We can now state our first main result.

Theorem 3.4.3 *Let $\Lambda \subset \mathbb{R}_+^*$ be a finite grid. Using the notation of Definition 3.3.4, let $\widehat{f}_{\mathcal{T}}^{\mathrm{ag}}$ be the output of Agghoo, applied to the collection $(\mathcal{A}_\lambda)_{\lambda\in\Lambda}$ given by Definition 3.4.1. Assume that $\lambda_m = \min\Lambda > 0$ and $\kappa = \sup_{x\in\mathcal{X}} K(x,x) < +\infty$. Assume that $Comp_C(g,c)$ holds for a constant $C > 0$ and that (g,X,Y) satisfies $SC_{\rho,\nu}$ with constants $\rho \geqslant 0, \nu \geqslant 0$. Assume that c and g are convex and Lipschitz in their first argument, with Lipschitz constant less than L. Assume also that $n_v \geqslant 100$ and $3 \leqslant |\Lambda| \leqslant \mathrm{e}^{\sqrt{n_v}}$. Then, for any $\theta \in (0;1]$,*

$$(1-\theta)\mathbb{E}\Big[\ell\big(s,\widehat{f}_{\mathcal{T}}^{\mathrm{ag}}\big)\Big] \leqslant (1+\theta)\mathbb{E}\Big[\min_{\lambda\in\Lambda}\ell\big(s,\mathcal{A}_\lambda(D_{n_t})\big)\Big] + \max\left\{18\rho\frac{\log\big(n_v|\Lambda|\big)}{\theta n_v}, b_1\frac{\log^2\big(n_v|\Lambda|\big)}{\theta^3\lambda_m n_v^2}, b_2\frac{\log^{\frac{3}{2}}\big(n_v|\Lambda|\big)}{\theta\lambda_m n_v\sqrt{n_t}}\right\}, \tag{3.6}$$

where b_1, b_2 do not depend on n_v, n_t, λ_m or θ but only on κ, L, ν and C.

Theorem 3.4.3 is proved in Appendix 3.8 as a consequence of a result valid in the general framework of Section 3.2.1 (Theorem 3.7.3). It shows that $\widehat{f}_{\mathcal{T}}^{\mathrm{ag}}$ satisfies an oracle inequality of the form (3.1), with $\mathcal{A}_\lambda(D_{n_t})$ instead of $\mathcal{A}_\lambda(D_n)$ on the right-hand side of the inequality. The fact that D_{n_t} appears in the bound instead of D_n is a limitation of our result, but it is natural since predictors aggregated by Agghoo are only trained on part of the data. In most cases, it can be expected that $\ell(s,\mathcal{A}_\lambda(D_{n_t}))$ is close to $\ell(s,\mathcal{A}_\lambda(D_n))$ whenever $\frac{n_t}{n}$ is close to 1.

The assumption that K is bounded is mild. For instance, popular kernels such as Gaussian kernels, $(x,x') \mapsto \exp[-\|x-x'\|^2/(2h^2)]$ for some $h > 0$, or Laplace kernels, $(x,x') \mapsto \exp(-\|x-x'\|/h)$ for some $h > 0$, are bounded by $\kappa = 1$.

Taking $|\mathcal{T}| = 1$ in Theorem 3.4.3 yields a new oracle inequality for the hold-out. Oracle inequalities for the hold-out have already been proved in a variety of settings (see [3] for a review), and used to obtain adaptive rates in regularized kernel regression [96]. However, this work has mostly been accomplished under the assumption that the contrast $\gamma\left(\mathcal{A}_\lambda(D_n),(X,Y)\right)$ is bounded uniformly (in n, D_n and $\lambda \in \Lambda$) by a constant. If this constant increases with n, bounds obtained in this manner may worsen considerably. As many "natural" regression procedures —including regularized kernel regression (Definition 3.4.1)— fail to satisfy such bounds, some theoreticians introduce "truncated" versions of standard procedures [96], but truncation has no basis in practice. Theorem 3.4.3 avoids these complications.

In order to be satisfactory, Theorem 3.4.3 should prove that Agghoo performs asymptotically as well as the best choice of $\lambda \in \Lambda$, at least for reasonable choices

of Λ. This is the case whenever the maximum in Equation (3.6) is negligible with respect to the oracle excess risk $\mathbb{E}[\min_{\lambda\in\Lambda} \ell(s, \mathcal{A}_\lambda(D_{n_t}))]$ as $n \to +\infty$. This depends on the range $[\lambda_m; +\infty)$ in which the hold out is allowed to search for the optimal λ. On the one hand, it is desirable that this interval be wide enough to contain the true optimal value. On the other hand, if $\lambda_m = 0$, then inequality (3.6) becomes vacuous. We now provide precise examples where Theorem 3.4.3 applies with a remainder term in Equation (3.6) that is negligible relative to the oracle excess risk.

Take the example of median regression, in which $c(u, y) = g(u, y) = |u - y|$. Then $Comp_1(g, c)$ holds trivially. Make also the same assumptions as in Proposition 3.4.2, which ensures that $SC_{\rho,\nu}$ holds for some finite values of ρ and ν. Theorem 3.4.3 therefore applies as long as the kernel K is bounded and $\lambda_m > 0$. Choose $n_v = n_t = \frac{n}{2}$ and Λ of cardinality at most polynomial in n (which is sufficient in theory and in practice). Then [96, Theorem 9.6] proves the consistency of $\mathcal{A}_{\lambda_n}(D_n)$ as $n \to +\infty$, provided that $\lambda_n^2 n \to +\infty$. This suggests choosing $\lambda_m = 1/\sqrt{n_t}$, in which case the remainder term of Equation (3.6) is of order $(\log n)^{3/2}/n$, which is negligible relative to nonparametric convergence rates in median regression.

In order to have a more precise idea of the order of magnitude of the oracle excess risk, let us consider median regression with a Gaussian kernel. Under some assumptions, one of which coincides with Proposition 3.4.2, [40, Corollary 4.12] shows that taking $\lambda_n = \frac{c_1}{n}$ leads to rates of order $n^{-\frac{2\alpha}{2\alpha+d}}$, where $d \in \mathbb{N}$ is the dimension of $\mathcal{X}$ and $\alpha > 0$ is the smoothness of s. Therefore, taking $\lambda_m = 1/n_t$ in Theorem 3.4.3, the remainder term of Equation (3.6) is at most of order $(\log n)^{3/2}/\sqrt{n}$, hence negligible relative to the above risk rates as soon as $2\alpha < d$.

Theorem 3.4.3 can handle situations where g is different from the training loss c, provided that $Comp(g, c)$ holds true. Such situations arise for instance in the case of support vector regression [95, Chapter 9], which uses for training Vapnik's ε-insensitive loss $c_\varepsilon^{eps}(u, y) = (|u - y| - \varepsilon)_+$. This loss depends on a parameter ε, the choice of which is usually motivated by a tradeoff between sparsity and prediction accuracy [95]. Therefore, some other loss is typically used to measure predictive performance, independently of ε. We state one possible application of Theorem 3.4.3 to this case, as a corollary.

Corollary 3.4.4 (ε-regression) *Let $c = c_\varepsilon^{eps} : (u, y) \mapsto (|y-u|-\varepsilon)_+$ be Vapnik's ε-insensitive loss and assume that the evaluation loss is $g = c_0^{eps} : (u, y) \mapsto |u - y|$. Assume that for every x the conditional distribution of Y given $X = x$ has a unimodal density with respect to the Lebesgue measure, symmetric around its mode.*

Introduce the robust noise parameter:

$$\sigma = \sup_{x\in\mathcal{X}} \left\{ \inf\left\{ y\in\mathbb{R} \,\middle|\, \mathbb{P}(Y\leqslant y \,|\, X=x) \geqslant \frac{3}{4} \right\} - \sup\left\{ y\in\mathbb{R} \,\middle|\, \mathbb{P}(Y\leqslant y \,|\, X=x) \leqslant \frac{1}{4} \right\} \right\} \quad . \tag{3.7}$$

Then, applying Agghoo to a finite subfamily $(\mathcal{A}_\lambda)_{\lambda\in\Lambda}$ *of the rules given by Definition 3.4.1 with* $c = c_\varepsilon^{eps}$ *and a kernel* K *such that* $\|K\|_\infty \leqslant 1$ *yields the following oracle inequality. Assuming* $n_v \geqslant 100$ *and* $3 \leqslant |\Lambda| \leqslant \mathrm{e}^{\sqrt{n_v}}$*, for any* $\theta \in (0;1]$*,*

$$(1-\theta)\mathbb{E}\Big[\ell\big(s,\widehat{f}_{\mathcal{T}}^{\mathrm{ag}}\big)\Big] \leqslant (1+\theta)\mathbb{E}\Big[\min_{\lambda\in\Lambda}\ell\big(s,\mathcal{A}_\lambda(D_{n_t})\big)\Big] + \max\left\{ 72\sigma\frac{\log\big(n_v|\Lambda|\big)}{\theta n_v} \,,\, b_1\frac{\log^2\big(n_v|\Lambda|\big)}{\theta^3\lambda_m n_v^2} \,,\, b_2\frac{\log^{\frac{3}{2}}\big(n_v|\Lambda|\big)}{\theta\lambda_m n_v\sqrt{n_t}} \right\} \quad ,$$

where b_1 *and* b_2 *are absolute constants.*

Corollary 3.4.4 is proved in Appendix 3.9.2.

When $\varepsilon = 0$, ε-regression becomes median regression, which is discussed above. The oracle inequality of Corollary 3.4.4 is then the same as that given by Theorem 3.4.3 and Proposition 3.4.2. Assumptions of unimodality and symmetry allow to give more explicit values of a_m and μ_m in terms of σ. When $\varepsilon > 0$, the unimodality and symmetry assumptions are used to prove hypothesis $Comp_C(g,c)$.

3.4.2 Classification

Loss functions are not all convex. When convexity fails, the aggregation procedure should be revised.

In classification, Majhoo is a possible solution (see Definition 3.3.5). By Proposition 3.10.1 in Appendix 3.10, majority voting satisfies a kind of "convexity inequality" with respect to the 0–1 loss; as a result, oracle inequalities for the hold-out imply oracle inequalities for majhoo.

Hold-out for binary classification with 0–1 loss has been studied by Massart [76]. In that work, Massart makes an assumption which is closely related to margin hypotheses, such as the Tsybakov noise condition [75] which we consider here. This approach allows to derive the following theorem.

Theorem 3.4.5 *Consider the classification setting described in Example 3.2.1 with* $M = 2$ *classes (binary classification). Let* $(\mathcal{A}_m)_{m\in\mathcal{M}}$ *be a collection of learning rules and* $\mathcal{T}$ *a collection of training sets satisfying assumption* (3.2)*.*

Assume that there exists $\beta \geqslant 0$ *and* $r \geq 1$ *such that for* $\xi = (X, Y)$ *with distribution* P*,*

$$\forall h > 0, \qquad \mathbb{P}\big(\big|2\eta(X) - 1\big| \leqslant h\big) \leqslant r h^{\beta} \tag{MA}$$

where $\eta(X) := \mathbb{P}(Y = 1 \,|\, X)$*. Then, we have*

$$\mathbb{E}\Big[\ell\big(s, \widehat{f}_{\mathcal{T}}^{\mathrm{mv}}\big)\Big] \leqslant 3\mathbb{E}\left[\inf_{m \in \mathcal{M}} \ell\big(s, \mathcal{A}_m(D_{n_t})\big)\right] + \frac{29 r^{\frac{1}{\beta+2}} \log\big(\mathrm{e}|\mathcal{M}|\big)}{n_v^{\frac{\beta+1}{\beta+2}}} \ .$$

Theorem 3.4.5 is proved in Appendix 3.10. It shows that $\widehat{f}_{\mathcal{T}}^{\mathrm{mv}}$, like $\widehat{f}_{\mathcal{T}}^{\mathrm{ag}}$, satisfies an oracle inequality of the form (3.1) with $\mathcal{A}_\lambda(D_{n_t})$ instead of $\mathcal{A}_\lambda(D_n)$. Tsybakov's noise condition (MA) only depends on the distribution of (X, Y) and not on the collection of learning rules. It is a standard hypothesis in classification, under which "fast" learning rates —faster than $n^{-1/2}$— are attainable [103]. In contrast with the results of Section 3.4.1, that are valid for various losses but only for a specific type of learning rule, Theorem 3.4.5 holds true for *any* family of classification rules.

The constant 3 in front of the oracle excess risk can be replaced by any constant larger than 2, at the price of increasing the constant in the remainder term, as can be seen from the proof (in Appendix 3.10). However, our approach cannot yield a constant lower than 2, because we use Proposition 3.10.1 instead of a convexity argument, since the 0–1 loss is not convex.

3.5 Numerical experiments

This section investigates how Agghoo and Majhoo's performance vary with their parameters V and $\tau = \frac{n_t}{n}$, and how it compares to CV's performance at a similar computational cost —that is, for the same values of V and τ. Two settings are considered, corresponding to Corollary 3.4.4 and Theorem 3.4.5.

3.5.1 ε-regression

Consider the collection $(\mathcal{A}_\lambda)_{\lambda \in \Lambda}$ of regularized kernel estimators (see Definition 3.4.1) with loss function $c_\varepsilon^{eps}(u, y) = (|u - y| - \varepsilon)_+$ and Gaussian kernel $K(x, x') = \exp[-(x - x')^2/(2h^2)]$ over $\mathcal{X} = \mathbb{R}$.

Experimental procedure Agghoo and CV training sets $T \in \mathcal{T}$ are chosen independently and uniformly among the subsets of $\{1, \ldots, n\}$ with cardinality $\lfloor \tau n \rfloor$, for different values of τ and $V = |\mathcal{T}|$; hence, CV corresponds to what is usually called "Monte-Carlo CV" [3]. Each algorithm is run on 1000 independent samples

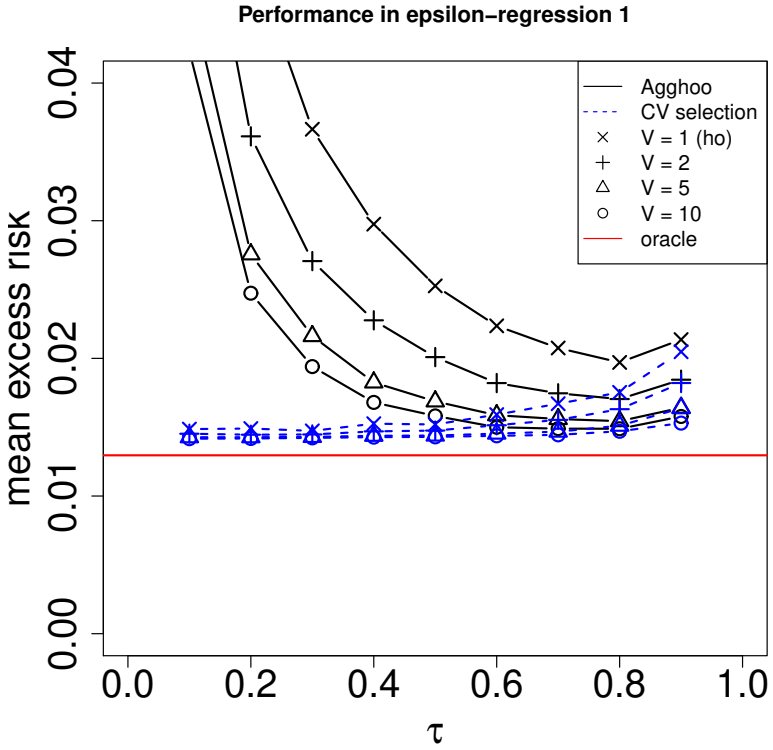

Figure 3.1: Performance of Agghoo and CV for ε-regression in setup 1

of size $n = 500$, and independent test samples of size 1000 are used for estimating the excess risks $\ell(s, \widehat{f}^{\mathrm{ag}}_{\mathcal{T}})$, $\ell(s, \widehat{f}^{\mathrm{cv}}_{\mathcal{T}})$ and the oracle excess risk $\inf_{\lambda\in\Lambda} \ell(s, \mathcal{A}_\lambda(D_n))$. The risks (and excess risks) are evaluated using the L^1 loss $g(u, y) = |u - y|$. Expectations of these quantities are estimated by taking an average over the 1000 samples; we also compute standard deviations for these estimates, which are not displayed, since they are sufficiently small to ensure that visible "gaps" on the graph are statistically significant.

Agghoo and CV are applied to $(\mathcal{A}_\lambda)_{\lambda\in\Lambda}$ over the grid $\Lambda = \{\frac{2^{j-1}}{500 n_t} \,|\, 0 \leqslant j \leqslant 17\}$, corresponding to the grid $\{\frac{500}{2^j} \,|\, 0 \leqslant j \leqslant 17\}$ over the "cost" parameter $C = \frac{1}{2\lambda n_t}$ of the R implementation "svm" from package e1071.

Experimental setup 1 Data $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent, with $X_i \sim \mathcal{N}(0, \pi^2)$, $Y_i = s(X_i) + Z_i$, with $Z_i \sim \mathcal{N}(0, 1/4)$ independent from X_i. The regression function is $s : x \mapsto e^{\cos(x)}$, the kernel parameter is $h = \frac{1}{2}$ and the threshold for the ε-insensitive loss is $\varepsilon = \frac{1}{4}$.

Results in setup 1 are shown on Figure 3.1. The performance of Agghoo strongly depends on both τ and V. For a fixed τ, increasing V significantly decreases the risk of the resulting estimator. This is not surprising and confirms that considering several data splits is always useful.

Most of the improvement occurs between $V = 1$ and $V = 5$, and taking V much larger seems useless —at least for $\tau \geqslant 0.5$—, a behavior previously observed for CV [5]. For a fixed V, the risk strongly decreases when τ increases from 0.1 to 0.5, decreases slowly over the interval $[0.5, 0.8]$ and seems to rise for $\tau > 0.8$. It seems that $\tau \in [0.6, 0.9]$ yields the best performance, while taking τ close to 0 should clearly be avoided (at least for $V \leqslant 10$). Taking V large enough, say $V = 10$, makes the choice of τ less crucial: a large region of values of τ yield (almost) optimal performance. We do not know whether taking V larger can make the performance of Agghoo with $\tau \leqslant 0.4$ close to the optimum.

As a function of τ, the risk of CV behaves quite differently from Agghoo's. The performance does not degrade significantly when τ is small. The optimum is located around $\tau = 0.1$, but the risk curve is so flat that there is no perceptible difference between the values of $\tau \in [0.1, 0.4]$. In any case, the optimum is much smaller than for Agghoo. A possible explanation is that the regressors produced by cross-validation are all trained on the whole sample, so that τ only impacts risk estimation. Furthermore, additional simulations show, as expected, that higher values of τ ($\tau = 0.8$ or $\tau = 0.9$) improve *risk estimation* while degrading the *hyperparameter selection* performance. Compared to Agghoo, CV's performance depends much less on V: only $V = 2$ appears to be significantly worse than $V \geqslant 5$.

Let us now compare Agghoo and CV. For small values of τ ($\tau \leqslant 0.5$), Agghoo generally performs much worse than CV for all values of V. In the case of the hold-out, this is unsurprising as the hold-out estimator is then trained on a much smaller sample than the CV estimator. Clearly, aggregation does not sufficiently compensate for this, at least for $V \leqslant 10$. On the other hand, for $\tau \in [0.6, 0.9]$, Agghoo with $V = 10$ approximately matches CV's performance. The risks of the two methods are indistinguishable for $V = 10, \tau = 0.8$.

The regression function $e^{\cos(x)}$ of setup 1 is very smooth (analytic) and bounded. Combined with a one-dimensional variable X and gaussian noise, this yields a comparatively "easy" non-parametric regression problem. Aggregation may prove more useful in harder problems where s is less smooth and the dimension is higher. To test this, we carry out a second simulation.

Experimental setup 2 Data $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent, with $X_i \in \mathbb{R}^2, X_i \sim \text{Cauchy}(0,1)^{\otimes 2}$, $Y_i = s(X_i) + Z_i$, with $Z_i \sim \mathcal{N}(0, 1/4)$ independent from X_i. The regression function is defined almost everywhere by $s(x_1, x_2) = \frac{2\sin(x_1 x_2)}{x_1^2 + x_2^2}$, the kernel parameter is $h = \frac{1}{2}$ and the threshold for the ε-insensitive loss is $\varepsilon = \frac{1}{4}$.

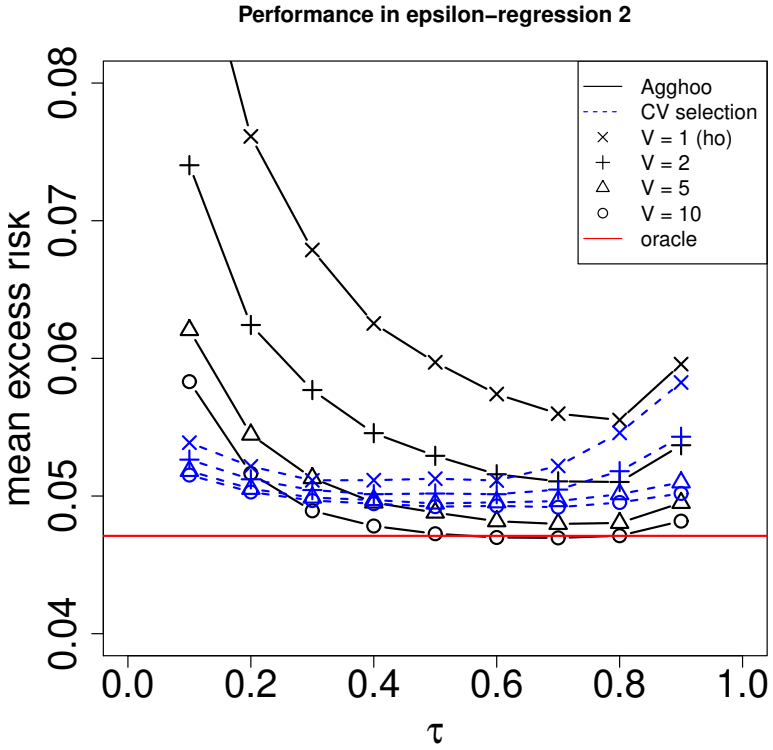

Figure 3.2: Performance of Agghoo and CV for ε-regression in setup 2

This regression function is less regular than in the previous setup, since it has a discontinuity at $(0, 0) \in \mathbb{R}^2$.

Results in setup 2 are shown on figure 3.2 The qualitative conclusions about the behaviour of Agghoo and CV, taken separately, are mostly the same as in setup 1, with the exception that CV now shows the expected increase in risk for the smallest values of τ.

The main difference with setup 1 is that Agghoo performs much better relative to CV and the oracle. For $V = 10$ and $\tau \in [0.4, 0.9]$, Agghoo outperforms CV by a significant margin; for $V = 10$ and $\tau \in [0.6, 0.8]$, Agghoo even matches the oracle's performance, up to statistical measurement error.

Part of the explanation is that, on a given dataset, Agghoo can perform better than the oracle using aggregation whereas CV, as a parameter selection method, naturally cannot. Indeed, for a randomly drawn dataset in setup 2, this situation can be observed to occur quite regularly.

Overall, if the computational cost of $V = 10$ data splits is not prohibitive, Agghoo with optimized parameters ($V = 10$, $\tau \in [0.6, 0.8]$) clearly improves over

CV with optimized parameters ($V = 10$, $\tau \in [0.5, 0.7]$). The same holds with $V = 5$. This advocates for the use of Agghoo instead of CV, unless we have to take $V \leqslant 5$ for computational reasons.

Computational complexity By Equation (3.3), regularized kernel regressors can be represented linearly by vectors of length n_t, therefore the aggregation step can be performed at training time by averaging these vectors. The complexity of this aggregation is at most $\mathcal{O}(V \times n_t)$. In general, this is negligible relative to the cost of computing the hold-out, as simply computing the kernel matrix requires $n_t(n_t + 1)/2$ kernel evaluations. Therefore, the aggregation step does not affect much the computational complexity of Agghoo, so the conclusion of Section 3.3.3 that Agghoo and CV have similar complexity applies in the present setting.

Evaluating Agghoo and CV on new data $x \in \mathcal{X}$ also takes the same time in general, as both are computed by evaluating the expression $\sum_{j=1}^{n_t} \theta_j K(X_j, x)$ with a pre-computed value of θ. A potential difference occurs when the $\widehat{\theta}_\lambda$ —given by Definition 3.4.1, Equation (3.3)— are sparse: aggregation increases the number of non-zero coefficients, so evaluating $\widehat{f}^{\mathrm{ag}}_{\mathcal{T}}$ on new data can be slower than evaluating $\widehat{f}^{\mathrm{cv}}_{\mathcal{T}}$ if the implementation is designed to take advantage of sparsity.

3.5.2 k-nearest neighbors classification

Consider the collection $(\mathcal{A}^{\mathrm{NN}}_k)_{k \geqslant 1,\, k \text{ odd}}$ of nearest-neighbors classifiers —assuming k is odd to avoid ties— on the following binary classification problem.

Experimental setup Data $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent, with X_i uniformly distributed over $\mathcal{X} = [0, 1]^2$ and

$$\mathbb{P}(Y_i = 1 \,|\, X_i) = \sigma\left(\frac{g(X_i) - b}{\lambda}\right)$$

$$\text{where } \forall u, v \in \mathbb{R}, \qquad \sigma(u) = \frac{1}{1 + \mathrm{e}^{-u}} \quad \text{and} \quad g(u, v) = \mathrm{e}^{-(u^2+v)^3} + u^2 + v^2 \ ,$$

$b = 1.18$ and $\lambda = 0.05$. The Bayes classifier is $s : x \mapsto \mathbb{I}_{g(x) \geqslant b}$ and the Bayes risk, computed numerically using the scipy.integrate python library, is approximately equal to 0.242. Majhoo (the classification version of Agghoo, see Definition 3.3.5) and CV are used with the collection $(\mathcal{A}^{\mathrm{NN}}_k)_{k \geqslant 1,\, k \text{ odd}}$ and "Monte Carlo" training sets as in Section 3.5.1. An experimental procedure similar to the one of Section 3.5.1 is used to evaluate the performance of Agghoo and to compare it with Monte-Carlo cross-validation. Standard deviations of the excess risk were computed; they are smaller than 3.6% of the estimated value.

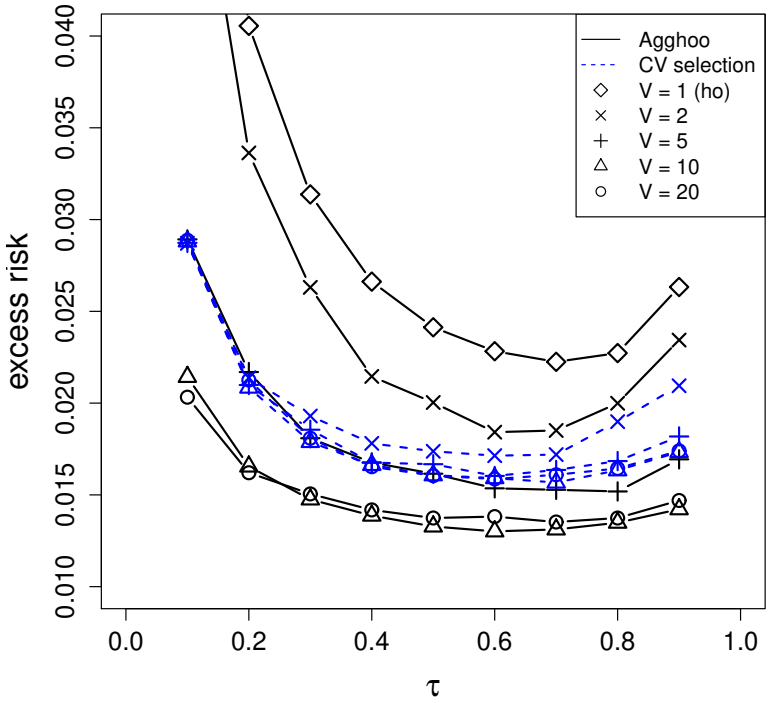

Figure 3.3: Classification performance of Majhoo and CV for the k-NN family

Results are shown on Figure 3.3. They are similar to the regression case (see Section 3.5.1), with a few differences. First, Agghoo does not perform better than the oracle. In fact, all methods considered here remain far from the oracle, which has an excess risk around 0.0034 ± 0.0004; both Agghoo and CV have excess risks at least 4 times larger. Second, risk curves as a function of τ for Agghoo are almost U-shaped, with a significant rise of the risk for $\tau > 0.6$. Therefore, less data is needed for training, compared to Section 3.5.1. The optimal value of τ here is 0.6, at least for some values of V, up to statistical error. Third, the performance of CV as a function of τ has a similar U-shape, which makes the comparison between Agghoo and CV easier. For a given τ, Agghoo performs significantly better if $V \geqslant 10$, while CV performs significantly better if $V = 2$; the difference is mild for $V = 5$.

Computational complexity As said in Section 3.3.3, the complexity of computing the optimal parameters for CV ($\hat{k}^{cv}_{\mathcal{T}}$) is the same as for Majhoo ($(\hat{k}^{ho}_{T})_{T\in\mathcal{T}}$). Here, there is no simple way to represent the aggregated estimator, so aggregation may have to be performed at test time. In that case, the complexity of evaluating Majhoo on new data is roughly V times greater than for CV, as explained in Section 3.3.3 for Agghoo.

3.6 Discussion

Theoretical and numerical results of the paper show that Agghoo can be used safely in RKHS regression, at least when its parameters are properly chosen; $V \geqslant 10$ and $\tau = 0.8$ seem to be safe choices. A variant, Majhoo, can be used in supervised classification with the 0–1 loss, with a general guarantee on its performance (Theorem 3.4.5). Experiments show that Agghoo actually performs much better than what the upper bounds of Section 3.4 suggest, with a significant improvement over cross-validation except when $V < 5$ splits are used. Proving theoretically that Agghoo can improve over CV is an open problem that deserves future works.

Since Agghoo and CV have the same training computational cost for fixed (V, τ), Agghoo —with properly chosen parameters V, τ— should be preferred to CV, unless aggregation is undesirable for some other reason, such as interpretability of the predictors, or computational complexity at test time.

Our results can be extended in several ways. First, our theoretical bounds directly apply to subagging hold-out, which also averages several hold-out selected estimators. The difference is that, in subagging, the training set size is $n-p-q$ and the validation set size is q, for some $q \in \{1, \dots, n-p-1\}$, leading to slightly worse bounds than those we obtained for Agghoo (at least if $\mathbb{E}\left[\ell(s, \mathcal{A}_m(D_n))\right]$ decreases with n). The difference should not be large in practice, if q is well chosen.

Oracle inequalities can also be obtained for Agghoo in other settings, as a consequence of our general theorems 3.7.2 and 3.7.3 in Appendix 3.7.

3.7 General Theorems

We need the following hypothesis, defined for two functions $w_i : \mathbb{R}_+ \to \mathbb{R}_+$, $i \in \{1;2\}$ and a family $(t_m)_{m\in\mathcal{M}} \in \mathbb{S}^{\mathcal{M}}$.

Hypothesis $H(w_1, w_2, (t_m)_{m\in\mathcal{M}})$: w_1 and w_2 are non-decreasing, and for any $(m, m') \in \mathcal{M}^2$, some $c_{m'}^m \in \mathbb{R}$ exists such that, for all $k \geqslant 2$,

$$P\Big(\big|\gamma(t_m) - \gamma(t_{m'}) - c_{m'}^m\big|^k\Big) \leqslant k!\Big[w_1\big(\sqrt{\ell(s,t_m)}\big) + w_1\big(\sqrt{\ell(s,t_{m'})}\big)\Big]^2 \times \Big[w_2\big(\sqrt{\ell(s,t_m)}\big) + w_2\big(\sqrt{\ell(s,t_{m'})}\big)\Big]^{k-2} .$$

This hypothesis is similar to those used by Massart [76] to study the hold-out and empirical risk minimizers. However, unlike [76], we intend to go beyond the setting of bounded risks.

We also need the following definition.

Definition 3.7.1 *Let $w : \mathbb{R}_+ \to \mathbb{R}_+$ and $r \in \mathbb{R}_+$. Let*

$$\delta(w, r) = \inf\left\{\delta \geqslant 0 : \forall x \geqslant \delta, w(x) \leqslant rx^2\right\},$$

with the convention $\inf \emptyset = +\infty$.

Remark 3.7.1

- *If $r > 0$ and $x \mapsto \frac{w(x)}{x}$ is nonincreasing, then $\delta(w, r)$ is the unique solution to the equation $\frac{w(x)}{x} = rx$.*
- *$r \mapsto \delta(w, r)$ is nonincreasing.*
- *If $w(x) = cx^\beta$ for $c > 0$ and $\beta \in [0; 2)$, then $\delta(w, r) = \left(\frac{c}{r}\right)^{\frac{1}{2-\beta}}$.*

3.7.1 Theorem statements

We can now state two general theorems from which we deduce all the theoretical results of the paper. The first theorem is a general oracle inequality for the hold-out.

Theorem 3.7.2 *Let $(t_m)_{m\in\mathcal{M}}$ be a finite collection in $\mathbb{S}$, and*

$$\widehat{m} \in \operatorname*{argmin}_{m\in\mathcal{M}} P_{n_v}\gamma(t_m, \cdot) .$$

Assume that $H(w_1, w_2, (t_m)_{m\in\mathcal{M}})$ holds true. Let $x > 0$. Then, with probability larger than $1 - \mathrm{e}^{-x}$, for any $\theta \in (0; 1]$, we have

$$(1-\theta)\,\ell(s, t_{\widehat{m}}) \leqslant (1+\theta) \min_{m\in\mathcal{M}} \ell(s, t_m) + \sqrt{2}\theta\delta^2\left(w_1, \frac{\theta}{2}\sqrt{\frac{n_v}{x+\log|\mathcal{M}|}}\right) + \frac{\theta^2}{2}\delta^2\left(w_2, \frac{\theta^2}{4}\frac{n_v}{x+\log|\mathcal{M}|}\right). \tag{3.8}$$

If in addition, the two functions $x \mapsto \frac{w_j(x)}{x}$, $j = 1, 2$, are nonincreasing, then for any $x > 0$, with probability larger than $1 - \mathrm{e}^{-x}$, for all $\theta \in (0; 1]$, we have

$$(1-\theta)\ell(s, t_{\widehat{m}}) \leqslant (1+\theta) \min_{m\in\mathcal{M}} \ell(s, t_m) + \delta^2(w_1, \sqrt{n_v})\left[\theta + \frac{2(x+\log|\mathcal{M}|)}{\theta}\right] \tag{3.9}$$

$$+ \delta^2(w_2, n_v)\left[\theta + \frac{(x+\log|\mathcal{M}|)^2}{\theta}\right] . \tag{3.10}$$

Using Theorem 3.7.2, we prove the following general oracle inequality for Agghoo.

Theorem 3.7.3 *Assume that the hyperparameter space $\mathbb{S}$ is convex and that the risk $\mathcal{L}$ is convex. Let $(\mathcal{A}_m)_{m\in\mathcal{M}}$ be a finite collection of learning rules of size $|\mathcal{M}| \geqslant 3$. Let $\widehat{f}^{\mathrm{ag}}_{\mathcal{T}}$ be an Agghoo estimator, according to Definition 3.3.4, with $\mathcal{T}$ satisfying assumption (3.2). Assume that $\widehat{w}_{1,1}, \widehat{w}_{1,2}$ are D_{n_t}-measurable random functions such that almost surely, $H\big(\widehat{w}_{1,1}, \widehat{w}_{1,2}, (\mathcal{A}_m(D_{n_t}))_{m\in\mathcal{M}}\big)$ holds true. Assume also that for $i \in \{1, 2\}$, $x \mapsto \frac{\widehat{w}_{1,i}(x)}{x}$ is non-increasing. Then for any $\theta \in (0; 1]$,*

$$(1-\theta)\mathbb{E}\Big[\ell\big(s, \widehat{f}^{\mathrm{ag}}_{\mathcal{T}}\big)\Big] \leqslant (1+\theta)\mathbb{E}\left[\min_{m\in\mathcal{M}} \ell\big(s, \mathcal{A}_m(D_{n_t})\big)\right] + R_1(\theta) \tag{3.11}$$

where $R_1(\theta) = R_{1,1}(\theta) + R_{1,2}(\theta)$ with

$$R_{1,1}(\theta) = \left(\theta + \frac{2\big(1+\log|\mathcal{M}|\big)}{\theta}\right)\mathbb{E}\Big[\delta^2\big(\widehat{w}_{1,1}, \sqrt{n_v}\big)\Big] ,$$

$$R_{1,2}(\theta) = \left(\theta + \frac{2\big(1+\log|\mathcal{M}|\big) + \log^2|\mathcal{M}|}{\theta}\right)\mathbb{E}\Big[\delta^2\big(\widehat{w}_{1,2}, n_v\big)\Big] .$$

Now, for any D_{n_t}-measurable functions $\widehat{w}_{2,1}$ and $\widehat{w}_{2,2}$ such that assumption $H(\widehat{w}_{2,1}, \widehat{w}_{2,2}, (\mathcal{A}_m(D_{n_t}))_{m\in\mathcal{M}})$ holds true almost surely, and any $x > 0$, $\theta \in (0; 1]$, we have

$$(1-\theta)\mathbb{E}\Big[\ell\big(s, \widehat{f}^{\mathrm{ag}}_{\mathcal{T}}\big)\Big] \leqslant (1+\theta)\mathbb{E}\left[\min_{m\in\mathcal{M}} \ell\big(s, \mathcal{A}_m(D_{n_t})\big)\right] + R_2(\theta) \tag{3.12}$$

where $R_2(\theta) = R_{2,1}(\theta) + R_{2,2}(\theta) + R_{2,3}(\theta) + R_{2,4}(\theta)$ *with*

$$R_{2,1}(\theta) = \sqrt{2}\theta\mathbb{E}\left[\delta^2\left(\widehat{w}_{2,1}, \frac{\theta}{2}\sqrt{\frac{n_v}{x+\log|\mathcal{M}|}}\right)\right] \ ,$$
$$R_{2,2}(\theta) = \frac{\theta^2}{2}\mathbb{E}\left[\delta^2\left(\widehat{w}_{2,2}, \frac{\theta^2}{4}\frac{n_v}{x+\log|\mathcal{M}|}\right)\right] \ ,$$
$$R_{2,3}(\theta) = \mathrm{e}^{-x}R_{1,1}(\theta) \ ,$$
and $$R_{2,4}(\theta) = \mathrm{e}^{-x}R_{1,2}(\theta) \ .$$

3.7.2 Proof of Theorem 3.7.2

We start by proving three lemmas.

Lemma 3.7.4 *Let* w *be a non-decreasing function on* $\mathbb{R}_+$*. Let* $r > 0$*. Then*

$$\forall u \geqslant 0, w(u) \leqslant r\big(u^2 \vee \delta^2(w,r)\big) \ ,$$

where $\delta(w,r)$ *is given by Definition 3.7.1.*

Proof If $u > \delta(w,r)$, by Definition 3.7.1,

$$w(u) \leqslant ru^2.$$

If $u \leqslant \delta(w,r)$, since w is non-decreasing, for all $v > \delta(w,r)$,

$$w(u) \leqslant w(v) \leqslant rv^2.$$

By taking the infimum over v, we recover $w(u) \leqslant r\delta(w,r)^2$. ■

Lemma 3.7.5 *Let* w *be a nondecreasing function such that* $x \mapsto \frac{w(x)}{x}$ *is nonincreasing over* $(0;+\infty)$*. Let* $a \in \mathbb{R}_+$ *and* $b \in (0;+\infty)$*. For any* $\theta \in (0;1]$ *and* $u \geqslant 0$*,*

$$\frac{a}{b}w(\sqrt{u}) \leqslant \frac{\theta}{2}\big[u + \delta^2(w,b)\big] + \frac{a^2\delta^2(w,b)}{\theta} \ .$$

Proof Since w is nondecreasing,

$$w(\sqrt{u}) \leqslant w(\sqrt{u+\delta^2(w,b)})$$
$$= \sqrt{u+\delta^2(w,b)}\frac{w(\sqrt{u+\delta^2(w,b)})}{\sqrt{u+\delta^2(w,b)}}.$$

Since $\frac{w(x)}{x}$ is nonincreasing and $\delta(w,b) > 0$,

$$\begin{aligned} w(\sqrt{u}) &\leqslant \sqrt{u+\delta^2(w,b)}\frac{w(\delta(w,b))}{\delta(w,b)} \\ &\leqslant \sqrt{u+\delta^2(w,b)}b\delta(w,b) \text{ by Definition 3.7.1.} \end{aligned}$$

Therefore, using the inequality $\sqrt{ab} \leqslant \frac{\theta}{2}a + \frac{b}{2\theta}$, valid for any $a > 0, b > 0$,

$$\frac{a}{b}w(\sqrt{u}) \leqslant \sqrt{a^2(u+\delta(w,b)^2)\delta(w,b)^2} \leqslant \frac{\theta}{2}(u+\delta(w,b)^2) + \frac{a^2\delta(w,b)^2}{\theta}.$$

■

Lemma 3.7.6 *Let $n_v \in \mathbb{N}^*$. Let $\mathcal{M}$ be a finite set and let $(t_m)_{m\in\mathcal{M}} \in \mathbb{S}^{\mathcal{M}}$. Assume that there exists $p \in [0; 1/|\mathcal{M}|)$ and a function $R : (0; 1] \to \mathbb{R}_+$ such that for any m, m' in $\mathcal{M}$, with probability greater than $1-p$,*

$$\forall \theta \in (0;1], \qquad (P_{n_v} - P)[\gamma(t_m,\cdot) - \gamma(t_{m'},\cdot)] \leqslant \theta\ell(s,t_m) + \theta\ell(s,t_{m'}) + R(\theta) \ .$$

Then for $\widehat{m} \in \operatorname{argmin}_{m\in\mathcal{M}} P_{n_v}\gamma(t_m,\cdot)$, with probability greater than $1 - |\mathcal{M}|p$,

$$\forall \theta \in (0;1], \qquad (1-\theta)\ell(s,t_{\widehat{m}}) \leqslant (1+\theta)\min_{m\in\mathcal{M}} \ell(s,t_m) + R(\theta) \ .$$

Proof Let $m_* \in \operatorname{argmin}_{m\in\mathcal{M}} P\gamma(t_m,\cdot)$. Then for any $m \in \mathcal{M}$, with probability greater than $1-p$,

$$\forall \theta \in (0;1], (P_{n_v} - P)[\gamma(t_{m_*},\cdot) - \gamma(t_m,\cdot)] \leqslant \theta\ell(s,t_{m_*}) + \theta\ell(s,t_m) + R(\theta).$$

So by the union bound, with probability greater than $1 - |\mathcal{M}|p$,

$$\forall \theta \in (0;1], \forall m \in \mathcal{M}, (P_{n_v} - P)[\gamma(t_{m_*},\cdot) - \gamma(t_m,\cdot)] \leqslant \theta\ell(s,t_{m_*}) + \theta\ell(s,t_m) + R(\theta).$$

On that event, for all $\theta \in (0;1]$,

$$\begin{aligned} P\gamma(t_{\widehat{m}},\cdot) &= P_{n_v}\gamma(t_{\widehat{m}},\cdot) + (P - P_{n_v})\gamma(t_{\widehat{m}},\cdot) \\ &\leqslant P_{n_v}\gamma(t_{m_*},\cdot) + (P - P_{n_v})\gamma(t_{\widehat{m}},\cdot) \\ &= P\gamma(t_{m_*},\cdot) + (P - P_{n_v})[\gamma(t_{\widehat{m}},\cdot) - \gamma(t_{m_*},\cdot)] \\ &\leqslant P\gamma(t_{m_*},\cdot) + \theta\ell(s,t_{m_*}) + \theta\ell(s,t_{\widehat{m}}) + R(\theta). \end{aligned}$$

Substracting the Bayes risk $P\gamma(s,\cdot)$ on both sides, we get with probability greater than $1 - |\mathcal{M}|p$, for all $\theta \in (0;1]$,

$$\begin{aligned} \ell(s,t_{\widehat{m}}) &\leqslant \ell(s,t_{m_*}) + \theta\ell(s,t_{m_*}) + \theta\ell(s,t_{\widehat{m}}) + R(\theta), \\ \text{that is, } (1-\theta)\ell(s,t_{\widehat{m}}) &\leqslant (1+\theta)\min_{m\in\mathcal{M}} \ell(s,t_m) + R(\theta). \end{aligned}$$

■

We now prove Theorem 3.7.2. Let $(m, m') \in \mathcal{M}^2$ be fixed. Let

$$\begin{aligned} \sigma &:= w_1(\sqrt{\ell(s,t_m)}) + w_1(\sqrt{\ell(s,t_{m'})}), \\ \text{and} \qquad c &:= w_2(\sqrt{\ell(s,t_m)}) + w_2(\sqrt{\ell(s,t_{m'})}) \ . \end{aligned} \tag{3.13}$$

By hypothesis $H\big(w_1, w_2, (t_m)_{m\in\mathcal{M}}\big)$,

$$\exists c_{m,m'} \text{ such that } \forall k \geqslant 2, P\left(\gamma(t_m,\cdot) - \gamma(t_{m'},\cdot) - c_{m,m'}\right)^k \leqslant k!\sigma^2 c^{k-2} \ . \tag{3.14}$$

For all $y > 0$, let $\Omega_y(m, m')$ be the event on which

$$(P_{n_v} - P)\big[\gamma(t_m,\cdot) - \gamma(t_{m'},\cdot)\big] \leqslant \sqrt{\frac{2y}{n_v}}\sigma + \frac{cy}{n_v} \ . \tag{3.15}$$

By Bernstein's inequality, $\mathbb{P}\big(\Omega_y(m,m')\big) \geqslant 1 - \mathrm{e}^{-y}$.

Let $q = \frac{\theta}{2}\sqrt{\frac{n_v}{x+\log|\mathcal{M}|}}$. By Lemma 3.7.4 with $r = q$,

$$\sigma := w_1(\sqrt{\ell(s,t_m)}) + w_1(\sqrt{\ell(s,t_{m'})}) \leqslant q\left(\ell(s,t_m) \vee \delta^2(w_1,q) + \ell(s,t_{m'}) \vee \delta^2(w_1,q)\right).$$

Set $y = x + \log|\mathcal{M}|$ in (3.15). Then

$$\begin{aligned} \sqrt{\frac{2y}{n_v}}\sigma &:= \sqrt{\frac{2(x+\log|\mathcal{M}|)}{n_v}}\sigma \\ &\leqslant \sqrt{\frac{2(x+\log|\mathcal{M}|)}{n_v}}\frac{\theta}{2}\sqrt{\frac{n_v}{x+\log|\mathcal{M}|}}\left(\ell(s,t_m) \vee \delta^2(w_1,q) + \ell(s,t_{m'}) \vee \delta^2(w_1,q)\right) \\ &\leqslant \frac{\theta}{\sqrt{2}}\left(\ell(s,t_m) + \ell(s,t_{m'}) + 2\delta^2\left(w_1, \frac{\theta}{2}\sqrt{\frac{n_v}{x+\log|\mathcal{M}|}}\right)\right). \end{aligned} \tag{3.16}$$

As for the second term of (3.15), by Lemma 3.7.4 with $r = q^2$, we have

$$c := w_2(\sqrt{\ell(s,t_m)}) + w_2(\sqrt{\ell(s,t_{m'})}) \leqslant q^2\left(\ell(s,t_m) \vee \delta^2(w_2,q^2) + \ell(s,t_{m'}) \vee \delta^2(w_2,q^2)\right).$$

Recall that q is shorthand for $\frac{\theta}{2}\sqrt{\frac{n_v}{x+\log|\mathcal{M}|}}$. Therefore:

$$\begin{aligned} c\frac{y}{n_v} &\leqslant \frac{x+\log|\mathcal{M}|}{n_v}\frac{\theta^2}{4}\frac{n_v}{x+\log|\mathcal{M}|}\left(\ell(s,t_m) \vee \delta^2(w_2,q^2) + \ell(s,t_{m'}) \vee \delta^2(w_2,q^2)\right) \\ &= \frac{\theta^2}{4}\left(\ell(s,t_m) \vee \delta^2(w_2,q^2) + \ell(s,t_{m'}) \vee \delta^2(w_2,q^2)\right) \\ &\leqslant \frac{\theta^2}{4}\left(\ell(s,t_m) + \ell(s,t_{m'}) + 2\delta^2\left(w_2, \frac{\theta^2}{4}\frac{n_v}{x+\log|\mathcal{M}|}\right)\right). \end{aligned} \tag{3.17}$$

Since $\sqrt{\frac{1}{2}} + \frac{1}{4} \leqslant 1$ and $\theta \in (0;1]$, plugging (3.16) and (3.17) in (3.15) yields, on the event $\Omega_{x+\log|\mathcal{M}|}(m,m')$, for all $\theta \in (0;1]$,

$$(P_{n_v} - P)[\gamma(t_m,\cdot) - \gamma(t_{m'},\cdot)] \leqslant \theta\big(\ell(s,t_m) + \ell(s,t_{m'})\big) + \sqrt{2}\theta\delta^2\left(w_1, \frac{\theta}{2}\sqrt{\frac{n_v}{x+\log|\mathcal{M}|}}\right) + \frac{\theta^2}{2}\delta^2\left(w_2, \frac{\theta^2}{4}\frac{n_v}{x+\log|\mathcal{M}|}\right). \tag{3.18}$$

Suppose now that $x \mapsto \frac{w_j(x)}{x}$ is nonincreasing for $j \in \{1;2\}$. Let $\theta \in [0;1]$. Let $y \geqslant 0$. By Lemma 3.7.5 with $a = \sqrt{2y}$ and $b = \sqrt{n_v}$,

$$\begin{aligned}\sqrt{\frac{2y}{n_v}}\sigma &= \sqrt{\frac{2y}{n_v}}\left(w_1(\sqrt{\ell(s,t_m)}) + w_1(\sqrt{\ell(s,t_{m'})})\right)\\ &\leqslant \frac{\theta}{2}\ell(s,t_m) + \frac{\theta}{2}\ell(s,t_{m'}) + \delta^2(w_1,\sqrt{n_v})\left[\theta + \frac{2y}{\theta}\right]\ .\end{aligned} \tag{3.19}$$

By Lemma 3.7.5 with $a = y$ and $b = n_v$,

$$\begin{aligned}c\frac{y}{n_v} &= \frac{y}{n_v}\left(w_2(\sqrt{\ell(s,t_m)}) + w_2(\sqrt{\ell(s,t_{m'})})\right)\\ &\leqslant \frac{\theta}{2}\ell(s,t_m) + \frac{\theta}{2}\ell(s,t_{m'}) + \delta^2(w_2,n_v)\left[\theta + \frac{y^2}{\theta}\right].\end{aligned} \tag{3.20}$$

Plugging (3.19) and (3.20) in (3.15) yields, on the event $\Omega_y(m,m')$, for all $\theta \in (0;1]$,

$$\begin{aligned}&(P_{n_v} - P)[\gamma(t_m,\cdot) - \gamma(t_{m'},\cdot)]\\ &\leqslant \theta\ell(s,t_m) + \theta\ell(s,t_{m'}) + \delta^2(w_1,\sqrt{n_v})\left[\theta + \frac{2y}{\theta}\right] + \delta^2(w_2,n_v)\left[\theta + \frac{y^2}{\theta}\right]\ .\end{aligned} \tag{3.21}$$

By (3.18), Lemma 3.7.6 applies with $p = \frac{\mathrm{e}^{-x}}{|\mathcal{M}|}$ and

$$R(\theta) = \sqrt{2}\theta\delta^2\left(w_1, \frac{\theta}{2}\sqrt{\frac{n_v}{x+\log|\mathcal{M}|}}\right) + \frac{\theta^2}{2}\delta^2\left(w_2, \frac{\theta^2}{4}\frac{n_v}{x+\log|\mathcal{M}|}\right).$$

This yields (3.8). By (3.21), Lemma 3.7.6 applies with $p = \mathrm{e}^{-y}$ and

$$R(\theta) = \theta\left[\delta_1^2 + \delta_2^2\right] + \frac{1}{\theta}\left[2y\delta_1^2 + y^2\delta_2^2\right].$$

Setting $y = \log|\mathcal{M}| + x$ yields (3.10). ■

3.7.3 Proof of Theorem 3.7.3

We start by proving two lemmas.

Lemma 3.7.7 *Let $f \in L^1(\mathbb{R}_+, \mathrm{e}^{-x}\mathrm{d}x)$ be a non-negative, non-decreasing function such that $\lim_{x\to+\infty} f(x) = +\infty$. Let X be a random variable such that*

$$\forall x \in \mathbb{R}_+, \mathbb{P}\big(X > f(x)\big) \leqslant \mathrm{e}^{-x} \ .$$

Then

$$\mathbb{E}[X] \leqslant \int_0^{+\infty} f(x)\mathrm{e}^{-x}\mathrm{d}x \ .$$

Proof Let $g \in L^1(\mathbb{R}_+, \mathrm{e}^{-x}\mathrm{d}x)$ be a non-decreasing, differentiable function such that $g \geqslant f$. Then

$$\begin{aligned}
\mathbb{E}[X] &\leqslant \int_0^{+\infty} \mathbb{P}[X > t]\mathrm{d}t \\
&= \int_0^{g(0)} \mathbb{P}[X > t]\mathrm{d}t + \int_0^{+\infty} \mathbb{P}[X > g(x)]g'(x)\mathrm{d}x \\
&\leqslant g(0) + \int_0^{+\infty} \mathrm{e}^{-x}g'(x)\mathrm{d}x \quad \text{since } g \geqslant f \\
&= g(0) + [\mathrm{e}^{-x}g(x)]_0^{\infty} + \int_0^{+\infty} \mathrm{e}^{-x}g(x)\mathrm{d}x \\
&= \int_0^{+\infty} \mathrm{e}^{-x}g(x)\mathrm{d}x \ .
\end{aligned}$$

It remains to show that g can approximate f in $L^1(\mathbb{I}_{x\geqslant 0}\mathrm{e}^{-x}\mathrm{d}x)$. Let K be a nonnegative smooth function vanishing outside $[-1;1]$, normalized such that $\int K(t)\mathrm{d}t = 1$. Let $\varepsilon > 0$. Define

$$f_\varepsilon(x) = \frac{1}{\varepsilon}\int f(t)K\left(\frac{x+\varepsilon-t}{\varepsilon}\right)\mathrm{d}t \tag{3.22}$$

$$= \frac{1}{\varepsilon}\int f(x+\varepsilon-t)K\left(\frac{t}{\varepsilon}\right)\mathrm{d}t \tag{3.23}$$

By (3.22), f_ε is smooth. By (3.23), f_ε is nondecreasing, moreover

$$\begin{aligned}
f_\varepsilon(x) - f(x) &= \frac{1}{\varepsilon}\int \big[f(x+\varepsilon-t) - f(x)\big]K\left(\frac{t}{\varepsilon}\right)\mathrm{d}t \text{ since } \int K = 1 \\
&= \frac{1}{\varepsilon}\int_{-\varepsilon}^{\varepsilon} \big[f(x+\varepsilon-t) - f(x)\big]K\left(\frac{t}{\varepsilon}\right)\mathrm{d}t \text{ since } K(u) = 0 \text{ when } |u| \geqslant 1 \\
&\geqslant 0 \text{ since } f \text{ is nondecreasing and } K \geqslant 0 \ .
\end{aligned}$$

Thus $f_\varepsilon \geqslant f$. Finally, by Jensen's inequality and Fubini's theorem,

$$\begin{aligned}\int |f_\varepsilon(x) - f(x)|\mathrm{e}^{-x}\mathrm{d}x &\leqslant \frac{1}{\varepsilon}\int_{-\varepsilon}^{\varepsilon} K\left(\frac{t}{\varepsilon}\right)\int |f(x+\varepsilon-t) - f(x)|\mathrm{e}^{-x}\mathrm{d}x \\ &\leqslant \sup_{|\tau|\leqslant 2\varepsilon}\int |f(x+\tau) - f(x)|\mathrm{e}^{-x}\mathrm{d}x \ ,\end{aligned}$$

which converges to 0 when $\varepsilon \to 0$ since $f \in L^1(\mathbb{R}_+, \mathrm{e}^{-x}\mathrm{d}x)$. ■

We use the following additional notation:

Definition 3.7.8 *Let g be the function defined by*

$$\forall(\theta, y, p, q) \in (0;1] \times \mathbb{R}_+^3, \qquad g(\theta, y, p, q) = \theta[p+q] + \frac{1}{\theta}\left[2yp + y^2 q\right] \ .$$

This function satisfies the following properties.

Lemma 3.7.9 *Let g be the function given in Definition 3.7.8. For any $\theta \in [0;1]$ and any $u > 0, p \geqslant 0, q \geqslant 0$,*

$$\mathrm{e}^u \int_u^{+\infty} g(\theta, y, p, q)\mathrm{e}^{-y}\mathrm{d}y = \left(\theta + \frac{2(1+u)}{\theta}\right)p + \left(\theta + \frac{2+2u+u^2}{\theta}\right)q \ .$$

Proof of Lemma 3.7.9

Using the formulas

$$\begin{aligned}&\int_u^{+\infty} \mathrm{e}^{-x}\mathrm{d}x = \mathrm{e}^{-u}, \int_u^{+\infty} x\mathrm{e}^{-x}\mathrm{d}x = (1+u)\mathrm{e}^{-u}, \\ &\int_u^{+\infty} x^2\mathrm{e}^{-x}\mathrm{d}x = (u^2+2u+2)\mathrm{e}^{-u} \ ,\end{aligned}$$

we get:

$$\begin{aligned}\mathrm{e}^u \int_u^{+\infty} g(\theta, y, p, q)\mathrm{e}^{-y}\mathrm{d}y &= \theta[p+q] + \frac{2}{\theta}(1+u)p + (u^2+2u+2)\frac{q}{\theta} \\ &= \left(\theta + \frac{2(1+u)}{\theta}\right)p + \left(\theta + \frac{2+2u+u^2}{\theta}\right)q \ .\end{aligned}$$

■

We can now proceed with the proof of Theorem 3.7.3. Let $\theta \in (0;1]$ be fixed. Let $(\widehat{f}_T^{\text{ho}})_{T\in\mathcal{T}}$ be the individual hold out estimators, so that $\widehat{f}_{\mathcal{T}}^{\text{ag}} = \frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}} \widehat{f}_T^{\text{ho}}$. By convexity of the risk functional $\mathcal{L}$, we have

$$\mathcal{L}(\widehat{f}_{\mathcal{T}}^{\text{ag}}) \leqslant \frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}} \mathcal{L}(\widehat{f}_T^{\text{ho}}) \ .$$

It follows by substracting $\mathcal{L}(s)$ that:

$$\ell(s, \widehat{f}_{\mathcal{T}}^{\text{ag}}) \leqslant \frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}} \ell(s, \widehat{f}_T^{\text{ho}}) \ .$$

Since the data are i.i.d, by assumption (3.2), all $\widehat{f}_T^{\text{ho}}$ have the same distribution. Let $T_1 = \{1, \dots, n_t\}$, so that $D_n^{T_1} = D_{n_t}$. Taking expectations yields

$$\mathbb{E}[\ell(s, \widehat{f}_{\mathcal{T}}^{\text{ag}})] \leqslant \mathbb{E}[\ell(s, \widehat{f}_{T_1}^{\text{ho}})] \ . \tag{3.24}$$

Since $H\left(\widehat{w}_{1,1}, \widehat{w}_{1,2}, (\mathcal{A}_m(D_{n_t})_{m\in\mathcal{M}})\right)$ holds, we can apply Theorem 3.7.2 conditionally on D_{n_t}, with $t_m = \mathcal{A}_m(D_{n_t})$.

Proof of (3.11) For $i \in \{1;2\}$, let $\widehat{\delta}_{1,i} = \delta(\widehat{w}_{1,i}, \sqrt{n_v}^i)$. Let g be given in Definition 3.7.8. By Theorem 3.7.2, Equation (3.10), for any $z \geqslant 0$, with probability greater than $1 - \mathrm{e}^{-z}$,

$$(1-\theta)\ell(s, \widehat{f}_{T_1}^{\text{ho}}) \leqslant (1+\theta)\min_{m\in\mathcal{M}} \ell(s, t_m) + g\big(\theta, z + \log|\mathcal{M}|, \widehat{\delta}_{1,1}^2, \widehat{\delta}_{1,2}^2\big) \ . \tag{3.25}$$

As g is nondecreasing in its second variable, Lemma 3.7.7 applied to the random variable $(1-\theta)\ell(s, \widehat{f}_{T_1}^{\text{ho}})$ yields:

$$(1-\theta)\mathbb{E}\left[\ell(s, \widehat{f}_{T_1}^{\text{ho}})\middle| D_n^{T_1}\right] \leqslant (1+\theta)\min_{m\in\mathcal{M}} \ell(s, t_m) + \int_{\log|\mathcal{M}|}^{+\infty} g\big(\theta, y, \widehat{\delta}_{1,1}^2, \widehat{\delta}_{1,2}^2\big)\mathrm{e}^{-(y-\log|\mathcal{M}|)}\mathrm{d}y \ .$$

Lemma 3.7.9 yields

$$\begin{aligned}(1-\theta)\mathbb{E}\left[\ell(s, \widehat{f}_{T_1}^{\text{ho}})\middle| D_n^{T_1}\right] \leqslant (1+\theta)\min_{m\in\mathcal{M}} \ell(s, t_m) &+ \left(\theta + \frac{2\left(1+\log|\mathcal{M}|\right)}{\theta}\right)\widehat{\delta}_{1,1}^2 \\ &+ \left(\theta + \frac{2\left(1+\log|\mathcal{M}|\right) + \log^2|\mathcal{M}|}{\theta}\right)\widehat{\delta}_{1,2}^2 \ .\end{aligned}$$

Taking expectations with respect to $D_n^{T_1} = D_{n_t}$,

$$\begin{aligned}(1-\theta)\mathbb{E}\left[\ell(s, \widehat{f}_{T_1}^{\text{ho}})\right] \leqslant (1+\theta)\mathbb{E}\big[\min_{m\in\mathcal{M}} \ell(s, \mathcal{A}_m(D_{n_t}))\big] &+ \left(\theta + \frac{2\left(1+\log|\mathcal{M}|\right)}{\theta}\right)\mathbb{E}\big[\widehat{\delta}_{1,1}^2\big] \\ &+ \left(\theta + \frac{2\left(1+\log|\mathcal{M}|\right) + \log^2|\mathcal{M}|}{\theta}\right)\mathbb{E}\left[\widehat{\delta}_{1,2}^2\right] \ .\end{aligned}$$

Equation (3.11) then follows from Equation (3.24).

Proof of (3.12) Fix $x \geqslant 0$. For $i \in \{1;2\}$, let $\widehat{\delta}_{2,i} = \delta\left(\widehat{w}_{2,i}, \left(\frac{\theta}{2}\sqrt{\frac{n_v}{x+\log|\mathcal{M}|}}\right)^i\right)$.

By Theorem 3.7.2, Equation (3.8), with probability larger than $1-\mathrm{e}^{-x}$,

$$(1-\theta)\ell(s,\widehat{f}^{\mathrm{ho}}_{T_1}) \leqslant (1+\theta)\min_{m\in\mathcal{M}}\ell(s,t_m) + \sqrt{2}\theta\widehat{\delta}^2_{2,1} + \frac{\theta^2}{2}\widehat{\delta}^2_{2,2} \ . \tag{3.26}$$

Combining (3.25) and (3.26), for any $z \geqslant 0$, with probability larger than $1-\mathrm{e}^{-z}$,

$$(1-\theta)\ell(s,\widehat{f}^{\mathrm{ho}}_{T_1}) \leqslant (1+\theta)\min_{m\in\mathcal{M}}\ell(s,t_m) + \sqrt{2}\theta\widehat{\delta}^2_{2,1} + \frac{\theta^2}{2}\widehat{\delta}^2_{2,2} + \mathbb{I}_{z\geqslant x}g\big(\theta, z+\log|\mathcal{M}|, \widehat{\delta}^2_{1,1}, \widehat{\delta}^2_{1,2}\big) \ .$$

By Lemma 3.7.7,

$$\begin{aligned}(1-\theta)\mathbb{E}\left[\ell(s,\widehat{f}^{\mathrm{ho}}_{T_1})\middle|D_n^{T_1}\right] \leqslant (1+\theta)\min_{m\in\mathcal{M}}\ell(s,t_m) + \sqrt{2}\theta\widehat{\delta}^2_{2,1} + \frac{\theta^2}{2}\widehat{\delta}^2_{2,2} \\ + \int_{x+\log|\mathcal{M}|}^{+\infty} g\big(\theta, y, \widehat{\delta}^2_{1,1}, \widehat{\delta}^2_{1,2}\big)\mathrm{e}^{-(y-\log|\mathcal{M}|)}\mathrm{d}y \ .\end{aligned}$$

By Lemma 3.7.9, it follows that

$$\begin{aligned}(1-\theta)\mathbb{E}\left[\ell(s,\widehat{f}^{\mathrm{ho}}_{T_1})\middle|D_n^{T_1}\right] \leqslant (1+\theta)\min_{m\in\mathcal{M}}\ell(s,t_m) + \sqrt{2}\theta\widehat{\delta}^2_{2,1} + \frac{\theta^2}{2}\widehat{\delta}^2_{2,2} \\ + \mathrm{e}^{-x}\left(\theta + \frac{2(1+x+\log|\mathcal{M}|)}{\theta}\right)\widehat{\delta}^2_{1,1} \\ + \mathrm{e}^{-x}\left(\theta + \frac{2(1+x+\log|\mathcal{M}|) + (x+\log|\mathcal{M}|)^2}{\theta}\right)\widehat{\delta}^2_{1,2} \ .\end{aligned}$$

Taking expectations with respect to $D_n^{T_1}$ and using inequality (3.24) yields Equation (3.12) of Theorem 3.7.3. ■

3.8 RKHS regression: proof of Theorem 3.4.3

In the following, for any $g : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ and $t : \mathcal{X} \to \mathbb{R}$, the function $(x,y) \mapsto g(t(x),y)$ is denoted by $g \circ t$.

3.8.1 Preliminary results

Remark first that the RKHS norm dominates the supremum norm:

Lemma 3.8.1 *If $\kappa = \sup_x K(x,x) < +\infty$ then for any $t \in \mathcal{H}$,*

$$\|t\|_\infty \leqslant \sqrt{\kappa}\, \|t\|_{\mathcal{H}} \ .$$

Proof By definition of an RKHS, $\forall t \in \mathcal{H}, \forall x \in \mathcal{X}, \langle t, K(x,\cdot)\rangle_{\mathcal{H}} = t(x)$. It follows that, for any $t \in \mathcal{H}$,

$$\begin{aligned}\|t\|_\infty^2 = \sup_x t(x)^2 &= \sup_x \langle t, K(x,\cdot)\rangle_{\mathcal{H}}^2 \\ &\leqslant \|t\|_{\mathcal{H}}^2 \sup_x \langle K(x,\cdot), K(x,\cdot)\rangle \\ &\leqslant \|t\|_{\mathcal{H}}^2 \sup_x K(x,x).\end{aligned}$$

■

Using standard arguments, the following deviation inequality can be derived.

Proposition 3.8.2 *Let $\mathcal{H}$ denote a RKHS with bounded kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $\kappa = \sup_x K(x,x)$ and $h : \mathbb{R}^2 \to \mathbb{R}$ be Lipschitz in its first argument with Lipschitz constant L. For any $t \in \mathcal{H}$ and $r > 0$, denote*

$$B_{\mathcal{H}}(t,r) = \left\{ t' \in \mathcal{H} \mid \|t' - t\|_{\mathcal{H}} \leqslant r \right\} \ .$$

Let $t_0 \in \mathcal{H}$. Then for any probability measure P on $\mathcal{X} \times \mathbb{R}$ and any $y > 0$,

$$P^{\otimes n}\left[\sup_{(t_1,t_2) \in B_{\mathcal{H}}(t_0,r)^2} (P_n - P)\big(h \circ t_1 - h \circ t_2\big) \geqslant 2(2 + \sqrt{2y}) L \frac{r\sqrt{\kappa}}{\sqrt{n}} \right] \leqslant \mathrm{e}^{-y} \ .$$

Proof Let $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n}$ be a dataset drawn from P. Let $(\sigma_i)_{1 \leqslant i \leqslant n}$ be i.i.d Rademacher variables independent from D_n. Denote by $R_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i)\right]$ the Rademacher complexity of a class $\mathcal{F}$ of real valued functions.

By Lemma 3.8.1, for any $(t_1, t_2) \in B_{\mathcal{H}}(t_0, r)^2$,

$$\|h \circ t_1 - h \circ t_2\|_\infty \leqslant L \|t_1 - t_2\|_\infty \leqslant L\left[\|t_1 - t_0\|_\infty + \|t_2 - t_0\|_\infty\right] \leqslant 2L\sqrt{\kappa} r \ .$$

By symmetry under exchange of t_1 and t_2, notice that

$$R_n\left(\{h \circ t_1 - h \circ t_2 | (t_1,t_2) \in B_{\mathcal{H}}(t_0,r)^2\}\right) = \sup_{(t_1,t_2) \in B_{\mathcal{H}}(t_0,r)^2} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i (h \circ t_1 - h \circ t_2)(X_i) \right| \ .$$

By the bounded difference inequality and [20], Theorem 3.2, it follows that for any $y > 0$, with probability greater than $1 - \mathrm{e}^{-y}$,

$$\sup_{(t_1,t_2) \in B_{\mathcal{H}}(t_0,r)^2} (P_n - P)(h \circ t_1 - h \circ t_2) \leqslant 2R_n\left(\{h \circ t_1 - h \circ t_2 | (t_1,t_2) \in B_{\mathcal{H}}(t_0,r)^2\}\right) + 2Lr\sqrt{\frac{2\kappa y}{n}}.$$

Moreover,

$$
\begin{aligned}
&R_n\left(\{h\circ t_1 - h\circ t_2 | (t_1,t_2)\in B_{\mathcal{H}}(t_0,r)^2\}\right)\\
&\leqslant R_n(\{h\circ t | t\in B_{\mathcal{H}}(t_0,r)\}) + R_n(\{-h\circ t | t\in B_{\mathcal{H}}(t_0,r)\})\\
&\leqslant 2LR_n(B_{\mathcal{H}}(t_0,r)) \text{by the contraction lemma (relevant version: [80], Theorem 7),}\\
&= 2LR_n(B_{\mathcal{H}}(0,r)) \text{ (by translation invariance).}
\end{aligned}
$$

Finally, by a classical computation (see for example [20], Section 4.1.2),

$$
\begin{aligned}
&R_n\left(\{h\circ t_1 - h\circ t_2 | (t_1,t_2)\in B_{\mathcal{H}}(t_0,r)^2\}\right)\\
&\leqslant 2L\frac{r}{n}\mathbb{E}\sqrt{\sum_{i=1}^{n} K(X_i,X_i)}\\
&\leqslant 2Lr\sqrt{\frac{\kappa}{n}} \ .
\end{aligned}
$$

■

The proof of Theorem 3.4.3 also uses the following peeling lemma.

Lemma 3.8.3 *Let $(Z_u)_{u\in T}$ be a stochastic process and $d: T\to\mathbb{R}_+$ be a function. Let $a\geqslant 0$ and $b\in(0;2]$ and assume that*

$$
\forall r,y\geqslant 0, \mathbb{P}\left[\sup_{u\in T: d(u)\leqslant r} Z_u \geqslant r\frac{1+\sqrt{b(a+y)}}{\sqrt{n}}\right]\leqslant \mathrm{e}^{-y} \ . \tag{3.27}
$$

Then, for any $\theta\in(0;+\infty)$,

$$
\mathbb{P}\left[\exists u\in T, Z_u\geqslant \theta d^2(u) + \frac{2+b\big[1.1+2(a+y)\big]}{\theta n}\right]\leqslant \mathrm{e}^{-y} \ .
$$

Proof Let $x>0$. Let $\eta\in(1;2]$, $j_m\in\mathbb{N}^*$ and $y_0\in\mathbb{R}$ be absolute constants that

will be determined later. Then

$$
\begin{aligned}
&\mathbb{I}\left\{\sup_{u\in T}\frac{Z_u}{d^2(u)+x^2}\geqslant\frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right\}\\
&\leqslant\mathbb{I}\left\{\sup_{u\in T:d(u)\leqslant x}\frac{Z_u}{d^2(u)+x^2}\geqslant\frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right\}\\
&+\sum_{j=0}^{+\infty}\mathbb{I}\left\{\sup_{u\in T:\eta^j x\leqslant d(u)\leqslant\eta^{j+1}x}\frac{Z_u}{d^2(u)+x^2}\geqslant\frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right\}\\
&\leqslant\mathbb{I}\left\{\sup_{u\in T:d(u)\leqslant x}\frac{Z_u}{x^2}\geqslant\frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right\}\\
&+\sum_{j=0}^{+\infty}\mathbb{I}\left\{\sup_{u\in T:\eta^j x\leqslant d(u)\leqslant\eta^{j+1}x}\frac{Z_u}{(1+\eta^{2j})x^2}\geqslant\frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right\}\\
&\leqslant\mathbb{I}\left\{\sup_{u\in T:d(u)\leqslant x}Z_u\geqslant\frac{x(1+\sqrt{b(a+y)})}{\sqrt{n}}\right\}\\
&+\sum_{j=0}^{+\infty}\mathbb{I}\left\{\sup_{u\in T:d(u)\leqslant\eta^{j+1}x}Z_u\geqslant(1+\eta^{2j})\frac{x(1+\sqrt{b(a+y)})}{\sqrt{n}}\right\}\quad. \qquad (3.28)
\end{aligned}
$$

Notice that:

$$
\begin{aligned}
(1+\eta^{2j})\frac{x(1+\sqrt{b(a+y)})}{\sqrt{n}}&=x\eta^{j+1}\frac{\eta^{2j}+1}{\eta^{j+1}}\frac{1+\sqrt{b(a+y)}}{\sqrt{n}}\\
&=x\eta^{j+1}\frac{1+\sqrt{b(a+z_j)}}{\sqrt{n}}\quad,
\end{aligned}
$$

where:

$$
\begin{aligned}
z_j&=\frac{1}{b}\left(\frac{\eta^{2j}+1}{\eta^{j+1}}-1+\frac{\eta^{2j}+1}{\eta^{j+1}}\sqrt{b(a+y)}\right)^2-a\\
&\geqslant\frac{1}{b}\left[\frac{\eta^{2j}+1}{\eta^{j+1}}-1\right]^2+\left(\frac{\eta^{2j}+1}{\eta^{j+1}}\right)^2 y\qquad\text{since } a\geqslant 0 \text{ and } \eta^{2j}+1\geqslant\eta^{j+1}\quad.
\end{aligned}
$$

Taking expectations in (3.28) and using hypothesis (3.27), we obtain:

$$
\mathbb{P}\left[\sup_{u\in T}\frac{Z_u}{d^2(u)+x^2}\geqslant\frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right]\leqslant\mathrm{e}^{-y}+\sum_{j=0}^{+\infty}\mathrm{e}^{-z_j}\quad.
$$

So for any $y \geqslant y_0$,

$$\begin{aligned}
&\mathbb{P}\left[\sup_{u\in T}\frac{Z_u}{d^2(u)+x^2} \geqslant \frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right] \\
&\leqslant \mathrm{e}^{-y} + \mathrm{e}^{-y}\sum_{j=0}^{+\infty}\exp\left(-\frac{1}{b}\left[\frac{\eta^{2j}+1}{\eta^{j+1}}-1\right]^2 - \left(\frac{(\eta^{2j}+1)^2}{(\eta^{j+1})^2}-1\right)y\right) \\
&\leqslant \mathrm{e}^{-y} + \mathrm{e}^{-y}\sum_{j=0}^{+\infty}\exp\left(-\frac{1}{b}\left[\frac{\eta^{2j}+1}{\eta^{j+1}}-1\right]^2 - \left(\frac{(\eta^{2j}+1)^2}{(\eta^{j+1})^2}-1\right)y_0\right) \quad . \quad (3.29)
\end{aligned}$$

Now, we have

$$\begin{aligned}
\exp\left(-\frac{1}{b}\left[\frac{\eta^{2j}+1}{\eta^{j+1}}-1\right]^2 - \left(\frac{(\eta^{2j}+1)^2}{(\eta^{j+1})^2}-1\right)y_0\right) &\leqslant \exp\left(-\left(\frac{(\eta^{2j}+1)^2}{(\eta^{j+1})^2}-1\right)y_0\right) \\
&\leqslant \exp\left(y_0 - \eta^{2(j-1)}y_0\right) \quad . \quad (3.30)
\end{aligned}$$

Let u denote the sequence $u_j = \exp\left(y_0 - \eta^{2(j-1)}y_0\right)$. Then for $j \geqslant j_m$,

$$\begin{aligned}
\log u_{j+1} - \log u_j &= \eta^{2(j-1)}y_0 - \eta^{2j}y_0 \\
&= y_0(1-\eta^2)\eta^{2(j-1)} \\
&\leqslant y_0(1-\eta^2)\eta^{2(j_m-1)} \text{ since } \eta > 1 \quad .
\end{aligned}$$

Thus,

$$\forall j \geqslant j_m, \quad u_{j+1} \leqslant u_j \exp\left(-y_0(\eta^2-1)\eta^{2(j_m-1)}\right) \quad .$$

Therefore, we have

$$\forall j \geqslant 0, \quad u_{j+j_m} \leqslant u_{j_m}\exp\left(-jy_0(\eta^2-1)\eta^{2(j_m-1))}\right)$$

and

$$\sum_{j=j_m}^{+\infty} u_j \leqslant u_{j_m}\left[1-\exp\left(-y_0(\eta^2-1)\eta^{2(j_m-1)}\right)\right]^{-1} \quad .$$

It follows from (3.29) and (3.30) that for any $y \geqslant y_0$, since $b \leqslant 2$,

$$\begin{aligned}
&\mathrm{e}^y\mathbb{P}\left[\sup_{u}\frac{Z_u}{d^2(u)+x^2} \geqslant \frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right] \\
&\leqslant 1 + \sum_{j=0}^{j_m}\exp\left(-\frac{1}{2}\left[\frac{\eta^{2j}+1}{\eta^{j+1}}-1\right]^2 - \left(\frac{(\eta^{2j}+1)^2}{(\eta^{j+1})^2}-1\right)y_0\right) \\
&+ \frac{\exp\left(y_0 - \eta^{2(j_m-1)}y_0\right)}{1-\exp\left(-y_0(\eta^2-1)\eta^{2(j_m-1)}\right)} \quad . \quad (3.31)
\end{aligned}$$

On the other hand, when $y \leqslant y_0$, trivially,

$$\mathbb{P}\left[\sup_u \frac{Z_u}{d^2(u)+x^2} \geqslant \frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right] \leqslant 1 \leqslant \mathrm{e}^{y_0}\mathrm{e}^{-y}.$$

Taking $\eta = 1.18, j_m = 10, y_0 = 0.52$, the right-hand side of (3.31) evaluates to $1.6765 < 1.7$ whereas $\mathrm{e}^{y_0} \leqslant 1.683 < 1.7$. It follows that for all $y > 0$,

$$\mathbb{P}\left[\sup_u \frac{Z_u}{d^2(u)+x^2} \geqslant \frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right] \leqslant 1.7\mathrm{e}^{-y} \ . \tag{3.32}$$

Now take $x = \frac{1+\sqrt{b(a+y)}}{\theta\sqrt{n}}$ with $\theta > 0$. We can rewrite:

$$\begin{aligned}
\mathbb{P}\left[\sup_u \frac{Z_u}{d^2(u)+x^2} \geqslant \frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right] &= \mathbb{P}\left[\exists u \in T, \frac{Z_u}{d^2(u)+x^2} \geqslant \theta\right] \\
&= \mathbb{P}\left[\exists u \in T, Z_u \geqslant \theta d^2(u) + \frac{1}{\theta n}\left(1+\sqrt{b(a+y)}\right)^2\right] \\
&\geqslant \mathbb{P}\left[\exists u \in T, Z_u \geqslant \theta d^2(u) + \frac{2+2b(a+y)}{\theta n}\right] \ .
\end{aligned}$$

It follows from Equation (3.32), with y replaced by $y + 0.55$, that

$$\begin{aligned}
\mathbb{P}\left[\exists u \in T, Z_u \geqslant \theta d^2(u) + \frac{2+b(1.1+2(a+y))}{\theta n}\right] &\leqslant 1.7\mathrm{e}^{-0.55}\mathrm{e}^{-y} \\
&\leqslant \mathrm{e}^{-y} \ .
\end{aligned}$$

■

We need two other technical lemmas in the proof of Theorem 3.4.3.

Lemma 3.8.4 *For any nonnegative, continuous convex function h over a Hilbert space $\mathcal{H}$, and any $\lambda \in \mathbb{R}_+$, the elements of the regularization path,*

$$t_\lambda = \operatorname*{argmin}_{t\in\mathcal{H}} \left\{h(t) + \lambda \|t\|_{\mathcal{H}}^2\right\} \ ,$$

satisfy, for any $(\lambda, \mu) \in \mathbb{R}^2$ such that $0 < \lambda \leqslant \mu$,

$$\|t_\lambda - t_\mu\|_{\mathcal{H}}^2 \leqslant \|t_\lambda\|_{\mathcal{H}}^2 - \|t_\mu\|_{\mathcal{H}}^2 \ .$$

Proof By [9, Theorem 2.11], t_λ exists for any $\lambda \in \mathbb{R}_+$. Moreover, it is unique by strong convexity of $\|\cdot\|_{\mathcal{H}}^2$. For a closed convex set $\mathcal{C} \subset \mathcal{H}$, let $\Pi_{\mathcal{C}}$ denote the orthogonal projection onto $\mathcal{C}$.

Let $\mu > 0$. The set $\{t : h(t) \leqslant h(t_\mu)\}$ is closed by continuity of h and convex by convexity of h. Moreover, for any $t \in \mathcal{H}$ such that $h(t) \leqslant h(t_\mu)$,

$$\begin{aligned}\mu \|t_\mu\|_{\mathcal{H}}^2 &\leqslant h(t_\mu) - h(t) + \mu \|t_\mu\|_{\mathcal{H}}^2 \\ &\leqslant \mu \|t\|_{\mathcal{H}}^2 \text{ by definition of } t_\mu \ .\end{aligned}$$

Therefore, $t_\mu = \Pi_{\{t:h(t)\leqslant h(t_\mu)\}}(0)$. Let $\lambda \in (0;\mu)$. By definition of t_λ, t_μ,

$$\begin{aligned}\frac{h(t_\mu)}{\mu} + \|t_\mu\|_{\mathcal{H}}^2 &\leqslant \frac{h(t_\lambda)}{\mu} + \|t_\lambda\|_{\mathcal{H}}^2 \\ &= \frac{h(t_\lambda)}{\lambda} + \|t_\lambda\|_{\mathcal{H}}^2 + \left(\frac{1}{\mu} - \frac{1}{\lambda}\right) h(t_\lambda) \\ &\leqslant \frac{h(t_\mu)}{\lambda} + \|t_\mu\|_{\mathcal{H}}^2 + \left(\frac{1}{\mu} - \frac{1}{\lambda}\right) h(t_\lambda) \ ,\end{aligned}$$

which implies $(\mu^{-1} - \lambda^{-1})h(t_\mu) \leqslant (\mu^{-1} - \lambda^{-1})h(t_\lambda)$ and thus $h(t_\lambda) \leqslant h(t_\mu)$ since $\lambda < \mu$. For a projection $\Pi_{\mathcal{C}}$, it is well known that:

$$\forall t \in \mathcal{H}, \forall t' \in \mathcal{C}, \langle t - \Pi_{\mathcal{C}}(t), \Pi_{\mathcal{C}}(t) - t' \rangle_{\mathcal{H}} \geqslant 0 \ .$$

Choosing $\mathcal{C} = \{t : h(t) \leqslant h(t_\mu)\}, t' = t_\lambda \in \mathcal{C}, t = 0$ yields $\langle -t_\mu, t_\mu - t_\lambda \rangle_{\mathcal{H}} \geqslant 0$. Therefore

$$\begin{aligned}\|t_\lambda\|_{\mathcal{H}}^2 &= \|t_\mu + (t_\lambda - t_\mu)\|_{\mathcal{H}}^2 \\ &= \|t_\mu\|_{\mathcal{H}}^2 + \|t_\lambda - t_\mu\|_{\mathcal{H}}^2 + 2\langle t_\mu, t_\lambda - t_\mu \rangle_{\mathcal{H}} \\ &\geqslant \|t_\mu\|_{\mathcal{H}}^2 + \|t_\lambda - t_\mu\|_{\mathcal{H}}^2 \ .\end{aligned}$$

■

Lemma 3.8.5 *Let $(b,c) \in \mathbb{R}_+^2$ and $l_{b,c}(x) = bx + c$. Let δ be given by Definition 3.7.1. For any $r \in \mathbb{R}_+$,*

$$\delta^2(l_{b,c}, r) \leqslant \frac{b^2}{r^2} + \frac{2c}{r} \ . \tag{3.33}$$

For $(a,b,c) \in \mathbb{R}_+^3$, let $g_{a,b,c}(x) = ax \vee [bx^3 + cx^2]^{\frac{1}{2}}$. For any $r \in \mathbb{R}_+$,

$$\delta^2\big(g_{a,b,c}, r\big) \leqslant \frac{a^2}{r^2} \vee \left[\frac{b^2}{r^4} + \frac{2c}{r^2}\right] \leqslant \frac{a^2}{r^2} + \frac{b^2}{r^4} + \frac{2c}{r^2} \ . \tag{3.34}$$

Proof Since $x \mapsto \frac{l_{b,c}(x)}{x}$ is nonincreasing, we have by Remark 3.7.1:

$$b\delta(l_{b,c}, r) + c = r\delta^2(l_{b,c}, r), \text{ i.e}$$
$$\delta^2(l_{b,c}, r) - \frac{b\delta(l_{b,c}, r)}{r} - \frac{c}{r} = 0 \ .$$

Hence $\delta(l_{b,c}, r) = \frac{b}{2r} + \frac{1}{2}\sqrt{\frac{b^2}{r^2} + \frac{4c}{r}}$. Thus

$$\delta^2(l_{b,c}, r) \leqslant 2\left(\frac{b^2}{4r^2} + \frac{b^2}{4r^2} + \frac{c}{r}\right) \leqslant \frac{b^2}{r^2} + \frac{2c}{r}.$$

This proves (3.33). For any $x > 0$, $g_{a,b,c}(x) \leqslant rx^2$ is equivalent to

$$ax \leqslant rx^2 \tag{3.35}$$
$$\text{and } bx^3 + cx^2 \leqslant r^2x^4 \ . \tag{3.36}$$

Eq. (3.35) is equivalent to $x \geqslant \frac{a}{r}$. On the other hand,

$$\begin{aligned} x > \left[\frac{b^2}{r^4} + \frac{2c}{r^2}\right]^{\frac{1}{2}} &\implies x > \delta(l_{b,c}, r^2) \text{ by (3.33)} \\ &\implies bx + c \leqslant r^2x^2 \text{ by Definition 3.7.1} \\ &\implies \text{(3.36)}. \end{aligned}$$

Therefore, whenever

$$x > \frac{a}{r} \vee \left[\frac{b^2}{r^4} + \frac{2c}{r^2}\right]^{\frac{1}{2}} \ ,$$

it holds that $g_{a,b,c}(x) \leqslant rx^2$. (3.34) follows by Definition 3.7.1. ■

3.8.2 Uniform control on the empirical process

From now on until the end of the proof, the notation and hypotheses of Theorem 3.4.3 are used. Recall also the notation $g \circ t : (x, y) \mapsto g(t(x), y)$, for any $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and $t : \mathcal{X} \to \mathbb{R}$. Fix a training set D_{n_t}. Start with the following definition.

Definition 3.8.6 *For $t_1, t_2 \in \mathcal{H}$, let*

$$d(t_1, t_2) = \min_{\lambda \in \Lambda} \|t_1 - s_\lambda\|_{\mathcal{H}} + \|t_1 - t_2\|_{\mathcal{H}} \ , \tag{3.37}$$

where $s_\lambda = \operatorname{argmin}_{t\in\mathcal{H}} \left\{P(c\circ t) + \lambda \|t\|_{\mathcal{H}}^2\right\}$. *Furthermore, let*

$$\widehat{y} = \frac{\lambda_m n_t}{32\kappa L^2} \times \sup_{(t_1,t_2)\in\mathcal{H}^2} \left\{(P_{n_t} - P)(c\circ t_1 - c\circ t_2) - \frac{\lambda_m}{2} d(t_1,t_2)^2\right\} \ ,$$

so that

$$\forall (t_1,t_2)\in\mathcal{H}^2, (P_{n_t}-P)(c\circ t_1 - c\circ t_2) \leqslant \frac{\lambda_m}{2} d(t_1,t_2)^2 + \frac{32\kappa L^2 \widehat{y}}{\lambda_m n_t} \ . \tag{3.38}$$

We then have the following bounds on $\widehat{y}$.

Claim 3.8.6.1 *For all* $x \geqslant 0$,

$$\mathbb{P}\big(\widehat{y} \geqslant 2.6 + \log|\Lambda| + x\big) \leqslant \mathrm{e}^{-x} \ .$$

In particular, $\mathbb{E}[\widehat{y}] \leqslant 4 + \log|\Lambda|$.

Proof Let $(t_1,t_2)\in\mathcal{H}$ be such that $d(t_1,t_2)\leqslant r$. Let $\lambda\in\Lambda$ be such that $\|t_1 - s_\lambda\|_{\mathcal{H}} + \|t_1 - t_2\|_{\mathcal{H}} \leqslant r$. By the triangle inequality, $t_1, t_2 \in B(s_\lambda, r)$. Hence

$$\sup_{(t_1,t_2):d(t_1,t_2)\leqslant r} \{(P_{n_t}-P)(c\circ t_1 - c\circ t_2)\} \leqslant \max_{\lambda\in\Lambda} \sup_{(t_1,t_2)\in B(s_\lambda,r)^2} (P_{n_t}-P)(c\circ t_1 - c\circ t_2). \tag{3.39}$$

From Proposition 3.8.2 and the union bound, it follows that, for any $x\geqslant 0$,

$$\mathbb{P}\left[\max_{\lambda\in\Lambda} \sup_{(t_1,t_2)\in B(s_\lambda,r)^2} (P_{n_t}-P)(c\circ t_1 - c\circ t_2) \geqslant 2\left(2+\sqrt{2(x+\log|\Lambda|)}\right) L\frac{r\sqrt{\kappa}}{\sqrt{n_t}}\right] \leqslant \mathrm{e}^{-x}.$$

It follows by Equation (3.39) that, for all $x\geqslant 0$,

$$\mathbb{P}\left[\sup_{(t_1,t_2):d(t_1,t_2)\leqslant r} \frac{1}{4L\sqrt{\kappa}} (P_{n_t}-P)(c\circ t_1 - c\circ t_2) \geqslant \left(1+\sqrt{\frac{x+\log|\Lambda|}{2}}\right)\frac{r}{\sqrt{n_t}}\right] \leqslant \mathrm{e}^{-x}.$$

By Lemma 3.8.3 with $\theta = \frac{\lambda_m}{8L\sqrt{\kappa}}$, $a = \log|\Lambda|$, $b = \frac{1}{2}$, with probability larger than $1-\mathrm{e}^{-x}$,

$$\forall (t_1,t_2), (P_{n_t}-P)(c\circ t_1 - c\circ t_2) \leqslant \frac{\lambda_m}{2} d(t_1,t_2)^2 + 32L^2 \frac{\kappa(2.6 + x + \log|\Lambda|)}{\lambda_m n_t}.$$

On the same event, $\widehat{y} \leqslant 2.6 + x + \log|\Lambda|$ by Definition 3.8.6.

Therefore, by Lemma 3.7.7, $\mathbb{E}[\widehat{y}] \leqslant 3.6 + \log|\Lambda|$. ■

Definition 3.8.6 and Proposition 3.8.6.1 together imply a uniform control on the empirical process thanks to the drift term $\lambda_m d(t_1,t_2)^2$, whereas Proposition 3.8.6.2 only gave a bound on an RKHS ball of fixed radius.

3.8.3 Verifying the assumptions of Theorem 3.7.3

Theorem 3.4.3 is a consequence of Theorem 3.7.3. For all $\lambda \in \Lambda$, let $\widehat{t}_\lambda = \mathcal{A}_\lambda(D_{n_t})$, where $\mathcal{A}_\lambda$ is given by Definition 3.4.1. To verify the assumptions of Theorem 3.7.3, adequate functions $(\widehat{w}_{i,j})_{(i,j)\in\{1;2\}^2}$ must be found such that for $i \in \{1;2\}$, $H\left(\widehat{w}_{i,1}, \widehat{w}_{i,2}, (\widehat{t}_\lambda)_{\lambda\in\Lambda}\right)$ holds almost surely . This is the purpose of this section.

The core of the proof of Theorem 3.4.3 lies in the following deterministic claim.

Claim 3.8.6.2 *For all $\lambda, \mu \in \Lambda$ such that $\lambda \leqslant \mu$,*

$$\left\|\widehat{t}_\lambda - \widehat{t}_\mu\right\|_\infty^2 \leqslant \frac{\kappa C}{\lambda_m}\ell(s, \widehat{t}_\mu) + 96L^2\frac{\kappa^2\widehat{y}}{\lambda_m^2 n_t} \ .$$

Proof Let $(\lambda, \mu) \in \Lambda^2$ with $\lambda \leqslant \mu$. Let s_μ be as in Definition 3.8.6, Equation (3.37). By convexity of c, the function $t \mapsto P(c \circ t) + \mu \|t\|_{\mathcal{H}}^2$ is μ-strongly convex. Since s_μ is its optimum, we get

$$\forall t \in \mathcal{H}, P(c \circ t) + \mu \|t\|_{\mathcal{H}}^2 \geqslant P(c \circ s_\mu) + \mu \|s_\mu\|_{\mathcal{H}}^2 + \mu \|t - s_\mu\|_{\mathcal{H}}^2 \ .$$

Hence, taking $t = \widehat{t}_\mu$,

$$\begin{aligned}
\lambda_m \left\|\widehat{t}_\mu - s_\mu\right\|_{\mathcal{H}}^2 &\leqslant \mu \left\|\widehat{t}_\mu - s_\mu\right\|_{\mathcal{H}}^2 \\
&\leqslant P(c \circ \widehat{t}_\mu) + \mu \left\|\widehat{t}_\mu\right\|_{\mathcal{H}}^2 - P(c \circ s_\mu) - \mu \|s_\mu\|_{\mathcal{H}}^2 \\
&= P_{n_t}(c \circ \widehat{t}_\mu) + \mu \left\|\widehat{t}_\mu\right\|_{\mathcal{H}}^2 - P_{n_t}(c \circ s_\mu) - \mu \|s_\mu\|_{\mathcal{H}}^2 + (P - P_{n_t})(c \circ \widehat{t}_\mu - c \circ s_\mu) \ .
\end{aligned}$$

By Definition 3.4.1,

$$P_{n_t}(c \circ \widehat{t}_\mu) + \mu \left\|\widehat{t}_\mu\right\|_{\mathcal{H}}^2 \leqslant P_{n_t}(c \circ s_\mu) + \mu \|s_\mu\|_{\mathcal{H}}^2 .$$

Hence $\lambda_m \left\|\widehat{t}_\mu - s_\mu\right\|_{\mathcal{H}}^2 \leqslant (P - P_{n_t})(c \circ \widehat{t}_\mu - c \circ s_\mu) = (P_{n_t} - P)(c \circ s_\mu - c \circ \widehat{t}_\mu)$. Now take $t_1 = s_\mu$ and $t_2 = \widehat{t}_\mu$ in Equation (3.38) of Definition 3.8.6 to get

$$\begin{aligned}
\lambda_m \left\|\widehat{t}_\mu - s_\mu\right\|_{\mathcal{H}}^2 &\leqslant \frac{\lambda_m}{2} d(s_\mu, \widehat{t}_\mu)^2 + 32L^2\frac{\kappa\widehat{y}}{\lambda_m n_t} \\
&= \frac{\lambda_m}{2}\left\|s_\mu - \widehat{t}_\mu\right\|_{\mathcal{H}}^2 + 32L^2\frac{\kappa\widehat{y}}{\lambda_m n_t} \ .
\end{aligned}$$

Therefore,

$$\left\|\widehat{t}_\mu - s_\mu\right\|_{\mathcal{H}}^2 \leqslant 64L^2\frac{\widehat{y}\kappa}{\lambda_m^2 n_t} \ . \tag{3.40}$$

Now $\left\|\widehat{t}_\lambda - \widehat{t}_\mu\right\|_{\mathcal{H}}^2$ can be bounded as follows. Since $t \mapsto P_{n_t}(c \circ t) + \lambda \|t\|_{\mathcal{H}}^2$ is λ-strongly convex and $\widehat{t}_\lambda$ is its optimum,

$$\begin{aligned}\lambda_m \left\|\widehat{t}_\lambda - \widehat{t}_\mu\right\|_{\mathcal{H}}^2 &\leqslant \lambda \left\|\widehat{t}_\lambda - \widehat{t}_\mu\right\|_{\mathcal{H}}^2 \\ &\leqslant P_{n_t}(c \circ \widehat{t}_\mu) - P_{n_t}(c \circ \widehat{t}_\lambda) + \lambda \left\|\widehat{t}_\mu\right\|_{\mathcal{H}}^2 - \lambda \left\|\widehat{t}_\lambda\right\|_{\mathcal{H}}^2 \quad .\end{aligned}$$

By Lemma 3.8.4 with $h(t) = P_{n_t}(c \circ t)$, $\left\|\widehat{t}_\lambda - \widehat{t}_\mu\right\|_{\mathcal{H}}^2 \leqslant \left\|\widehat{t}_\lambda\right\|_{\mathcal{H}}^2 - \left\|\widehat{t}_\mu\right\|_{\mathcal{H}}^2$. Hence

$$\begin{aligned}(\lambda_m + \lambda) \left\|\widehat{t}_\lambda - \widehat{t}_\mu\right\|_{\mathcal{H}}^2 &\leqslant P_{n_t}(c \circ \widehat{t}_\mu) - P_{n_t}(c \circ \widehat{t}_\lambda) \\ &= P(c \circ \widehat{t}_\mu) - P(c \circ \widehat{t}_\lambda) + (P_{n_t} - P)\left[c \circ \widehat{t}_\mu - c \circ \widehat{t}_\lambda\right] \\ &\leqslant P(c \circ \widehat{t}_\mu) - \min_{t \in \mathbb{S}} P(c \circ t) + (P_{n_t} - P)\left[c \circ \widehat{t}_\mu - c \circ \widehat{t}_\lambda\right] \\ &\leqslant C\ell(s, \widehat{t}_\mu) + (P_{n_t} - P)\left[c \circ \widehat{t}_\mu - c \circ \widehat{t}_\lambda\right] \text{ by hypothesis } Comp_C(g, c) \quad .\end{aligned}$$

By Definition 3.8.6, Equation (3.38) with $t_1 = \widehat{t}_\mu$ and $t_2 = \widehat{t}_\lambda$,

$$\begin{aligned}(\lambda_m + \lambda) \left\|\widehat{t}_\lambda - \widehat{t}_\mu\right\|_{\mathcal{H}}^2 &\leqslant C\ell(s, \widehat{t}_\mu) + \frac{\lambda_m}{2}\left[\left\|\widehat{t}_\mu - s_\mu\right\|_{\mathcal{H}} + \left\|\widehat{t}_\lambda - \widehat{t}_\mu\right\|_{\mathcal{H}}\right]^2 + 32L^2\frac{\kappa\widehat{y}}{\lambda_m n_t} \\ &\leqslant C\ell(s, \widehat{t}_\mu) + \frac{\lambda_m}{2}\left[8\frac{L\sqrt{\widehat{y}\kappa}}{\lambda_m\sqrt{n_t}} + \left\|\widehat{t}_\lambda - \widehat{t}_\mu\right\|_{\mathcal{H}}\right]^2 + 32L^2\frac{\kappa\widehat{y}}{\lambda_m n_t} \\ &\quad \text{by equation (3.40).}\end{aligned}$$

For any (a, b), $(a+b)^2 \leqslant 2a^2 + 2b^2$, hence

$$(\lambda + \lambda_m) \left\|\widehat{t}_\lambda - \widehat{t}_\mu\right\|_{\mathcal{H}}^2 \leqslant C\ell(s, \widehat{t}_\mu) + \frac{\lambda_m}{2}\left[128L^2\frac{\widehat{y}\kappa}{\lambda_m^2 n_t} + 2\left\|\widehat{t}_\lambda - \widehat{t}_\mu\right\|_{\mathcal{H}}^2\right] + 32L^2\frac{\kappa\widehat{y}}{\lambda_m n_t}.$$

This yields:

$$\lambda \left\|\widehat{t}_\lambda - \widehat{t}_\mu\right\|_{\mathcal{H}}^2 \leqslant C\ell(s, \widehat{t}_\mu) + 96L^2\frac{\kappa\widehat{y}}{\lambda_m n_t},$$

and finally, since $\lambda \geqslant \lambda_m$:

$$\left\|\widehat{t}_\lambda - \widehat{t}_\mu\right\|_{\mathcal{H}}^2 \leqslant \frac{C\ell(s, \widehat{t}_\mu)}{\lambda_m} + 96L^2\frac{\kappa\widehat{y}}{\lambda_m^2 n_t}.$$

Now, by Lemma 3.8.1,

$$\begin{aligned}\left\|\widehat{t}_\lambda - \widehat{t}_\mu\right\|_\infty^2 &\leqslant \kappa \left\|\widehat{t}_\lambda - \widehat{t}_\mu\right\|_{\mathcal{H}}^2 \\ &\leqslant \frac{\kappa C}{\lambda_m}\ell(s, \widehat{t}_\mu) + 96L^2\frac{\kappa^2\widehat{y}}{\lambda_m^2 n_t}.\end{aligned}$$

This proves Claim 3.8.6.2. ■

Using hypothesis $SC_{\rho,\nu}$ —Equation (3.4)—, a refined bound can be obtained on $P\left[\left(g\circ\widehat{t}_\lambda - g\circ\widehat{t}_\mu\right)^2\right]$.

Claim 3.8.6.3 *For any* $(\lambda,\mu)\in\Lambda^2$,

$$P\left[\left(g\circ\widehat{t}_\lambda - g\circ\widehat{t}_\mu\right)^2\right] \leqslant \widehat{w}_B\left(\sqrt{\ell(s,\widehat{t}_\lambda)}\right)^2 + \widehat{w}_B\left(\sqrt{\ell(s,\widehat{t}_\mu)}\right)^2$$

where

$$\widehat{w}_B(x)^2 = \max\left\{\rho x^2,\ \nu\frac{4}{3}\sqrt{\frac{\kappa C}{\lambda_m}}x^3 + 10\nu L\frac{\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}}x^2\right\}\ .$$

Proof By hypothesis $SC_{\rho,\nu}$ —Equation (3.4)— with $u=\widehat{t}_\lambda(X)$ and $v=\widehat{t}_\mu(X)$,

$$\begin{aligned}\mathbb{E}\left[(g\circ\widehat{t}_\lambda - g\circ\widehat{t}_\mu)^2(X,Y)|X\right] &\leqslant \left[\rho\vee\left(\nu|\widehat{t}_\lambda(X)-\widehat{t}_\mu(X)|\right)\right]\left[\ell_X(\widehat{t}_\lambda(X))+\ell_X(\widehat{t}_\mu(X))\right]\\ &\leqslant \left[\rho\vee\left(\nu\left\|\widehat{t}_\lambda-\widehat{t}_\mu\right\|_\infty\right)\right]\left[\ell_X(\widehat{t}_\lambda(X))+\ell_X(\widehat{t}_\mu(X))\right],\end{aligned}$$

where $\ell_X(u) = \mathbb{E}[g(u,Y)|X] - \min_{v\in\mathbb{R}}\mathbb{E}[g(v,Y)|X]$. Integrating this inequality with respect to X, it follows that,

$$P\left[\left(g\circ\widehat{t}_\lambda - g\circ\widehat{t}_\mu\right)^2\right] \leqslant \left[\rho\vee\left(\nu\left\|\widehat{t}_\lambda-\widehat{t}_\mu\right\|_\infty\right)\right]\left[\ell(s,\widehat{t}_\lambda)+\ell(s,\widehat{t}_\mu)\right].$$

Assume without loss of generality that $\lambda\leqslant\mu$. By Claim 3.8.6.2,

$$\begin{aligned}P\left[\left(g\circ\widehat{t}_\lambda - g\circ\widehat{t}_\mu\right)^2\right] &\leqslant \left(\rho\vee\nu\left[\sqrt{\frac{\kappa C}{\lambda_m}}\sqrt{\ell(s,\widehat{t}_\mu)}+10\frac{L\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}}\right]\right)\left[\ell(s,\widehat{t}_\lambda)+\ell(s,\widehat{t}_\mu)\right]\\ &\leqslant \max\Bigg\{\rho\left[\ell(s,\widehat{t}_\lambda)+\ell(s,\widehat{t}_\mu)\right], \nu\Bigg[\sqrt{\frac{\kappa C}{\lambda_m}}\left(\sqrt{\ell(s,\widehat{t}_\mu)}\ell(s,\widehat{t}_\lambda)+\sqrt{\ell(s,\widehat{t}_\mu)}^3\right)\\ &\quad + 10\frac{L\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}}\left[\ell(s,\widehat{t}_\lambda)+\ell(s,\widehat{t}_\mu)\right]\Bigg]\Bigg\}. \qquad (3.41)\end{aligned}$$

Using the inequality $ab\leqslant\frac{a^p}{p}+\frac{b^q}{q}$ with Hölder conjugates $p=3$, $q=\frac{3}{2}$, we have:

$$\begin{aligned}\sqrt{\ell(s,\widehat{t}_\mu)}\ell(s,\widehat{t}_\lambda)+\sqrt{\ell(s,\widehat{t}_\mu)}^3 &\leqslant \frac{1}{3}\sqrt{\ell(s,\widehat{t}_\mu)}^3+\frac{2}{3}\ell(s,\widehat{t}_\lambda)^{\frac{3}{2}}+\sqrt{\ell(s,\widehat{t}_\mu)}^3\\ &\leqslant \frac{4}{3}\left[\sqrt{\ell(s,\widehat{t}_\lambda)}^3+\sqrt{\ell(s,\widehat{t}_\mu)}^3\right]. \qquad (3.42)\end{aligned}$$

Claim 3.8.6.3 then follows from inequalities (3.41) and (3.42) using the elementary inequality $(a+b)\vee(c+d)\leqslant a\vee c+b\vee d$. ■

As g is L-Lipschitz in its first argument, it follows from Claim 3.8.6.2 that for all $\lambda,\mu\in\Lambda$ s.t. $\lambda\leqslant\mu$,

$$\begin{aligned}\left\|g\circ\widehat{t}_\lambda-g\circ\widehat{t}_\mu\right\|_\infty &\leqslant L\left\|\widehat{t}_\lambda-\widehat{t}_\mu\right\|_\infty\\ &\leqslant L\sqrt{\frac{\kappa C}{\lambda_m}}\sqrt{\ell(s,\widehat{t}_\mu)}+10L^2\frac{\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}}\\ &\leqslant \widehat{w}_A\left(\sqrt{\ell(s,\widehat{t}_\mu)}\right)+\widehat{w}_A\left(\sqrt{\ell(s,\widehat{t}_\lambda)}\right)\quad,\end{aligned}\tag{3.43}$$

where

$$\widehat{w}_A(x)=L\sqrt{\frac{\kappa C}{\lambda_m}}x+5L^2\frac{\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}}.\tag{3.44}$$

If follows that for all $k\geqslant 2$,

$$\begin{aligned}P\left[\left(g\circ\widehat{t}_\lambda-g\circ\widehat{t}_\mu\right)^k\right] &\leqslant \left\|g\circ\widehat{t}_\lambda-g\circ\widehat{t}_\mu\right\|_\infty^k\\ &\leqslant \left[\widehat{w}_A\left(\sqrt{\ell(s,\widehat{t}_\mu)}\right)+\widehat{w}_A\left(\sqrt{\ell(s,\widehat{t}_\lambda)}\right)\right]^k\quad.\end{aligned}$$

This proves that hypothesis $H\left(\widehat{w}_A,\widehat{w}_A,(\widehat{t}_\lambda)_{\lambda\in\Lambda}\right)$, as defined in Appendix 3.7, holds true.

It follows from Claim 3.8.6.3 and Equation (3.43) that, for all $k\geqslant 2$,

$$\begin{aligned}P\left[\left|g\circ\widehat{t}_\lambda-g\circ\widehat{t}_\mu\right|^k\right] &\leqslant \left\|g\circ\widehat{t}_\lambda-g\circ\widehat{t}_\mu\right\|_\infty^{k-2}P\left[\left(g(\widehat{t}_\lambda(X),Y)-g(\widehat{t}_\mu(X),Y)\right)^2\right]\\ &\leqslant \left[\widehat{w}_A\Big(\sqrt{\ell(s,\widehat{t}_\lambda)}\Big)+\widehat{w}_A\Big(\sqrt{\ell(s,\widehat{t}_\mu)}\Big)\right]^{k-2}\\ &\quad\times\left[\widehat{w}_B\Big(\sqrt{\ell(s,\widehat{t}_\lambda)}\Big)+\widehat{w}_B\Big(\sqrt{\ell(s,\widehat{t}_\mu)}\Big)\right]^2\quad;\end{aligned}$$

which proves that $H\left(\widehat{w}_B,\widehat{w}_A,(\widehat{t}_\lambda)_{\lambda\in\Lambda}\right)$ holds true.

3.8.4 Conclusion of the proof

We have proved that $H\left(\widehat{w}_B,\widehat{w}_A,(\widehat{t}_\lambda)_{\lambda\in\Lambda}\right)$ and $H\left(\widehat{w}_A,\widehat{w}_A,(\widehat{t}_\lambda)_{\lambda\in\Lambda}\right)$ hold, where $\widehat{w}_B$ is defined in Proposition 3.8.6.3 and $\widehat{w}_A$ in Equation (3.44). Moreover, $x\mapsto\frac{\widehat{w}_A(x)}{x}$ is nonincreasing. Therefore, Theorem 3.7.3 applies with $\widehat{w}_{1,1}=\widehat{w}_A,\widehat{w}_{1,2}=\widehat{w}_A,\widehat{w}_{2,1}=$

$\widehat{w}_B, \widehat{w}_{2,2} = \widehat{w}_A$, $x = \log n_v$ and it remains to bound the remainder terms $(R_{2,i})_{1\leqslant i\leqslant 4}$ of Equation (3.12). For each i, we bound $R_{2,i}(\theta)$ by an absolute constant times $\max\{T_1(\theta), T_2(\theta), T_3(\theta)\}$, where

$$\begin{aligned} T_1(\theta) &= \frac{6\rho}{100}\frac{\log(n_v|\Lambda|)}{\theta n_v} \\ T_2(\theta) &= (\nu \vee L)^2\, \kappa C \frac{\log^2(n_v|\Lambda|)}{\theta^3 \lambda_m n_v^2} \\ T_3(\theta) &= L(\nu \vee L)\kappa \frac{\log^{\frac{3}{2}}(n_v|\Lambda|)}{\theta \lambda_m n_v \sqrt{n_t}} \quad . \end{aligned}$$

Summing up these bounds yields Theorem 3.4.3.

Bound on $R_{2,1}(\theta) = \sqrt{2}\theta\mathbb{E}\left[\delta^2\left(\widehat{w}_B, \frac{\theta}{2}\sqrt{\frac{n_v}{\log(n_v|\Lambda|)}}\right)\right]$

Recall that $\widehat{w}_B(x)^2 := \max\left\{\rho x^2,\ \nu\frac{4}{3}\sqrt{\frac{\kappa C}{\lambda_m}}x^3 + 10\nu L\frac{\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}}x^2\right\}$.

By Equation (3.34) in Lemma 3.8.5 with $a = \sqrt{\rho}$, $b = \nu\frac{4}{3}\sqrt{\frac{\kappa C}{\lambda_m}}$, $c = 10\nu L\frac{\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}}$,

$$\delta^2\left(\widehat{w}_B, \frac{\theta}{2}\sqrt{\frac{n_v}{\log(n_v|\Lambda|)}}\right) \leqslant 4\rho\frac{\log(n_v|\Lambda|)}{\theta^2 n_v} + 29\nu^2\kappa C\frac{\left[\log(n_v|\Lambda|)\right]^2}{\theta^4\lambda_m n_v^2} + 80\nu L\kappa\frac{\left[\log(n_v|\Lambda|)\right]\sqrt{\widehat{y}}}{\theta^2\lambda_m n_v\sqrt{n_t}} \quad . \tag{3.45}$$

Therefore,

$$R_{2,1}(\theta) \leqslant 4\sqrt{2}\rho\frac{\log(n_v|\Lambda|)}{\theta n_v} + 29\sqrt{2}\nu^2\kappa C\frac{\left[\log(n_v|\Lambda|)\right]^2}{\theta^3\lambda_m n_v^2} + 80\sqrt{2}\nu L\kappa\frac{\left[\log(n_v|\Lambda|)\right]\sqrt{\mathbb{E}[\widehat{y}]}}{\theta\lambda_m n_v\sqrt{n_t}} \quad .$$

By Proposition 3.8.6.1, $\mathbb{E}[\widehat{y}] \leqslant 4 + \log|\Lambda|$. Since $n_v \geqslant 100 \geqslant \mathrm{e}^4$, $\mathbb{E}[\widehat{y}] \leqslant \log(n_v|\Lambda|)$. As a result,

$$\begin{aligned} R_{2,1}(\theta) &\leqslant 6\rho\frac{\log(n_v|\Lambda|)}{\theta n_v} + 42\nu^2\kappa C\frac{\left[\log(n_v|\Lambda|)\right]^2}{\theta^3\lambda_m n_v^2} + 114\nu L\kappa\frac{\left[\log(n_v|\Lambda|)\right]^{\frac{3}{2}}}{\theta\lambda_m n_v\sqrt{n_t}} \\ &\leqslant 100T_1(\theta) + 42T_2(\theta) + 114T_3(\theta) \\ &\leqslant 256 \times \max\left\{T_1(\theta), T_2(\theta), T_3(\theta)\right\} \quad . \end{aligned}$$

Bound on $R_{2,2}(\theta) = \frac{\theta^2}{2}\mathbb{E}\left[\delta^2\left(\widehat{w}_A, \frac{\theta^2}{4}\frac{n_v}{\log(n_v|\Lambda|)}\right)\right]$

Recall that by definition, $\widehat{w}_A(x) = L\sqrt{\frac{\kappa C}{\lambda_m}}x + 5L^2\frac{\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}}$ (Equation (3.44)). By Equation (3.33) in Lemma 3.8.5 with $b = L\sqrt{\frac{\kappa C}{\lambda_m}}$ and $c = 5L^2\frac{\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}}$, we have

$$\delta^2\left(\widehat{w}_A, \frac{\theta^2}{4}\frac{n_v}{\log(n_v|\Lambda|)}\right) \leqslant 16L^2\kappa C\frac{\log^2(n_v|\Lambda|)}{\theta^4\lambda_m n_v^2} + 40L^2\kappa\frac{\left[\log(n_v|\Lambda|)\right]\sqrt{\widehat{y}}}{\theta^2\lambda_m n_v\sqrt{n_t}} \ . \quad (3.46)$$

As $\mathbb{E}[\widehat{y}] \leqslant \log(n_v|\Lambda|)$ by Proposition 3.8.6.1, it follows that

$$\begin{aligned} R_{2,2}(\theta) &\leqslant 8L^2\kappa C\frac{\log^2(n_v|\Lambda|)}{\theta^2\lambda_m n_v^2} + 20L^2\kappa\frac{\log^{\frac{3}{2}}(n_v|\Lambda|)}{\lambda_m n_v\sqrt{n_t}} \\ &\leqslant 8\theta T_2(\theta) + 20\theta T_3(\theta) \\ &\leqslant 28 \times \max\{T_1(\theta), T_2(\theta), T_3(\theta)\} \text{ since } \theta \in (0;1] \ . \end{aligned}$$

Bound on $R_{2,3}(\theta) = \frac{1}{n_v}\left(\theta + \frac{2\left[1+\log(|\Lambda|)\right]}{\theta}\right)\mathbb{E}\left[\widehat{\delta^2}\big(\widehat{w}_A, \sqrt{n_v}\big)\right]$

By Equation (3.33) in Lemma 3.8.5 with $b = L\sqrt{\frac{\kappa C}{\lambda_m}}, c = 5L^2\frac{\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}}$,

$$\delta^2(\widehat{w}_A, \sqrt{n_v}) \leqslant L^2\frac{\kappa C}{\lambda_m n_v} + L^2\frac{10\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_v n_t}} \ . \quad (3.47)$$

As $\theta \in (0;1]$ and $n_v \geqslant 100 \geqslant \mathrm{e}^{\frac{3}{2}}$, we have $\theta + \frac{2}{\theta} \leqslant \frac{3}{\theta} \leqslant \frac{2\log n_v}{\theta}$, hence

$$\theta + \frac{2(1+\log(|\Lambda|))}{\theta} \leqslant \frac{2\log(n_v|\Lambda|)}{\theta} \ . \quad (3.48)$$

Therefore,

$$R_{2,3}(\theta) \leqslant \frac{2\log(n_v|\Lambda|)}{\theta n_v}\left[L^2\frac{\kappa C}{\lambda_m n_v} + L^2\frac{10\kappa\sqrt{\mathbb{E}[\widehat{y}]}}{\lambda_m\sqrt{n_v n_t}}\right] \ .$$

Since $\mathbb{E}[\widehat{y}] \leqslant \log(n_v|\Lambda|)$ by Proposition 3.8.6.1,

$$\begin{aligned} R_{2,3}(\theta) &\leqslant 2\log(n_v|\Lambda|)\frac{L^2\kappa C}{\theta\lambda_m n_v^2} + 20L^2\kappa\frac{\log^{\frac{3}{2}}(n_v|\Lambda|)}{\theta\lambda_m n_v\sqrt{n_v n_t}} \\ &\leqslant \frac{2\theta^2}{\log(n_v|\Lambda|)}T_2(\theta) + \frac{20}{\sqrt{n_v}}T_3(\theta) \\ &\leqslant 0.4T_2(\theta) + 2T_3(\theta) \text{ since } n_v \geqslant 100 \text{ and } |\Lambda| \geqslant 2 \\ &\leqslant 2.4 \times \max\{T_1, T_2, T_3\} \ . \end{aligned}$$

Bound on $R_{2,4}(\theta) = \frac{1}{n_v}\left(\theta + \frac{2\left[1+\log(|\Lambda|)\right]+\log^2(|\Lambda|)}{\theta}\right)\mathbb{E}\left[\widehat{\delta}^2\big(\widehat{w}_A, n_v\big)\right]$

By Equation (3.33) in Lemma 3.8.5 with $b = L\sqrt{\frac{\kappa C}{\lambda_m}}$, $c = 5L^2\frac{\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}}$,

$$\delta^2(\widehat{w}_A, n_v) \leqslant L^2\frac{\kappa C}{\lambda_m n_v^2} + L^2\frac{10\kappa\sqrt{\widehat{y}}}{\lambda_m n_v\sqrt{n_t}} \ . \tag{3.49}$$

Since $\theta \in [0;1]$, $n_v \geqslant 100$ and $|\Lambda| \geqslant 2$, we have $\log(n_v|\Lambda|) \geqslant \log(200) \geqslant 5$ and

$$\begin{aligned}\theta + \frac{2\big[1+\log(|\Lambda|)\big]}{\theta} &\leqslant \frac{2\log(n_v|\Lambda|)}{\theta} \text{ by equation (3.48)}\\ &\leqslant \frac{2\log^2(n_v|\Lambda|)}{5\theta} \ .\end{aligned}$$

Hence, by Equation (3.49),

$$R_{2,4}(\theta) \leqslant \frac{1,4\log^2(n_v|\Lambda|)}{\theta n_v}\left[L^2\frac{\kappa C}{\lambda_m n_v^2} + L^2\frac{10\kappa\sqrt{\mathbb{E}[\widehat{y}]}}{\lambda_m n_v\sqrt{n_t}}\right] \ .$$

Since $\mathbb{E}[\widehat{y}] \leqslant \log(n_v|\Lambda|)$,

$$\begin{aligned}R_{2,4}(\theta) &\leqslant 1,4\log^2(n_v|\Lambda|)\frac{L^2\kappa C}{\theta\lambda_m n_v^3} + 14L^2\kappa\frac{\log^{\frac{5}{2}}(n_v|\Lambda|)}{\theta\lambda_m n_v^2\sqrt{n_t}}\\ &\leqslant \frac{1,4\theta^2}{n_v}T_2(\theta) + 14\frac{\log(n_v|\Lambda|)}{n_v}T_3(\theta) \ .\end{aligned}$$

Since $n_v \geqslant 100$ and $|\Lambda| \leqslant \mathrm{e}^{\sqrt{n_v}}$, we have $\frac{\log(n_v|\Lambda|)}{n_v} \leqslant \frac{\log(n_v)}{n_v} + \frac{\log(\mathrm{e}^{\sqrt{n_v}})}{n_v} \leqslant \frac{\log(100)}{100} + \frac{1}{10} \leqslant 0.15$ and so

$$\begin{aligned}R_{2,4}(\theta) &\leqslant 0.014T_2(\theta) + 2.1T_3(\theta)\\ &\leqslant 2.2 \times \max\{T_1(\theta), T_2(\theta), T_3(\theta)\} \ .\end{aligned}$$

Conclusion

Summing up the above inequalities, we get that for every $\theta \in (0;1]$,

$$\begin{aligned}R_2(\theta) &= R_{2,1}(\theta) + R_{2,2}(\theta) + R_{2,3}(\theta) + R_{2,4}(\theta)\\ &\leqslant 289\max\{T_1(\theta), T_2(\theta), T_3(\theta)\} \ .\end{aligned}$$

Equation (3.12) in Theorem 3.7.3 thus yields

$$\Big(1-\theta\Big)\mathbb{E}[\ell(s, \widehat{f}^{\mathrm{ag}}_{\mathcal{T}})] \leqslant \Big(1+\theta\Big)\mathbb{E}\Big[\min_{\lambda\in\Lambda}\ell(s, \mathcal{A}_\lambda(D_{n_t}))\Big] + 289\max\{T_1(\theta), T_2(\theta), T_3(\theta)\}$$

which proves Theorem 3.4.3 with $b_1 = 289(\nu \vee L)^2\kappa C$ and $b_2 = 289L(\nu \vee L)\kappa$. ■

3.9 Proof of Proposition 3.4.2 and Corollary 3.4.4

Let us start by two useful lemmas.

Lemma 3.9.1 *If ψ is a convex, Lipschitz-continuous, and even function, and Y is a random variable with a non-atomic distribution, the function*

$$R : u \mapsto \mathbb{E}\big[\psi(u - Y)\big]$$

is convex and differentiable with derivative $R'(u) = \mathbb{E}[\psi'(u - Y)]$. Moreover, if Y is symmetric around q, i.e $(q - Y) \sim (Y - q)$, then R reaches a minimum at q.

Proof First, remark that R is convex by convexity of ψ. Let $u \in \mathbb{R}$. For $h \neq 0$, let $k(h, Y) = \frac{\psi(u+h-Y)-\psi(u-Y)}{h}$. Let A be the set on which ψ is non-differentiable. Since ψ is convex, A is at most countable. By definition, $k(h, Y) \underset{h\to 0}{\longrightarrow} \psi'(u - Y)$ whenever $u - Y \notin A$, that is to say $Y \notin u - A$. Since Y is non-atomic, $\mathbb{P}(Y \notin u - A) = 1$. Moreover, since ψ is Lipschitz, there exists a constant L such that $\forall h \neq 0, |k(h, Y)| \leqslant L$. Therefore, by the dominated convergence theorem,

$$\frac{R(u + h) - R(u)}{h} = \mathbb{E}[k(h, Y)] \underset{h\to 0}{\longrightarrow} \mathbb{E}[\psi'(u - Y)] \ .$$

Thus, R is differentiable and for all $u \in \mathbb{R}$, $R'(u) = \mathbb{E}[\psi'(u - Y)]$.

Moreover, we have

$$\begin{aligned} R'(q) &= \mathbb{E}[\psi'(q - Y)] \\ &= -\mathbb{E}[\psi'(Y - q)] \text{ since } \psi'(-x) = -\psi'(x) \text{ on } \mathbb{R}\backslash A \\ &= -\mathbb{E}[\psi'(q - Y)] \text{ since } (Y - q) \sim (q - Y) \ , \end{aligned}$$

which implies that $R'(q) = 0$. Hence, R reaches a minimum at q since R is convex. ■

Lemma 3.9.2 *Let $g : \mathbb{R} \to \mathbb{R}$ be a differentiable convex function that reaches a minimum at $u_* \in \mathbb{R}$. If there exists ε, δ such that*

$$\forall u \in [u_* - \delta; u_* + \delta], \qquad |g'(u)| \geqslant \varepsilon |u - u_*| \ , \tag{3.50}$$

then for all $(u, v) \in \mathbb{R}^2$,

$$(u - v)^2 \leqslant \left[\frac{4}{\varepsilon} \vee \left(\frac{4}{\varepsilon\delta}|u - v|\right)\right] \big[g(u) + g(v) - 2g(u_*)\big] \ .$$

Proof By integrating Equation (3.50),

$$\forall u \in [u_* - \delta; u_* + \delta], (g(u) - g(u_*)) \geqslant \frac{\varepsilon}{2}(u - u_*)^2 \ . \tag{3.51}$$

Let

$$h(u) = \frac{1}{\delta}\left[g(u_* + \delta) - g(u_*)\right][u - u_*] \ . \tag{3.52}$$

By convexity of g, for any $u \geqslant u_* + \delta$, $g(u) - g(u_*) \geqslant h(u)$. Hence by Equation (3.51) with $u = u_* + \delta$ and Equation (3.52),

$$\forall u \geqslant u_* + \delta, g(u) - g(u_*) \geqslant \frac{1}{\delta}\frac{\varepsilon}{2}\delta^2[u - u_*] = \frac{\varepsilon\delta}{2}[u - u_*] \ . \tag{3.53}$$

The same argument applies to the convex function $g(-\cdot)$ with minimum $-u_*$, which yields

$$\forall u \in \mathbb{R}, |u - u_*| \geqslant \delta \implies g(u) - g(u_*) \geqslant \frac{\varepsilon\delta}{2}|u - u_*| \ . \tag{3.54}$$

Let $(u, v) \in \mathbb{R}^2$. Assume without loss of generality that $|u - u_*| \geqslant |v - u_*|$. If $|u - u_*| \leqslant \delta$ then by Equation (3.51),

$$\begin{aligned}(u - v)^2 &\leqslant 2\big[u - u_*\big]^2 + 2\big[v - u_*\big]^2 \\ &\leqslant \frac{4}{\varepsilon}\big[g(u) + g(v) - 2g\big(u_*\big)\big] \ ..\end{aligned} \tag{3.55}$$

Otherwise, by Equation (3.54),

$$\begin{aligned}(u - v)^2 &\leqslant |u - v|\big[|u - u_*| + |v - u_*|\big] \\ &\leqslant 2|u - v||u - u_*| \\ &\leqslant \frac{4}{\varepsilon\delta}|u - v|\big[g(u) - g\big(u_*\big)\big] \\ &\leqslant \frac{4}{\varepsilon\delta}|u - v|\big[g(u) + g(v) - 2g(u_*)\big] \ .\end{aligned} \tag{3.56}$$

■

3.9.1 Proof of Proposition 3.4.2

Now, we can prove Proposition 3.4.2. Let $R_x : u \mapsto \int |u - y| dF_x(y)$. By Lemma 3.9.1 with $\psi = |\cdot|$, for all $v \in \mathbb{R}$,

$$\begin{aligned}R'_x(v) &= \int \left[-\mathbb{I}_{v-y\leqslant 0} + \mathbb{I}_{v-y\geqslant 0}\right] dF_x(y) \\ &= F_x(v) - \big[1 - F_x(v)\big] \\ &= 2\left[F_x(v) - F_x(s(x))\right]\end{aligned}$$

since by definition, $F_x(s(x)) = \frac{1}{2}$. Hence by hypothesis (3.5), for all $u \in [s(x) - b(x); s(x) + b(x)]$,

$$|R'_x(u)| \geqslant 2a(x)|u - s(x)|.$$

Therefore by Lemma 3.9.2, for all $x \in \mathcal{X}$ and $(u, v) \in \mathbb{R}^2$,

$$\begin{aligned}(u-v)^2 &\leqslant \left(\frac{4}{a(x)} \vee \frac{4|u-v|}{a(x)b(x)}\right)\left[R_x(u) + R_x(v) - 2R_x\big(s(x)\big)\right] \\ &\leqslant \left(\frac{4}{a_m} \vee \left(\frac{4}{\mu_m}|u-v|\right)\right)\left[R_x(u) + R_x(v) - 2R_x\big(s(x)\big)\right].\end{aligned}$$

Since $g : (u, y) \mapsto |u - y|$, it follows by taking $x = X$ that

$$\big(g(u,Y) - g(v,Y)\big)^2 \leqslant (u-v)^2 \leqslant \left(\frac{4}{a_m} \vee \left(\frac{4}{\mu_m}|u-v|\right)\right)\big[\ell_X(u) + \ell_X(v)\big],$$

which implies hypothesis $SC_{\frac{4}{a_m},\frac{4}{\mu_m}}$. ■

3.9.2 Proof of Corollary 3.4.4

Corollary 3.4.4 is a consequence of Theorem 3.4.3. Let us check that its assumptions are satisfied.

Compatibility hypothesis $(Comp_1(c_0^{eps}, c_\varepsilon^{eps}))$ Fix $x \in \mathcal{X}$ and let p_x, F_x be the pdf and cdf corresponding to the distribution Y given $X = x$. By assumption, p_x is symmetric; $s(x)$ can be chosen equal to the center of symmetry (recall that the contrast function here is $\gamma(t,(x,y)) = c_0^{eps}(t(x), y) = |t(x) - y|$, so any conditional median is a possible value for $s(x)$). Let

$$R_{\varepsilon,x} : u \mapsto \int_y c_\varepsilon^{eps}(u, y)p_x(y)\mathrm{d}y = \int \psi_\varepsilon(u - y)p_x(y)\mathrm{d}y \ , \tag{3.57}$$

where $\psi_\varepsilon(z) = (|z| - \varepsilon)_+$ for any $z \in \mathbb{R}$. Lemma C.1 applies, since p_x is symmetric by assumption and ψ_ε is even, convex and 1-Lipschitz.

Hence for any $\varepsilon \geqslant 0$, $R_{\varepsilon,x}$ has a minimum at $s(x)$ and is differentiable, with

$$\begin{aligned}R'_{\varepsilon,x}(u) &= \int \psi'_\varepsilon(u - y)p_x(y)\mathrm{d}y = \int \left[-\mathbb{I}_{u-y\leqslant-\varepsilon} + \mathbb{I}_{u-y\geqslant\varepsilon}\right]p_x(y)\mathrm{d}y \\ &= F_x(u - \varepsilon) - [1 - F_x(u + \varepsilon)] \ .\end{aligned} \tag{3.58}$$

Therefore, for any $\varepsilon \geqslant 0$ and $u \in \mathbb{R}$,

$$R'_{\varepsilon,x}(u) - R'_{0,x}(u) = \int_0^\varepsilon [-p_x(u - t) + p_x(u + t)]\,\mathrm{d}t \ . \tag{3.59}$$

Now, assume that $u \geqslant s(x)$. By symmetry of p_x around $s(x)$, for all $t \geqslant 0$,

$$\begin{aligned} p_x(u-t) &= p_x(s(x) + (u - s(x) - t)) \\ &= p_x(s(x) + |u - s(x) - t|) \ . \end{aligned} \tag{3.60}$$

Since p_x is unimodal, its mode is $s(x)$ and p_x is non-increasing on $[s(x); +\infty)$. It follows from Equation (3.60) that for all $u \geqslant s(x)$ and $t \geqslant 0$,

$$\begin{aligned} p_x(u-t) &\geqslant p_x(s(x) + |u - s(x)| + t) \\ &= p_x(u+t). \end{aligned} \tag{3.61}$$

Therefore, by Eq. (3.59) and (3.61), for all $u \geqslant s(x)$ and $\varepsilon \geqslant 0$, $R'_{\varepsilon,x}(u) \leqslant R'_{0,x}(u)$. By integration, this implies that for all $u \geqslant s(x)$,

$$R_{\varepsilon,x}(u) - R_{\varepsilon,x}(s(x)) \leqslant R_{0,x}(u) - R_{0,x}(s(x)) \ . \tag{3.62}$$

By Equation (3.57) and symmetry of p_x, $R_{\varepsilon,x}$ and $R_{0,x}$ are symmetric around $s(x)$, hence inequality (3.62) is also valid when $u \leqslant s(x)$. Taking $x = X$, $u = t(X)$ and integrating, we get $\mathcal{L}_{c_\varepsilon^{eps}}(t) - \mathcal{L}_{c_\varepsilon^{eps}}(s) \leqslant \mathcal{L}_{c_0^{eps}}(t) - \mathcal{L}_{c_0^{eps}}(s)$ which proves $Comp_1(c_0^{eps}, c_\varepsilon^{eps})$.

Hypothesis $SC_{4\sigma,8}$ We first compute a lower bound on $R_{0,x}$.

Let $q_{x,\frac{1}{4}} = \sup\{y | F_x(y) \leqslant \frac{1}{4}\}$ and $q_{x,\frac{3}{4}} = \inf\{y | F_x(y) \geqslant \frac{3}{4}\}$. By continuity of F_x, $F_x(q_{x,\frac{1}{4}}) = \frac{1}{4}$ and $F_x(q_{x,\frac{3}{4}}) = \frac{3}{4}$. Let $\sigma(x) = q_{x,\frac{3}{4}} - q_{x,\frac{1}{4}}$, which is the smallest determination of the interquartile range. By symmetry of p_x around $s(x)$, $\frac{1}{2}\big[q_{x,\frac{1}{4}} + q_{x,\frac{3}{4}}\big] = s(x)$, therefore $q_{x,\frac{3}{4}} = s(x) + \frac{\sigma(x)}{2}$ and $q_{x,\frac{1}{4}} = s(x) - \frac{\sigma(x)}{2}$.

For any $u \in \Big[s(x) - \frac{\sigma(x)}{2}; s(x) + \frac{\sigma(x)}{2}\Big]$, by symmetry of p_x around $s(x)$,

$$\begin{aligned} \big|F_x(u) - F_x\big(s(x)\big)\big| &= \int_{s(x)}^{s(x)+|u-s(x)|} 2p_x(v)\mathrm{d}v \\ &= |u - s(x)| \frac{1}{|u - s(x)|} \int_{s(x)}^{s(x)+|u-s(x)|} 2p_x(v)\mathrm{d}v \ . \end{aligned}$$

Since p_x is non-increasing on $[s(x); +\infty)$ and $|u - s(x)| \leqslant \frac{\sigma(x)}{2}$,

$$\begin{aligned} \big|F_x(u) - F_x\big(s(x)\big)\big| &\geqslant |u - s(x)| \frac{2}{\sigma(x)} \int_{s(x)}^{s(x)+\frac{\sigma(x)}{2}} 2p_x(v)\mathrm{d}v \\ &= |u - s(x)| \frac{4}{\sigma(x)} \Big[F_x\big(q_{x,\frac{3}{4}}\big) - F_x\big(s(x)\big)\Big] \\ &= \frac{|u - s(x)|}{\sigma(x)} \ . \end{aligned}$$

Hence, by Proposition 3.4.2 with $a(x) = \frac{1}{\sigma(x)}$ and $b(x) = \frac{\sigma(x)}{2}$, (g, X, Y) satisfies hypothesis $SC_{4\sigma,8}$.

Conclusion To conclude, we apply Theorem 3.4.3 with $\kappa = 1, C = 1, L = 1$ (since c_0^{eps} and c_ε^{eps} are 1-Lipschitz), $\rho = 4\sigma$ and $\nu = 8$. Since constants b_1, b_2 of Theorem 3.4.3 only depend on κ, L, C, ν and all these parameters have now received explicit values, the constants b_1, b_2 are now absolute.

3.10 Classification: proof of Theorem 3.4.5

In the proof of Theorem 3.7.3, we used convexity of the risk to show that the risk of the average was less than the average of the risk. A property of this type also holds in the setting of classification, with the average replaced by the majority vote.

Proposition 3.10.1 *In the classification classification —see Example 3.2.1—, let* $(\widehat{f}_i)_{1\leqslant i\leqslant V}$ *denote a finite family of functions* $\mathcal{X} \to \mathcal{Y}$ *and let* $\widehat{f}^{mv}$ *be some majority vote rule:* $\forall x \in \mathcal{X}$, $\widehat{f}^{mv}(x) \in \operatorname{argmax}_{y\in\mathcal{Y}} |\{i \in [V] : \widehat{f}_i(x) = m\}|$. *Then,*

$$\ell(s, \widehat{f}^{mv}) \leqslant \frac{M}{V} \sum_{i=1}^{V} \ell(s, \widehat{f}_i) \qquad \textit{and} \qquad \mathcal{L}(\widehat{f}^{mv}) \leqslant \frac{2}{V} \sum_{i=1}^{V} \mathcal{L}(\widehat{f}_i) \ .$$

Proof For any $y \in \mathcal{Y}$, define $\eta_y : x \mapsto \mathbb{P}[Y = y | X = x]$. Then, for any $f \in \mathbb{S}$, $\mathcal{L}(f) = \mathbb{E}[1 - \eta_{f(X)}(X)]$ hence $s(X) \in \operatorname{argmax}_{y\in\mathcal{Y}} \eta_y(X)$ and

$$\ell(s, f) = \mathbb{E}\Big[\max_{y\in\mathcal{Y}} \eta_y(X) - \eta_{f(X)}(X)\Big] = \mathbb{E}\big[\eta_{s(X)}(X) - \eta_{f(X)}(X)\big] \ .$$

We now fix some $x \in \mathcal{X}$ and define $\mathcal{C}_x(y) = \{i \in [V] : \widehat{f}_i(x) = y\}$ and $C_x = \max_{y\in\mathcal{Y}} |\mathcal{C}_x(y)|$. Since $C_x M \geq \sum_{y\in\mathcal{Y}} |\mathcal{C}_x(y)| = V$, it holds $C_x \geqslant V/M$. On the other hand, by definition of $\widehat{f}^{\mathrm{mv}}$,

$$\frac{1}{V} \sum_{i=1}^{V} \big[\underbrace{\eta_{s(x)}(x) - \eta_{\widehat{f}_i(x)}(x)}_{\geqslant 0}\big] \geqslant \frac{C_x}{V}\big(\eta_{s(x)}(x) - \eta_{\widehat{f}^{\mathrm{mv}}(x)}(x)\big) \geqslant \frac{1}{M}\big(\eta_{s(x)}(x) - \eta_{\widehat{f}^{\mathrm{mv}}(x)}(x)\big) \ .$$

Integrating over x (with respect to the distribution of X) yields the first bound.

For the second bound, fix $x \in \mathcal{X}$ and define $\mathcal{C}_x(y)$ and C_x as above. Let $y \in \mathcal{Y}$ be such that $\widehat{f}^{\mathrm{mv}}(x) \neq y$. Since y occurs less often than $\widehat{f}^{\mathrm{mv}}(x)$ among $\widehat{f}_1(x), \ldots, \widehat{f}_V(x)$, we have $|\mathcal{C}_x(y)| \leqslant V/2$. Therefore,

$$\frac{1}{V} \sum_{i=1}^{V} \mathbb{I}_{\{\widehat{f}_i(x)\neq y\}} = \frac{V - |\mathcal{C}_x(y)|}{V} \geqslant \frac{1}{2} \ .$$

Thus

$$\widehat{f}^{\mathrm{mv}}(x) \neq y \implies \frac{1}{V}\sum_{i=1}^{V} \mathbb{I}_{\{\widehat{f}_i(x)\neq y\}} \geqslant \frac{1}{2} \ .$$

Hence, for any $y \in \mathcal{Y}$,

$$\mathbb{I}_{\{\widehat{f}^{\mathrm{mv}}(x)\neq y\}} \leqslant \frac{2}{V}\sum_{i=1}^{V} \mathbb{I}_{\{\widehat{f}_i(x)\neq y\}} \ .$$

Taking expectations with respect to (x, y) yields $\mathcal{L}(\widehat{f}^{\mathrm{mv}}) \leqslant 2V^{-1}\sum_{i=1}^{V}\mathcal{L}(\widehat{f}_i)$. ■

We can now proceed with the proof of Theorem 3.4.5.

Proof The proof relies on a result by [76, Eq. (8.60), which is itself a consequence of Corollary 8.8], which holds true as soon as

$$\forall t \in \mathbb{S}, \qquad \mathrm{Var}\big(\mathbb{I}_{\{t(X)\neq Y\}} - \mathbb{I}_{\{s(X)\neq Y\}}\big) \leqslant \Big[w\big(\sqrt{\ell(s,t)}\big)\Big]^2 \tag{3.63}$$

for some nonnegative and nondecreasing continuous function w on $\mathbb{R}^+$, such that $x \mapsto w(x)/x$ is nonincreasing on $(0, +\infty)$ and $w(1) \geqslant 1$.

Let us first prove that assumption (3.63) holds true. On one hand, since $\mathcal{Y} = \{0, 1\}$, for any $t \in \mathbb{S}$,

$$\begin{aligned}\mathrm{Var}\big(\mathbb{I}_{\{t(X)\neq Y\}} - \mathbb{I}_{\{s(X)\neq Y\}}\big) &\leqslant \mathbb{E}[|\mathbb{I}_{\{t(X)\neq Y\}} - \mathbb{I}_{\{s(X)\neq Y\}}|^2] \\ &= \mathbb{E}[\mathbb{I}_{\{t(X)\neq s(X)\}}] = \mathbb{E}\big[|t(X) - s(X)|\big] \ . \end{aligned} \tag{3.64}$$

On the other hand, since we consider binary classification with the 0–1 loss, for any $t \in \mathbb{S}$ and $h > 0$,

$$\begin{aligned}
\ell(s,t) &= \mathbb{E}\big[|2\eta(X) - 1| \cdot |t(X) - s(X)|\big] && \text{by [38, Theorem 2.2]} \\
&\geqslant h\mathbb{E}\big[|t(X) - s(X)|\mathbb{I}_{\{|2\eta(X)-1|\geqslant h\}}\big] \\
&\geqslant h\mathbb{E}\big[|t(X) - s(X)| - \mathbb{I}_{\{|2\eta(X)-1|<h\}}\big] && \text{since } \|t - s\|_\infty \leqslant 1 \\
&\geqslant h\mathbb{E}\big[|t(X) - s(X)|\big] - rh^{\beta+1} && \text{by (MA).}
\end{aligned}$$

This lower bound is maximized by taking

$$h = h_* := \left(\frac{\mathbb{E}\big[|t(X) - s(X)|\big]}{r(\beta+1)}\right)^{\frac{1}{\beta}} \ ,$$

which belongs to $[0, 1]$ since $r \geqslant 1$ and $\mathbb{E}\big[|t(X) - s(X)|\big] \leqslant 1$. Thus, we obtain

$$\ell(s,t) \geqslant h_*\frac{\beta}{\beta+1}\mathbb{E}\big[|t(X) - s(X)|\big] = \frac{\beta}{(\beta+1)^{(\beta+1)/\beta}r^{1/\beta}}\mathbb{E}\big[|t(X) - s(X)|\big]^{(\beta+1)/\beta}$$

hence Eq. (3.64) leads to

$$\mathrm{Var}\big(\mathbb{I}_{\{t(X)\neq Y\}} - \mathbb{I}_{\{s(X)\neq Y\}}\big) \leqslant \mathbb{E}\big[|t(X) - s(X)|\big] \leqslant \frac{\beta+1}{\beta^{\beta/(\beta+1)}} r^{\frac{1}{\beta+1}} \ell(s,t)^{\frac{\beta}{\beta+1}} \leqslant 2r^{\frac{1}{\beta+1}} \ell(s,t)^{\frac{\beta}{\beta+1}} \ .$$

Therefore, Eq. (3.63) holds true with $w(u) = \sqrt{r_1} u^{\frac{\beta}{\beta+1}}$ and $r_1 = 2r^{\frac{1}{\beta+1}}$, which statisfies the required conditions. So, by [76, Eq. (8.60)], for any $\theta \in (0,1)$,

$$\mathbb{E}\big[\ell\big(s, \widehat{f}_T^{\,\mathrm{ho}}\big) \,|\, D_n^T\big] \leqslant \frac{1+\theta}{1-\theta} \inf_{m\in\mathcal{M}} \ell\big(s, \mathcal{A}_m(D_n^T)\big) + \frac{\delta_*^2}{1-\theta}\left[2\theta + \log(e|\mathcal{M}|)\left(\frac{1}{3} + \theta^{-1}\right)\right] \tag{3.65}$$

where δ_* is the positive solution of the fixed-point equation $w(\delta_*) = \sqrt{n_v}\delta_*^2$, that is $\delta_*^2 = (r_1/n_v)^{\frac{\beta+1}{\beta+2}}$. Taking expectations with respect to the training data D_n^T, we obtain

$$\mathbb{E}\big[\ell(s, \widehat{f}_T^{\,\mathrm{ho}})\big] \leqslant \frac{1+\theta}{1-\theta}\mathbb{E}\left[\inf_{m\in\mathcal{M}} \ell\big(s, \mathcal{A}_m(D_n^T)\big)\right] + \frac{2r^{\frac{1}{\beta+2}}}{1-\theta} \frac{2\theta + \log(e|\mathcal{M}|)\left(\frac{1}{3}+\theta^{-1}\right)}{n_v^{\frac{\beta+1}{\beta+2}}} \ .$$

Under assumptions (3.2), $\mathbb{E}\big[\ell(s, \widehat{f}_T^{\,\mathrm{ho}})\big]$ and $\mathbb{E}\big[\mathcal{L}(\widehat{f}_T^{\,\mathrm{ho}})\big]$ do not depend on $T \in \mathcal{T}$ (they only depend on T through its cardinality n_t).

Now, by Proposition 3.10.1 applied to $(\widehat{f}_T^{\,\mathrm{ho}})_{T\in\mathcal{T}}$,

$$\mathbb{E}\big[\ell(s, \widehat{f}_{\mathcal{T}}^{\,\mathrm{mv}})\big] \leqslant 2\mathbb{E}\big[\ell(s, \widehat{f}_{T_1}^{\,\mathrm{ho}})\big] \leqslant 2\frac{1+\theta}{1-\theta}\mathbb{E}\left[\inf_{m\in\mathcal{M}} \ell\big(s, \mathcal{A}_m(D_n^T)\big)\right] + \frac{4r^{\frac{1}{\beta+2}}}{1-\theta} \frac{2\theta + \log(e|\mathcal{M}|)\left(\frac{1}{3}+\theta^{-1}\right)}{n_v^{\frac{\beta+1}{\beta+2}}} \ .$$

Taking $\theta = 1/5$ leads to the result. ■

Chapter 4

Aggregated hold out for sparse linear regression with a robust loss function

4.1 Introduction

From the statistical learning point of view, linear regression is a risk-minimization problem wherein the aim is to minimize the average *prediction error* $\phi(Y - \theta^T X)$ on a new, independent data-point (X, Y), as measured by a *loss function* ϕ. When $\phi(x) = x^2$, this yields classical least-squares regression; however, Lipschitz-continuous loss functions have better robustness properties and are therefore preferred in the presence of heavy-tailed noise, since they require fewer moment assumptions on Y [34, 56]. In general, substracting the risk of the (distribution-dependent) optimal predictor yields a measure of performance for estimators, called the *excess risk*, which coïncides with the L^2 norm in the least-squares case.

In the high-dimensional setting, where $X \in \mathbb{R}^d$ with potentially $d >> n$, minimizing the risk over all θs is impossible; some assumptions must be made. A popular approach is to suppose that only a small number k_* of covariates are relevant to the prediction of Y, so that θ may be sought among the *sparse* vectors with less than k_* non-zero components. Estimators which target such problems include the Lasso and its variants, LARS, stagewise regression and the classical greedy procedures of stepwise regression. In the robust setting, variants of the Lasso with robust loss functions have been investigated by a number of authors [64, 91, 32, 111].

Such methods generally introduce a free hyperparameter that regulates the "sparsity" of the estimator; sometimes this is directly the number of non-zero components, as in stepwise procedures, sometimes not, as in the Lasso. In any

case, the user is left with the problem of calibrating this hyperparameter.

Several goals are conceivable for a hyperparameter selection method, such as support recovery or estimation of a "true" underlying regression coefficient. From a prediction perspective, hyperparameters should be chosen so as to minimize the risk, and a good method should approach this minimum. As a consequence, the proposed data-driven choice of hyperparameter should allow the estimator to attain all known convergence rates without any a priori knowledge, effectively adapting to the difficulty of the problem.

For the Lasso and some variants, such as the fused Lasso, Zou, Wang, Tibshirani and coauthors have proposed [120] and investigated [110, 101] a method based on Mallow's C_p and estimation of the "degrees of freedom of the Lasso". However, consistency of this method has only been proven [110] in an asymptotic where the dimension is fixed while n grows, hence not the setting considered here. Moreover, the method depends on specific properties of the Lasso, and may not be readily applicable to other sparse regression procedures.

A much more widely applicable procedure is to choose the hyperparameter by cross-validation. For the Lasso, this approach has been recommended by Tibshirani [99], van de Geer and Lederer [104] and Greenshtein [47], among many others. More generally, cross-validation is the default method for calibrating hyperparameters in practice. For exemple, R implementations of the elastic net (package glmnet), LARS (package lars) and the huberized lasso (package hqreg) all incorporate a cross-validation subroutine to automatically choose the hyperparameter.

Theoretically, cross-validation has been shown to perform well in a variety of settings [3]. For cross-validation with one split, also known as the hold-out, and for a bagged variant of v-fold cross-validation [69], some general oracle inequalities are available in least squares regression [76, Corollary 8.8] [114] [69]. However, they rely on uniform boundedness assumptions on the estimators which may not hold in high-dimensional linear regression. For the more popular V-fold procedure, results are only available in specific settings. Of particular interest here is the article [83] which proves oracle inequalities for linear model selection in least squares regression, since linear model selection is very similar to sparse regression (the main difference being that in sparse regression, the "models" are not fixed a priori but depend on the data). This suggests that similar results could hold for sparse regression.

However, in the case of the Lasso at least, no such theoretical guarantees exist, to the best of my knowledge. Some oracle inequalities [69, 82] and also fast rates [53, Theorem 1] have been obtained, but only under very strong assumptions: [69] assumes that X is log-concave, [82] that X is a gaussian vector, and [53, Theorem 1] assumes that there is a true model and that the variance-covariance matrix is diagonal dominant. In contrast, there are also theorems [31, 33] [53, Theorem 2]

which make much weaker distributional assumptions but only prove convergence of the risk at the "slow" rate $\mathcal{O}(\sqrt{\frac{s_* \log p}{n}})$ or slower. Though this rate is minimax [31], a hyperparameter selection method should adapt also to the favorable cases where the Lasso converges faster; these results do not show that CV has this property.

Thus, the theoretical justification for the use of standard CV in sparse regression is somewhat lacking. In fact, two of the articles mentioned above do not study standard CV applied to the Lasso but introduce a variant; a bagged CV in [69] and the aggregation of two hold-out predictors in [31]. In practice too, there is reason to consider alternatives to CV-based hyperparameter selection in sparse regression: sparse estimators are unstable, and selecting only one estimator can result in arbitrarily ignoring certain variables among a correlated group with similar predictive power [115]. For the Lasso, these difficulties have motivated researchers to introduce several aggregation schemes, such as the Bolasso [8], stability selection [79], the lasso-zero [36] and the random lasso [112], which are shown to have some better properties than the standard Lasso.

Since aggregating the Lasso seems to be advantageous, it seems logical to consider aggregation rather than cross-validation to handle the free hyperparameters. In this article, I consider the application to sparse regression of the aggregated hold-out procedure. Aggregated hold-out (agghoo) is a general aggregation method which mixes cross-validation with bagging. It is an alternative to cross-validation, with a comparable level of generality. In a previous article with Sylvain Arlot and Matthieu Lerasle (Chapter 3), we formally defined and studied Agghoo, and showed empirically that it can improve on cross-validation when calibrating the level of regularization for kernel regression. Though we came up with the name and the general mathematical definition, Agghoo has already appeared in the applied litterature in combination with sparse regression procedures [54], among others [107], under the name "CV + averaging" in this case.

In the present article, the aim is to study the application of Agghoo to sparse regression with a robust loss function. Theoretically, assuming an $L^\infty - L^2$ norm inequality to hold on the set of sparse linear predictors, it is proven that Agghoo satisfies an asymptotically optimal oracle inequality. This result applies also to cross-validation with one split (the so-called hold-out), yielding a new oracle inequality which allows norms of the sparse linear predictors to grow polynomially with the sample size. Empirically, Agghoo is compared to cross-validation in a number of simulations, which investigate the impact of correlations in the design matrix and sparsity of the ground truth on the performance of aggregated hold-out and cross-validation. Agghoo appears to perform better than cross-validation when the number of non-zero coefficients to be estimated is not much smaller than the sample size. The presence of confounders correlated to the predictive variables also favours Agghoo relative to cross-validation.

4.2 Setting and Definitions

The problem of non-parametric regression is to infer a predictor $t : \mathcal{X} \to \mathbb{R}$ from a dataset $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ of pairs, where $X_i \in \mathcal{X}$ and $Y_i \in \mathbb{R}$. The pairs will be assumed to be i.i.d, with joint distribution P. The prediction error made at a point $(x, y) \in \mathcal{X} \times \mathbb{R}$ is measured using a non-negative function of the residual $\phi(y - t(x))$. The global performance of a predictor is assessed on a new, independent data point (X, Y) drawn from the same distribution P using the risk $\mathcal{L}(t) = \mathbb{E}[\phi(Y - t(X))]$. The optimal predictors s are characterized by $s(x) \in \operatorname{argmin}_u \mathbb{E}[\phi(Y - u)|X = x]$ a.s. The risk of any optimal predictor is (in general) a non-zero quantity which characterizes the intrinsic amount of "noise" in Y unaccounted for by the knowledge of X. A predictor t can be compared with this benchmark by using the *excess risk* $\ell(s, t) = \mathcal{L}(t) - \mathcal{L}(s)$. Taking $\phi(x) = x^2$ yields the usual least-squares regression, where $s(x) = \mathbb{E}[Y|X = x]$ and $\ell(s, t) = \|s - t\|^2_{L^2(X)}$. However, the least-squares approach is known to suffer from a lack of robustness. For this reason, in the field of robust statistics, a number of alternative loss functions are used. One popular choice was introduced by Huber [55].

Definition 4.2.1 *Let $c > 0$. Huber's loss function is $\phi_c(u) = \frac{u^2}{2}\mathbb{I}_{|u| \leqslant c} + c\left(|u| - \frac{c}{2}\right)\mathbb{I}_{|u| > c}$.*

When $c \to +\infty$, ϕ_c converges to the least-squares loss. When $c \to 0$, $\frac{1}{c}\phi_c$ converges to the absolute value loss $x \to |x|$ of median regression. Thus, the c parameter allows a trade-off between robustness and approximation of the least squares loss.

The rest of the article will focus on sparse linear regression with the loss function ϕ_c. Thus, notations s, $\ell(s, t)$ and $\mathcal{L}$ are to be understood with respect to ϕ_c.

4.2.1 Sparse linear regression

With finite data, it is impossible to solve the optimization problem $\min \mathcal{L}(t)$ over the set of all predictors t. Some modeling assumptions must be made to make the problem tractable. A popular approach is to build a finite set of features $(\psi_j(X))_{1 \leqslant j \leqslant d}$ and consider predictors that are linear in these features: $\exists \theta \in \mathbb{R}^d, \forall x \in \mathcal{X}, t(x) = \sum_{j=1}^d \theta_j \psi_j(x)$. This is equivalent to replacing $X \in \mathcal{X}$ with $\tilde{X} = (\psi_j(X))_{1 \leqslant j \leqslant d} \in \mathbb{R}^d$ and regressing Y on $\tilde{X}$. For theoretical purposes, it is thus equivalent to assume that $\mathcal{X} = \mathbb{R}^d$ for some d and predictors are linear: $t(x) = \theta^T x$.

As the aim is to reduce the average prediction error $\mathcal{L}(t)$, a logical way to choose θ is by *empirical risk minimization*:

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \phi_c(Y_i - \theta^T X_i).$$

Empirical risk minimization works well when $d << n$ but will lead to overfitting in large dimensions [106]. Sparse regression attempts instead to locate a "good" subset of variables in order to optimize risk for a given model dimension. Lasso penalization [99] is now a standard method of achieving sparsity. The specific version of the Lasso which we consider here is given by the following Definition.

Definition 4.2.2 *Let $n \in \mathbb{N}$ and let $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n}$ be a dataset such that $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$ for all $i \in [|1; n|]$ and some $d \in \mathbb{N}$. Let ϕ_c be the Huber loss defined in Definition 4.2.1. For any $\lambda > 0$, let*

$$\hat{\mathcal{C}}(\lambda) = \underset{(q,\theta)\in\mathbb{R}^{d+1}:\|\theta\|_1 \leqslant n^\alpha}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \phi_c\big(Y_i - q - \theta^T X_i\big) + \lambda \|\theta\|_1 \quad \textit{and}$$

$$(\hat{q}(\lambda), \hat{\theta}(\lambda)) \in \underset{(q,\theta)\in\hat{\mathcal{C}}(\lambda)}{\operatorname{argmin}} \big|q + < \theta, \frac{1}{n}\sum_{i=1}^{n} X_i >\big|. \tag{4.1}$$

Now let

$$\mathcal{A}(\lambda)(D_n) : x \to \hat{q}(\lambda) + \hat{\theta}(\lambda)^T x.$$

The restriction $\|\theta\|_1 \leqslant n^\alpha$ ensures that the solution does not become too large when the design matrix is ill-conditioned. It can be seen that the effect of this restriction is, potentially, to truncate the Lasso solution path at the value of λ at which the bound is attained (i.e, $\hat{\theta}(\cdot)$ becomes constant for smaller values of λ). Without this, the ℓ^1 norm of Lasso solutions is upper bounded by the minimal ℓ^1 norm $\hat{r}$ of an empirical risk minimizer on the whole set of variables. Hence, if $\hat{r} \leqslant n^\alpha$, Definition 4.2.2 coïncides with the huberized lasso along the whole regularization path. In the least squares case, [53] discuss conditions under which $\|\hat{r}\|_{L^4} \leqslant n^{\frac{1}{4}}$, which ensures that for $\alpha > \frac{1}{4}$, $\hat{r} < n^\alpha$ with high probability. It seems reasonable to expect a similar result to hold true in the case of the huberized lasso. However, rather than make further technical assumptions on the design to make sure of this, it seems simpler to introduce this slight modification to the standard definition of the huberized lasso, which may even be statistically beneficial, since it diminishes the variance by restricting the hypothesis space.

A suitable choice of α should guarantee that an optimal excess risk $\mathbb{E}\left[\ell(s, q + \langle\theta, \cdot\rangle)\right]$ can be obtained for some θ such that $\|\theta\|_1 \leqslant n^\alpha$. For example, if the features X form an orthonormal set and $Y \in L^2$, then the least-squares optimal coefficient θ_* belongs to $\left\{\theta : \|\theta\|_2 \leqslant \sqrt{\mathbb{E}[Y^2]}\right\}$. Assume that only sparse predictors θ^T, with less than n non-zero components, are considered. Since for such θ, $\|\theta\|_1 \leqslant \sqrt{n}\,\|\theta\|_2$, it is reasonable to restrict the optimization to the set $\{\theta : \|\theta\|_1 \leqslant n^\alpha\}$ for some $\alpha > \frac{1}{2}$.

The intercept q is left unpenalized in definition 4.2.2, as is usually the case in practice [119]. Equation (4.1) is a tiebreaking rule which is required for the proof to work.

4.2.2 Hyperparameter tuning

The zero-norm of a vector θ is the integer $\|\theta\|_0 = |\{i : \theta_i \neq 0\}|$. Many sparse estimators, such as best subset or forward stagewise, are directly parametrized by their desired zero-norm, which must be chosen by the practitioner. It controls the "complexity" of the estimator, and hence the bias-variance tradeoff. In the case of the standard Lasso (Definition 4.2.2 with $\phi(x) = x^2$), Zou, Hastie and Tibshirani [120] showed that $\left\|\hat{\theta}(\lambda)\right\|_0$ is an unbiased estimator of the "degrees of freedom" of the estimator $\mathcal{A}(\lambda)$. As a consequence, [120] suggests reparametrizing the lasso by its zero-norm. Applying their definition to the present setting yields the following.

Definition 4.2.3 *For any dataset D_n, let $(\hat{q}, \hat{\theta})$ be given by Definition 4.2.2, equation (4.1) . Let $M \in \mathbb{N}$ and $(\lambda_m)_{1 \leqslant m \leqslant M}$ be the finite decreasing sequence at which the sets $\{i : \hat{\theta}(\lambda)_i \neq 0\}$ change. Let $\lambda_0 = +\infty$. For any $k \in \mathbb{N}$ let*

$$\hat{m}_k^{last} = \max\left\{m \in \mathbb{N} | \left\|\hat{\theta}(\lambda_m)\right\|_0 = k\right\},$$

with the convention $\max \emptyset = 0$. Let then

$$\mathcal{A}_k(D_n) = \mathcal{A}\left(\lambda_{\hat{m}_k^{last}}\right)(D_n). \tag{4.2}$$

More generally, consider sequences $(\mathcal{A}_k)_{k\in\mathbb{N}}$ of linear regression estimators $\mathcal{A}_k : D_n \to \left(x \to \hat{q}_k(D_n) + \langle \hat{\theta}_k(D_n), x\rangle\right)$, such that the following hypothesis holds.

Hypothesis 4.2.1 *For any $n \in \mathbb{N}$, let $D_n \sim P^{\otimes n}$ denote a dataset of size n. Assume that*

1. *Almost surely, for all $k \in [|1; n|]$, $\left\|\hat{\theta}_k(D_n)\right\|_0 \leqslant k$.*
2. *There exist L, α such that $\forall n \in \mathbb{N}, \mathbb{E}\left[\sup_{1\leqslant k\leqslant n}\left\|\hat{\theta}_k(D_n)\right\|_1\right] \leqslant Ln^\alpha$.*
3. *For all $k \in [|1; n|]$, $\hat{q}_k(D_n) \in \operatorname{argmin}_{q\in\hat{Q}\left(D_n, \hat{\theta}_k(D_n)\right)}\left|q + \langle\hat{\theta}_k(D_n), \frac{1}{n}\sum_{i=1}^n X_i\rangle\right|$, where $\hat{Q}(D_n, \theta) = \operatorname{argmin}_{q\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n \phi_c\left(Y_i - \langle\theta, X_i\rangle - q\right)$.*

These hypotheses hold for the reparametrized Lasso given by definition 4.2.2 and 4.2.3, by construction.

Moreover, Condition 1 is naturally satisfied by such sparse regression methods as forward stepwise and best subset. Condition 2 can be enforced by restricting the set of θs over which the optimization is conducted, similarly to Definition 4.2.2. Condition 3 states that the intercept q is chosen by empirical risk minimization, with a specific tie-breaking rule in case the minimum is not unique.

4.2.3 Aggregated hold out applied to the zero-norm parameter

The tuning of the zero-norm k is important to ensure good prediction performance by optimizing the bias-variance tradeoff. For the Lasso and other methods based on empirical risk minimization, such as forward stepwise, there is little interest in considering values of $k > n$, since n non-zero coefficients suffice for perfect interpolation of the (X_i, Y_i) and yield an empirical risk of 0. Practicioners may also want to impose additional limitations on the zero-norm in order to reduce the computational load or improve interpretability. For this reason, we consider the problem of selecting the zero-norm among the K_n first values, where $K_n \leqslant n$. This article investigates the use of Agghoo in this context, as an alternative to cross-validation. Agghoo is a general hyperparameter aggregation method which was defined in Chapter 3, in a general statistical learning context. Let us briefly recall its definition in the present setting. For a more detailed introductory discussion of this procedure, we refer the reader to Chapter 3. To simplify notations, fix a collection $(\hat{q}_k, \hat{\theta}_k)_{1\leqslant k\leqslant K}$ of linear regression estimators. First, we need to define *hold-out* selection of the zero-norm parameter.

Definition 4.2.4 *Let $D_n = (X_i, Y_i)_{1\leqslant i\leqslant n}$ be a dataset. For any $T \subset \{1, \dots, n\}$, denote $D_n^T = (X_i, Y_i)_{i\in T}$. Let then*

$$\hat{k}_T(D_n) = \min \operatorname*{argmin}_{1\leqslant k\leqslant K} \frac{1}{|T^c|} \sum_{i\notin T} \phi_c \left(Y_i - \hat{q}_k(D_n^T) - \langle \hat{\theta}_k(D_n^T), X_i \rangle \right).$$

Using the hyperparameter $\hat{k}_T(D_n)$ together with the dataset D_n^T to train a linear regressor yields the hold-out predictor

$$\widehat{f}_T^{\,\mathrm{ho}}(D_n) : x \to \hat{q}_{\hat{k}_T(D_n)}(D_n^T) + \langle \hat{\theta}_{\hat{k}_T(D_n)}(D_n^T), x \rangle.$$

Aggregation of hold-out predictors is performed in the following manner.

Definition 4.2.5 *Let $\mathcal{T} \subset \mathcal{P}(\{1, \dots, n\})$. Let:*

$$\hat{\theta}_{\mathcal{T}}^{ag} = \frac{1}{|\mathcal{T}|} \sum_{T\in\mathcal{T}} \hat{\theta}_{\hat{k}_T(D_n)}(D_n^T)$$
$$\hat{q}_{\mathcal{T}}^{ag} = \frac{1}{|\mathcal{T}|} \sum_{T\in\mathcal{T}} \hat{q}_{\hat{k}_T(D_n)}(D_n^T).$$

The Agghoo predictor is the linear regressor:

$$\widehat{f}_{\mathcal{T}}^{\,\mathrm{ag}}(D_n) : x \to \hat{q}_{\mathcal{T}}^{ag} + \langle \hat{\theta}_{\mathcal{T}}^{ag}, x \rangle.$$

Thus, Agghoo also yields a linear predictor, which means that it can be efficiently evaluated on new data. If the $\hat{\theta}_{\hat{k}_T(D_n)}$ have similar support, $\hat{\theta}^{ag}_{\mathcal{T}}$ will also be sparse: this will happen if the hold-out reliably identifies a true model. On the other hand, if the supports have little overlap, the Agghoo coefficient will lose sparsity, but it can be expected to be more stable and to perform better.

The linear regressors $x \to \hat{q}_{\hat{k}_T(D_n)}(D_n^T) + \langle \hat{\theta}_{\hat{k}_T(D_n)}(D_n^T), x\rangle$ aggregated by Agghoo are only trained on part of the data. This subsampling (typically) decreases the performance of each individual estimator, but combined with aggregation, it may stabilize an unstable procedure and improve its performance, similarly to bagging.

An alternative would be to *retrain* each regressor on the whole data-set D_n, yielding the following procedure, which we call "Aggregated cross-validation" (Agcv).

Definition 4.2.6 *Let $\mathcal{T} \subset \mathcal{P}(\{1,\dots,n\})$. Let:*

$$\hat{\theta}^{acv}_{\mathcal{T}} = \frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}} \hat{\theta}_{\hat{k}_T(D_n)}(D_n)$$
$$\hat{q}^{acv}_{\mathcal{T}} = \frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}} \hat{q}_{\hat{k}_T(D_n)}(D_n).$$

The Agcv predictor is the linear regressor:

$$\widehat{f}^{\,\mathrm{acv}}_{\mathcal{T}}(D_n) : x \to \hat{q}^{acv}_{\mathcal{T}} + \langle \hat{\theta}^{acv}_{\mathcal{T}}, x\rangle.$$

Agghoo is easier to study theoretically than Agcv due to the conditional independence: $\left(\hat{\theta}_k\left(D_n^T\right)\right)_{1\leqslant k\leqslant n_t} \perp\!\!\!\perp \hat{k}_T(D_n) \,\middle|\, D_n^T$. For this reason, the theoretical section will focus on Agghoo, while in the simulation study, both Agghoo and Agcv will be considered.

4.3 Theoretical results

Let $n \in \mathbb{N}$ and $D_n = (X_i, Y_i)_{1\leqslant i\leqslant n}$ denote an i.i.d dataset with common distribution P. In this section, we make the following assumption on $\mathcal{T}$: there is an integer $n_t < n$ such that

$$\begin{aligned} &\mathcal{T} \subset \{T \subset \{1,\dots,n\} : |T| = n_t\} \\ &\mathcal{T} \text{ independent from } D_n \ . \end{aligned} \tag{4.3}$$

Independence of $\mathcal{T}$ from D_n ensures that for $T \in \mathcal{T}$, D_n^T is also iid with distribution P. The assumption that $\mathcal{T}$ contain sets of equal size ensures that the pairs $\hat{q}_{\hat{k}_T(D_n)}(D_n^T), \hat{\theta}_{\hat{k}_T(D_n)}(D_n^T)$ are equidistributed for $T \in \mathcal{T}$. Most of the data partitioning procedures used for cross-validation satisfy hypothesis (4.3), including

leave-p-out, V-fold cross-validation (with $n - n_t = n_v = n/V$) and Monte-Carlo cross-validation [3].

In the following, we will use the notion of support of a random variable, for which we introduce the following definition.

Definition 4.3.1 *Let X be a random variable belonging to $\mathbb{R}^d$ for some $d \in \mathbb{N}$. Then the support of X is*

$$supp(X) = \{x \in \mathbb{R}^d : \forall \varepsilon > 0, \mathbb{P}\left(\|x - X\| \leqslant \varepsilon\right) > 0\}.$$

The support is closed and has full measure: $\mathbb{P}(X \in supp(X)) = 1$.

When Agghoo is used on a collection $(\mathcal{A}_k)_{1\leqslant k\leqslant K}$ of linear regression estimators satisfying Hypothesis (4.2.1), such as the Lasso parametrized by the number of non-zero coefficients, as in Definition 4.2.3, the following Theorem applies.

Theorem 4.3.2 *Let $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ be random variables with joint distribution P. Let $D_n = (X_i, Y_i)_{1\leqslant i\leqslant n} \sim P^{\otimes n}$ be a dataset of size n. Let $n_v = n - n_t$, where n_t is given by assumption* (4.3).

Assume that for some function s minimizing the risk $\mathbb{E}[\phi_c(Y - s(X))]$, there exists $\eta > 0$ such that almost surely,

$$\mathbb{P}\left[|Y - s(X)| \leqslant \frac{c}{2} \,\middle|\, X\right] \geqslant \eta. \tag{4.4}$$

Let $\bar{X} = X - \mathbb{E}[X]$ and let $supp(\bar{X})$ be its support (in the sense of Definition 4.3.1). Let $R = \sup_{x\in supp(\bar{X})} \|x\|_\infty$. For any $K \in \{1, \dots, n_t\}$, let

$$\kappa(K) = \sup_{v\neq 0, \|v\|_0 \leqslant 2K} \frac{\left\|\langle \bar{X}, v\rangle\right\|_{L^\infty}}{\left\|\langle \bar{X}, v\rangle\right\|_{L^2}}. \tag{4.5}$$

If $b_0 > 1$ and $K \in \{1, \dots, n_t\}$ are such that

$$\kappa(K) \leqslant \frac{\eta}{8}\sqrt{\frac{n_v}{8b_0 \log n_t}}, \tag{4.6}$$

applying Agghoo to a collection $(\mathcal{A}_k)_{1\leqslant k\leqslant K}$ of linear regression estimators which satisfies hypothesis (4.2.1) *yields the following oracle inequality.*

For any $\theta \in \left[\frac{1}{\sqrt{b_0}}; 1\right]$,

$$(1-\theta)\mathbb{E}\left[\ell(s, \widehat{f}^{\mathrm{ag}}_{\mathcal{T}})\right] \leqslant (1+\theta)\mathbb{E}\left[\min_{1\leqslant k\leqslant K} \ell(s, \mathcal{A}_k(D_{n_t}))\right] + 24\theta b_0 \frac{c \log n_t}{\eta n_v}\left[c + \frac{cK}{n_t^{\theta^2 b_0}} + \frac{16KLR}{n_t^{\theta^2 b_0 - \alpha}}\right]. \tag{4.7}$$

Theorem 4.3.2 is proved in appendix 4.6. It is, to the best of my knowledge, the first theoretical guarantee on hyperparameter selection for the huberized Lasso. Theorem 4.3.2 compares the excess risk of Agghoo to that of the best linear predictor in the collection $\mathcal{A}_k(D_{n_t})$, trained on a subset of the data of size n_t. That n_t appears in the oracle instead of n is a limitation, but it is logical, since estimators aggregated by Agghoo are only trained on samples of size n_t. Typically, the excess risk increases at most by a constant factor when a dataset of size n is replaced by a subset of size τn, and this constant tends to 1 as $\tau \to 1$. This allows to take n_v of order n ($n_v = (1-\tau)n$), while losing only a constant factor in the oracle term.

Taking $|\mathcal{T}| = 1$ in Theorem 4.3.2 yields an oracle inequality for the hold-out, which is also cross-validation with one split. Compared to previously known oracle inequalities for the hold-out, Theorem 4.3.2 distinguishes itself by only requiring some polynomial upper bound on $\|\theta\|_1$ and $\left\|\bar{X}\right\|_\infty$, instead of a uniform upper bound on some norm independent of n. Indeed, the prevailing approach to proving oracle inequalities for the hold-out (applied to the Lasso by Lecué [69]) uses a *margin assumption* which requires a uniform upper bound on the loss function [69, Assumption (A)], leading to a bound on $\langle \hat{\theta}_k, \bar{X} \rangle$. Theorem 4.3.2 relaxes this constraint by exploiting an $L^\infty - L^2$ norm inequality (equation(4.6)).

In order to fulfill its purpose, Theorem 4.3.2 should imply that Agghoo performs as well as the best of the sparse estimators $\mathcal{A}_k(D_{n_t})$, at least asymptotically. Here, we are interested in the high-dimensional, non-parametric case where the dimension grows with the amount of data n as a power of n. More precisely, consider a sequence of problems $(Y, \psi_n(X_0))$ where $X_0 \in \mathcal{X}$ and $\psi_n : \mathcal{X} \to \mathbb{R}^{d_n}$, where $d_n > n^\beta$ for some $\beta > 0$. Assume that $R = 1$, which can be achieved by renormalizing ψ_n - as long as $\|\psi_n(X_0)\|_{\infty, L^\infty}$ grows at most polynomially in n, this simply yields an increase in L and α. If additionally, equation (4.6) holds with $b_0 > \alpha + 1$, choosing $\theta \in \left(\sqrt{\frac{1+\alpha}{b_0}}; 1\right)$ yields a remainder term of order $\mathcal{O}(\frac{\log n}{n})$ in equation (4.7). By comparison, in the least squares setting the minimax excess risk for sparse regression with k_n predictive covariates among a total of d_n is of order $\frac{k_n \log\left(\frac{d_n}{k_n}\right)}{n}$ [113]. For large c, the huber loss approximates the least squares loss, so it is reasonable to expect this lower bound to apply also in huber regression. Assuming that the minimax is attained, the remainder term of equation (4.7) is negligible compared to the oracle whenever $k_n \to +\infty$, i.e when the problem is non-parametric.

Now if for any n, assumption (4.6) holds with $X = \psi_n(X_0)$ and $b_0 = b_{1,n} \to +\infty$, then Theorem 4.3.2 yields an asymptotically optimal oracle inequality. More precisely, applying Theorem 4.3.2 with $b_0 = b_{0,n} = b_{1,n} \wedge \sqrt{\frac{k_n}{1+\alpha}}$ (for which assumption (4.6) also holds) and $\theta = \theta_n = \sqrt{\frac{1+\alpha}{b_{0,n}}} \wedge 1$ yields a bounded term in the square

brackets of equation (4.7). Moreover, $\theta_n b_{0,n} \leqslant \sqrt{k_n}$. This implies that

$$\limsup_{n\to+\infty} \frac{\mathbb{E}\Big[\ell(s, \widehat{f}_{\mathcal{T}}^{\text{ag}})\Big]}{\mathbb{E}\Big[\min_{1\leqslant k\leqslant K_{n_t}} \ell(s, \mathcal{A}_k(D_{n_t}))\Big]} \leqslant 1$$

since $\frac{\sqrt{k_n}\log n}{n} = o\left(\mathbb{E}\Big[\min_{1\leqslant k\leqslant K_{n_t}} \ell(s, \mathcal{A}_k(D_{n_t}))\Big]\right)$.

To summarize, Theorem 4.3.2 proves an asymptotically optimal oracle inequality whenever

- Equation (4.4) holds.
- $\left\|\|\psi_n(X_0)\|_\infty\right\|_{L^\infty}$ grows at most polynomially in n.
- The problem is non-parametric, i.e the risk converges slower than $\frac{\log n}{n}$.
- For any $b_0 > 1$, equation (4.6) holds for all n large enough.

Equation (4.4) is specific to the Huber loss: it requires the conditional distribution of the residual $Y - s(x)$ to put sufficient mass in a region where the huber function is quadratic. If $Y = s_0(X_0) + \sigma\varepsilon$, with ε independent of X_0 and if ψ_n is injective, then $s(X) = s_0(X_0)$ and η depends only on c, σ and the distribution of ε. In particular, it is constant with respect to n. Moreover, if the Huber parameter c is proportional to σ, then η is independent also of σ and the remainder term of equation (4.7) is proportional to σ^2, as in least-squares regression. Injectivity of ψ_n will typically hold for nonparametric function bases (trigonometric, splines of degree greater than 1) as soon as n is large enough.

The norm inequality (Equation (4.6)) requires more clarification. It is worth giving some background on such hypotheses, which are relatively classical in the model selection litterature. They were introduced by Birgé and Massart in the context of least-squares density estimation [17, Section 3.1], where loosely speaking, it is assumed that $\kappa(K) = \mathcal{O}(\sqrt{K})$. A similar assumption was made by Arlot and Lerasle [5, Section 3.3, hypothesis (H1)] to prove oracle inequalities for cross-validation, also in least squares density estimation. In the regression setting, [83] proves an oracle inequality for cross-validation based on the assumption that the models have a "strongly localized basis", which implies in particular that $\kappa(K) = \mathcal{O}(\sqrt{K})$ when the model collection consists of all $\left\{\langle\theta, (\bar{X}_i)_{i\in I}\rangle : \theta \in \mathbb{R}^{|I|}\right\}$ for $I \subset [|1; n|]$.

These assumptions have been shown to hold for several standard model collections. In particular, in the regression setting, [94, Lemma 7] implies that linear models m consisting of piecewise-polynomial functions on an interval partition $(I_i)_{1\leqslant i\leqslant d}$ satisfy $\|\cdot\|_{L^\infty(X)} \leqslant C\sqrt{d}\,\|\cdot\|_{L^2(X)}$, provided that $\min_i \mathbb{P}(X \in I_i) \geqslant \frac{a}{d}$ for

some constant a and that the distribution of X has a lower-bounded density with respect to the Lebesgue measure.

The assumption (4.6) differs from its analogs in the model selection litterature in two ways: first, because of the sparse variable selection setting, the "models" which give rise to $\kappa(K)$ are the $\left\{\langle\theta, (\bar{X}_i)_{i\in I}\rangle : \theta \in \mathbb{R}^{|I|}\right\}$ for $I \subset [|1;d|]$ of cardinality $|I| = K$. Second, because an additional intercept term is included, the feature vector X has to be replaced by its centered version $\bar{X}$ in the definition of $\kappa(K)$.

We give below two simple examples where the hypotheses of Theorem 4.6 hold. In the case where the variables are independent and binary valued, we have the following.

Corollary 4.3.3 *Let $X = (X_1, \dots, X_p)$ where the X_i are independent Bernoulli random variables with parameters $p_i \in [0;1]$ Then*

$$\kappa(K) \leqslant \sqrt{\frac{2K}{\min_{1\leqslant i\leqslant p} p_i(1-p_i)}}.$$

Corollary 4.3.3 is proved in appendix 4.7.1. In the setting it describes, Assumption (4.6) is equivalent to choosing K of order $\frac{n_t}{\log n_t}$ or less, provided that the classes are well-balanced ($p_i \in [\varepsilon, 1-\varepsilon]$).

Despite the fact that Theorem 4.3.2 is formulated in the setting of sparse linear regression, it can also be applied to other regression problems, such as adaptive piecewise constant regression. In that case, the equivalent of the zero-norm of a vector is the number of discontinuities of a piecewise constant function, as can be seen from the following definition.

Definition 4.3.4 *Let $(I_j)_{1\leqslant j\leqslant d}$ denote a partition of $\mathbb{R}$ into disjoint intervals, indexed such that $\forall j, \sup I_j = \inf I_{j+1}$. For any $u \in \mathbb{R}^d$, let $t_u = \sum_j u_j \mathbb{I}_{I_j}$. Then the number of jumps of the piecewise constant function t_u is*

$$k(u) = \left|\{j \in [|1;d|] : u_{j+1} \neq u_j\}\right|,$$

and we say that t_u has k jumps *if and only if $k(u) = k$. Let $(j_r(u))_{0\leqslant r\leqslant k(u)}$ denote the ordered sequence of jump indices, i.e $(j_r(u))_{1\leqslant r\leqslant k(u)}$ is increasing, $j_0(u) = 0$ and*

$$\{j_r(u) | 1 \leqslant r \leqslant k(u)\} = \{j \in [|1;d|] : u_{j+1} \neq u_j\}.$$

For any $r \leqslant k(u)$, let $A_r(u) = \cup_{i=j_{r-1}(u)+1}^{j_r(u)} I_i$ be the largest intervals on which t_u is constant. Let $D_n = (U_i, Y_i)_{1\leqslant i\leqslant n}$ denote a dataset, with $U_i \in \mathbb{R}$ and $Y_i \in \mathbb{R}$.

Let now $(\hat{u}_k)_{0\leqslant k\leqslant d-1}$ denote a sequence of estimators such that $\hat{u}_k(D_n)$ has k jumps, and such that its coefficients $\hat{u}_{k,j}$ are obtained by empirical risk minimization on the minimal partition $(A_r(\hat{u}_k))_{1\leqslant r\leqslant k}$, i.e $\hat{u}_k \in \hat{C}_t(D_n, \hat{u}_k)$ where

$$\hat{C}_t(D_n,u) = \Big\{ u' \text{ s.t. } k(u') = k \text{ and } \forall r \in [|1;k|], j_r(u') = j_r(u) \\ \text{and } u'_{j_r(u')} \in \operatorname*{argmin}_{q\in\mathbb{R}} \sum_{i:X_i\in A_r(u)} \phi_c(Y_i - q) \Big\},$$

Assume also that the following tie-breaking rule applies:

$$\hat{u}_k \in \operatorname*{argmin}_{u\in\hat{C}_t(D_n,\hat{u}_k)} \left| \sum_{j=1}^{d} P_n(I_j) u_j \right|. \tag{4.8}$$

Estimators $\hat{u}_k$ which meet definition 4.3.4 can be obtained by a variety of model selection methods, including wavelet thresholding, empirical risk minimization over the set $\{u : k(u) = k\}$ [65] and the fused lasso [100] or total variation penalties [92, 19] (if the penalty is used only for estimating the change points).

Applied to such estimators, Theorem 4.3.2 allows to prove the following.

Proposition 4.3.5 *Consider the problem of tuning k so as to minimize the risk of the one-dimensional regression problem: $\mathbb{E}[\phi_c(Y - t_u(U))]$, where U is a random variable with distribution P. Assume that*

$$\forall x, \mathbb{P}\left[|Y - t_{u_*}(x)| \leqslant \frac{c}{2} \,\middle|\, X = x \right] \geqslant \eta, \tag{4.9}$$

where $u_* \in \operatorname{argmin}_{u\in\mathbb{R}^d} \mathbb{E}\left[\phi_c(Y - t_u(X))\right]$. Using the notations of definition 4.3.4, assume that:

$$\min_{1\leqslant j\leqslant d} P(I_j) \geqslant \frac{1536 \log^2 n_t}{\eta^2 n_v}. \tag{4.10}$$

Then, assuming that $n_t \geqslant n_v$,

$$\left(1 - \frac{1}{\sqrt{\log n_t}}\right) \mathbb{E}\left[\ell\big(t_{u_*}, \widehat{f}^{\text{ag}}_{\mathcal{T}}\big)\right] \leqslant \left(1 + \frac{1}{\sqrt{\log n_t}}\right) \mathbb{E}\left[\inf_{1\leqslant k\leqslant d} \ell\big(t_{u_*}, t_{\hat{u}_k(D_{n_t})}\big)\right] \\ + 72 \frac{c \log^{\frac{3}{2}} n_t}{\eta n_v} \left[c + \frac{3c}{n_t} + \frac{4\mathbb{E}[|Y|]}{n_t}\right]. \tag{4.11}$$

Proposition 4.3.5 is proved in appendix 4.7.2. The specific setting of Proposition 4.3.5 allows to state a fairly explicit oracle inequality for Agghoo under few conditions. Assuming as before that n_t and n_v are both of order n, the remainder term in equation (4.3.5) is of order $\frac{\log^{\frac{3}{2}} n}{n}$. This is negligible compared to the minimax rates achievable under regularity assumptions (eg. Hölder or Besov balls),

which are of order $n^{-\alpha}$ with $\alpha \in (0; \frac{2}{3}]$. Hence, Proposition 4.3.5 shows that Agghoo adapts to the unknown level of regularity, achieving the correct convergence rate.

Three assumptions are made to obtain this oracle inequality. The distributional assumption on the residuals (equation (4.9)) is identical to the one made in Theorem 4.3.2, and has already been discussed. The moment condition $\mathbb{E}[|Y|] < +\infty$, without which inequality (4.11) becomes vacuous, is self-explanatory.

Finally, hypothesis (4.10) requires that the intervals of the partition contain at least $cst \times \log^2 n_t$ points, on average. This is a mild requirement, since partitions finer than this cannot be expected to perform well anyway due to high variance of the empirical average.

Though hypothesis (4.10) involves the unknown distribution P, Bernstein's inequality shows that if $\min_{1 \leqslant j \leqslant d} P_{n_t}(I_j) \geqslant \frac{C \log^2 n_t}{\eta^2 n_v}$, where $C > 1536$ and P_{n_t} denotes the empirical measure on a sample of size n_t, then equation (4.10) holds with high probability. Thus, provided a lower bound on η is known, it is possible to guarante empirically that (4.10) holds, with a high degree of confidence.

4.3.1 Effect of V

The upper bound given by Theorem 4.3.2 only depends on $\mathcal{T}$ through n_v and n_t. The purpose of this section is to show that for a given value of n_v, increasing $V = |\mathcal{T}|$ cannot increase the risk. This is proved in the case of monte carlo subset generation defined below.

Definition 4.3.6 *For $\tau \in \left[\frac{1}{n}; 1\right]$ and $V \in \mathbb{N}^*$, let $\mathcal{T}^{mc}_{\tau,V}$ be generated independently of the data D_n by drawing V elements independently and uniformly in the set*

$$\{T \subset [|1; n|] : |T| = \lfloor \tau n \rfloor\}.$$

For fixed τ, the excess risk of Agghoo is a non-increasing function of V.

Proposition 4.3.7 *Let $U \leqslant V$ be two non-zero integers. Let $\tau \in \left[\frac{1}{n}; 1\right]$. Then:*

$$\mathbb{E}\left[\ell(s, \widehat{f}^{\text{ag}}_{\mathcal{T}^{mc}_{\tau,V}})\right] \leqslant \mathbb{E}\left[\ell(s, \widehat{f}^{\text{ag}}_{\mathcal{T}^{mc}_{\tau,U}})\right].$$

Proof Let $(T_i)_{i=1,\dots,V} = \mathcal{T}^{mc}_{\tau,U}$. Let $\mathcal{I} = \{I \subset [|1;V|] : |I| = U\}$. Then

$$\begin{aligned}
\widehat{f}^{\mathrm{ag}}_{\mathcal{T}^{mc}_{\tau,U}} &= \sum_{i=1}^{V} \frac{1}{V} \widehat{f}^{\mathrm{ho}}_{T_i} \\
&= \sum_{i=1}^{V} \frac{\binom{V-1}{U-1}}{U\binom{V}{U}} \widehat{f}^{\mathrm{ho}}_{T_i} \\
&= \frac{1}{U} \sum_{i=1}^{V} \frac{\sum_{I\in\mathcal{I}} \mathbb{I}_{i\in I}}{|\mathcal{I}|} \widehat{f}^{\mathrm{ho}}_{T_i} \\
&= \frac{1}{|\mathcal{I}|} \sum_{I\in\mathcal{I}} \frac{1}{U} \sum_{i\in I} \widehat{f}^{\mathrm{ho}}_{T_i}.
\end{aligned}$$

It follows by convexity of $f \mapsto \ell(s,f)$ that

$$\mathbb{E}\left[\ell(s, \widehat{f}^{\mathrm{ag}}_{\mathcal{T}^{mc}_{\tau,U}})\right] \leqslant \frac{1}{|\mathcal{I}|} \sum_{I\in\mathcal{I}} \mathbb{E}\left[\ell(s, \frac{1}{U}\sum_{i\in I} \widehat{f}^{\mathrm{ho}}_{T_i})\right].$$

For any $I \in \mathcal{I}$, $(T_i)_{i\in I} \sim \mathcal{T}^{mc}_{\tau,U}$ and is independent of D_n, therefore $\frac{1}{U}\sum_{i\in I} \widehat{f}^{\mathrm{ho}}_{T_i} \sim \widehat{f}^{\mathrm{ag}}_{\mathcal{T}^{mc}_{\tau,U}}$. This yields the result. ■

It can be seen from the proof that the proposition also holds for Agcv. Therefore, increasing V can only improve the performance of these methods. On the other hand, no such theoretical result is known for CV, even though increasing the number of CV splits (for given τ) almost always improves performance in practice.

4.4 Simulation study

This section focuses on hyperparameter selection for the Lasso with Huber loss, either using a fixed grid or using the reparametrization from Definition 4.2.3. The methods considered for this task are Aggregated hold-out given by Definition 4.2.5, Aggregated cross-validation given by Definition 4.2.6 and standard cross-validation. In all cases, the subsamples are generated independently from the data and uniformly among subsets of a given size τn, as in Definition 4.3.6. Thus, all three methods share the same two hyperparameters: τ, the fraction of data used for training the Lasso, and V, the number of subsets used by the method.

For the huberized Lasso with a fixed grid, the hqreg_raw function from the R package hqreg is used with a fixed grid designed to emulate the default choice: a geometrically decreasing sequence of length 100, with maximum value λ_{max} and

minimum value $\lambda_{min} = 0.05\lambda_{max}$. The fixed value of λ_{max} is obtained by averaging the (data-dependent) default value chosen by hqreg_raw over 10 independent datasets. To compute the reparametrization given by Definition 4.2.3, I implemented the LARS-based algorithm described by Rosset and Zhu [91], which allows to compute the whole regularization path.

I.i.d training samples of size $n = 100$ are generated according to a distribution (X, Y), where $X \in \mathbb{R}^{1000}$ and $Y = w_*^T X + \varepsilon$, with ε independent from X. To illustrate the robustness of the estimators, Cauchy noise is used: $\varepsilon \sim \text{Cauchy}(0, \sigma)$. The performance of Agghoo and cross-validation may depend on the presence of correlations between the covariates X and the sparsity of the ground truth w_*. To investigate these effects, three parametric families of distribution are considered for X, in sections 4.4.1, 4.4.2 and 4.4.3.

The risk of each method is evaluated on an independent training set of size 500, and results are averaged over 1000 repetitions of the simulation. More precisely, 1000 training sets D_j of size $n = 100$ are generated, along with 1000 test sets $(X'_{i,j}, Y'_{i,j})_{1 \leqslant i \leqslant 500}$, each of size 500. For each simulation j and any learning rule $\mathcal{A}_{\tau,V}$ among the six obtained by combining Agghoo, monte carlo CV and AGCV with either a fixed grid or the zero-norm parametrization, the average excess risk

$$\hat{R}_j(\mathcal{A}, \tau, V) = \frac{1}{500} \sum_{i=1}^{500} \left[\phi_c \left(Y'_{i,j} - \mathcal{A}_{\tau,V}(D_j)(X'_{i,j}) \right) - \phi_c \left(Y'_{i,j} - s(X'_{i,j}) \right) \right]$$

is computed on the test set for all values of $V \in \{1, 2, 5, 10\}$ and $\tau \in \left\{ \frac{i}{10} : 1 \leqslant i \leqslant 9 \right\}$.

4.4.1 Experimental setup 1

X is generated using the formula $X_i = \frac{1}{\|u\|_2} \sum_{j=1}^{d} u_{i-j} Z_j$, where Z_j are independent standard Gaussian random variables, $u_i = \mathbb{I}_{|i| \leqslant cor} e^{-\frac{2.33^2 i^2}{2cor^2}}$ and $cor \in \mathbb{N}$ is a parameter regulating the strength of the correlations. The regression coefficient has a support of size $r = 3 * k$ drawn at random from $[|1; 1000|]$, and is defined by $w_{*,j} = u_{*,g(j)}$, where g is a uniform random permutation, $u_{*,j} = b$ if $1 \leqslant j \leqslant k$ and $u_{*,j} = \frac{b}{4}$ if $2k + 1 \leqslant j \leqslant 3k$, with b calibrated so that $\|X w_*\|_{L^2} = 1$. The noise parameter is $\sigma = 0.08$, while the huber loss parameter c is set to 2 – a sub-optimal choice in this setting, but convenient for computing the huberized Lasso regularization path.

Choice of τ parameter For all methods, in most cases the optimal value of τ is 0.8 or 0.9, similarly to what was observed in the rkhs case, where $\tau = 0.8$ was

recommended. Table 1 displays the quantity

$$\hat{G}(\mathcal{A},\tau,V)=\frac{\text{Mean}\left[(\hat{R}_j(\mathcal{A},\tau,V)-\hat{R}_j(\mathcal{A},\tau_*,V))_{1\leqslant j\leqslant 1000}\right]}{\text{Sd}\left[(\hat{R}_j(\mathcal{A},\tau,V)-\hat{R}_j(\mathcal{A},\tau_*,V))_{1\leqslant j\leqslant 1000}\right]},$$

where Sd denotes the (empirical) standard deviation and τ_* the optimal choice of τ, $\tau_*=\text{argmin}_{\tau\in\{0.1,\dots,0.9\}}\,\text{Mean}\left[(\hat{R}_j(\mathcal{A},\tau,V))_{1\leqslant j\leqslant 1000}\right]$. Thus, values of $\hat{G}(\mathcal{A},\tau,V)$ bigger than a few units suggest that τ is suboptimal to a statistically significant degree. When $\tau_*=0.9$, $\hat{G}(\mathcal{A},0.8,V)$ is displayed in black on table 1. When $\tau_*=0.8$, $\hat{G}(\mathcal{A},0.9,V)$ is displayed in blue on table 1. Exceptions where $\tau_*\notin\{0.8,0.9\}$ are highlighted in red, with the value $\min\left(\hat{G}(\mathcal{A},0.8,V),\hat{G}(\mathcal{A},0.9,V)\right)$.

Most of the exceptions $\tau_*\notin\{0.8,0.9\}$ occur on the column $r=150$, $cor=1$, while most of the others are of low statistical significance, with values less than 1.1 on the fourth column ($r=60$ and $cor=1$). Thus, table 1 confirms the claim that $\tau_*\in\{0.8,0.9\}$ for all methods, in most cases. For grid agghoo, 0−norm agghoo, grid agcv and $V\geqslant 5$, $\tau_*\in\{0.8,0.9\}$ for all simulations. Comparing now $\tau=0.8$ and $\tau=0.9$, grid agghoo and 0−norm agghoo with $V\geqslant 5$ show a clear pattern: $\tau=0.9$ is better or as good as $\tau=0.8$ in all cases except $r=150, cor=1$ where $\tau=0.8$ is significantly better. For other methods, results are not so clear and the difference in risk between the two values of τ is often insignificant.

Choice of V For all methods considered, performance is expected to improve when V is increased, but by how much? If the performance increase is too slight, it may not be worth the additional computational cost. In figure 1, the mean excess risk for the optimal value of τ is displayed as a function of V, with error bars corresponding to one standard deviation. The scale used for the vertical axis in each graph is the average excess risk of the oracle with respect to the fixed grid over the λ parameter. Quantifying performance as a percentage of the oracle risk, when $cor=15$, Agghoo improves by roughly 20% from $V=1$ to $V=2$, by roughly 10% from $V=2$ to $V=5$ and by a few percent more from $V=5$ to $V=10$. CV with the standard grid behaves similarly in these two simulations, while CV with the zero-norm parametrization shows much less improvement when V is increased. Thus, taking $V\geqslant 5$ is advantageous, but there are clearly diminishing returns to choosing V much larger than this. For CV with the zero-norm parametrization, $V=2$ seems sufficient in these simulations .

Comparison between methods From figure 1, it appears that grid agcv is a very poor choice, being worse than both grid agghoo and grid cv for all values of

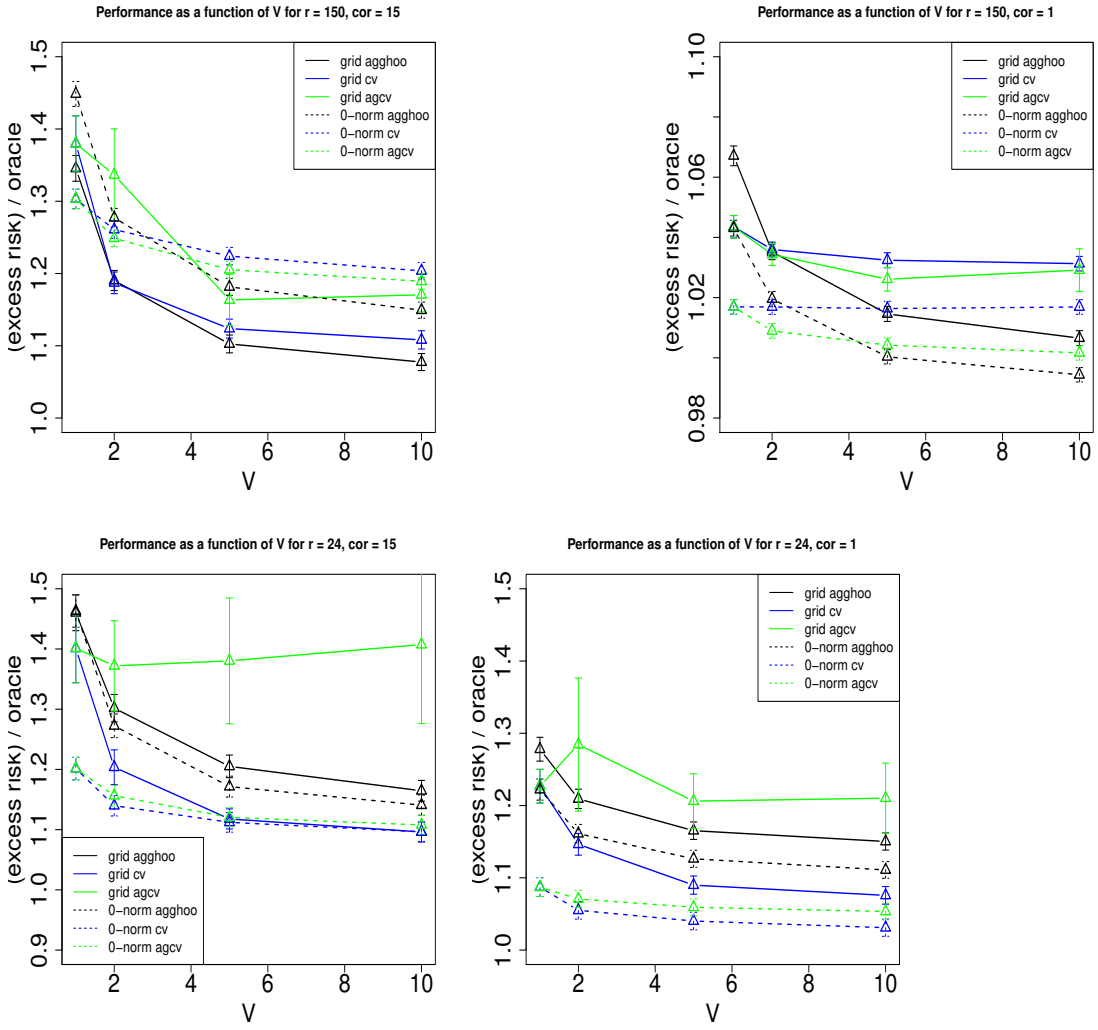

Figure 4.1: Performance relative to the oracle, as a function of V

V when $r = 150, cor = 15$, and being the worst of all the methods for $V \geqslant 2$ when $r = 24$, as well as highly unstable, as the size of the error bars clearly shows.

Interestingly, 0−norm agcv behaves much better, being the second best method when $cor = 1$, and very close to the best when $r = 24$ and $cor = 15$.

Generally speaking, of the two types of parametrization of the Lasso, the zero-norm parametrization appears to perform better than the standard grid when correlations are small ($cor = 1$), while the performance is significantly worse when $r = 150$ and $cor = 15$.

Comparing now Agghoo and CV, Agghoo appears to be better than CV when $V \geqslant 2$ in situations where r is larger ($r = 150$). This seems to hold for both the standard parametrization (grid agghoo) and the zero-norm one (0−norm agghoo). The relation is reversed for small r, with CV performing better than Agghoo for all values of V when $r = 24$.

Further studies The previous simulations suggest that Agghoo performs better than CV in the case of high intrinsic dimension. However, the effect of correlations is unclear. Experimental setup 1 mixes different types of correlations: correlations between predictive variables, correlations between predictive and non-predictive variables, and correlations among non-predictive variables. It is possible that one type of correlation favours Agghoo while another favours CV.

To gain a more accurate idea of when Agghoo is advantageous over CV, two more settings are studied, considering separately correlations among predictive variables, and between predictive and non-predictive variables. Since previous simulations showed that $\tau = 0.8, 0.9$ and $V = 10$ were the optimal parameters, only those parameters will be considered in the following.

Since the choice of lasso parametrization did not seem to affect the relative performance of Agghoo and CV, we only consider the standard parametrization, as it is more popular and also easier to use in our simulations. Agcv is not considered either, since it was discovered to be unreliable in previous simulations.

4.4.2 Experimental setup 2: correlations between predictive and noise variables

Let r be the number of predictive variables and let each predictive covariate have s "noise" covariates which are correlated with it at level $\rho = 0.8$. Assume that $rs \leqslant d$, where d is the total number of variables. Let $(Z_i^0)_{1 \leqslant i \leqslant r}$, $(Z_{i,j})_{1 \leqslant i \leqslant r, 1 \leqslant j \leqslant s}$ and $(W_k)_{1 \leqslant k \leqslant d-rs}$ be independent standard gaussian variables. For any $j \in [|0 : r-1|]$ and any $i \in [|1; s|]$, let $Z_{jr+i} = \sqrt{0.8} Y_j^0 + \sqrt{0.2} Z_{i,j}$ and for $rs < i \leqslant d$, let $X_i = W_{i-rs}$. For the regression coefficient, choose $w_* = \frac{3*u}{\|Xu\|_{L_2}}$,

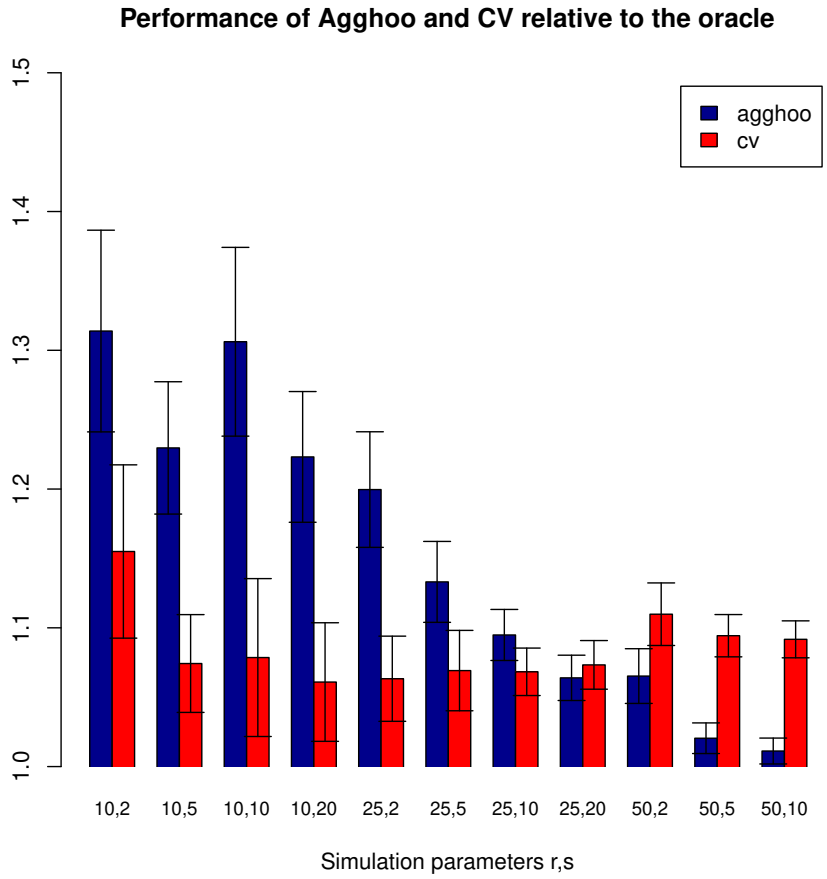

Figure 4.2: Relative risk in experimental setup 2 (section 4.4.2)

where $u = (\mathbb{I}_{r|(j-1)}\mathbb{I}_{j\leqslant rs})_{1\leqslant j\leqslant d}$. Let then Y be distributed conditionnally on X as Cauchy$(\langle w_*, X\rangle, 0.3)$. The loss function used here is ϕ_c with $c = 2$.

Results Figure 4.2 shows a bar plot of the average excess risk of CV and Agghoo as a fraction of the average risk of the oracle. 90 % error bars were estimated using asymptotic theory. Parameters used for Agghoo and CV were $\tau = 0.9$ and $V = 10$ ($\tau = 0.8$ yields similar result).

Overall, Agghoo's risk relative to the oracle significantly decreases as the zero-norm of w_* increases from $r = 10$ to $r = 50$, as was observed in section 4.4.1 . For $r = 25$ and $r = 50$ separately, the risk relative to the oracle significantly decreases as s increases from 2 to 10. For $r = 10$, this trend is unclear due to the random errors.

In contrast, CV's performance relative to the oracle shows no clear trend either

as a function of r or as as function of s, and could be constant when taking error bars into accounts.

As a result of these trends, Agghoo performs significantly worse than CV for $r = 10$ and significantly better when $r = 50$, especially when $s \geqslant 5$. When $r = 25$, CV performs significantly better than Agghoo for $s = 2$ and $s = 5$ and they perform similarly when $s = 10$ and $s = 20$.

4.4.3 Experimental setup 3: correlations between predictive variables

We consider now predictive covariates which are correlated between them, and independent from the unpredictive covariates. As above, let r denote the number of predictive variables and $\rho > 0$ be the level of correlations. Let Z_0, $(Z_i)_{1\leqslant i\leqslant r}$ and $(W_i)_{1\leqslant i\leqslant d-r}$ be standard Gaussian random variables. The random variable X is then defined by $X_i = \sqrt{\rho}Z_0 + \sqrt{1-\rho}Z_i$ for $1 \leqslant i \leqslant r$ and $X_i = W_{i-r}$ for $r+1 \leqslant i \leqslant d$. As in section 4.4.2, the regression coefficient w_* is a constant vector of the form $\frac{3*u}{\|Xu\|_{L^2}}$, where this time $u = (\mathbb{I}_{1\leqslant i\leqslant r})_{1\leqslant i\leqslant d}$.

Y is distributed conditionnally on X as Cauchy$(\langle X, w_*\rangle, 0.3)$ and the loss function used is the Huber loss ϕ_2.

Results Figure 4.3 shows a barplot generated in the same way as in section 4.4.2. Parameters used for Agghoo and CV were $V = 10$ and $\tau = 0.8$, which is optimal in this case for both Agghoo and CV.

As in previous simulations, Agghoo's performance relative to the oracle improves significantly when the intrinsic dimension r grows from 25 to 200, for a given value of ρ. The decrease in relative risk is faster for small values of ρ. As a result, Agghoo performs best, relative to the oracle, when $\rho = 0.2$ for $r = 200$, whereas best performance seems to occur at $\rho = 0.5$ for smaller values of r, up to random errors.

For cross-validation, the relative risk seems more or less unaffected by the dimension r, but shows an increasing trend as a function of ρ for all values of r.

As a result, Agghoo performs better than CV for $r = 200$ and for $r = 100$ and $\rho = 0.2, 0.5$. For $r = 200$ and $\rho = 0.2$, Agghoo even performs significantly better than the oracle! This is possible, since the Agghoo regression coefficient $\hat{\theta}^{ag}_{\mathcal{T}}$ does not itself belong to the Lasso regularization path.

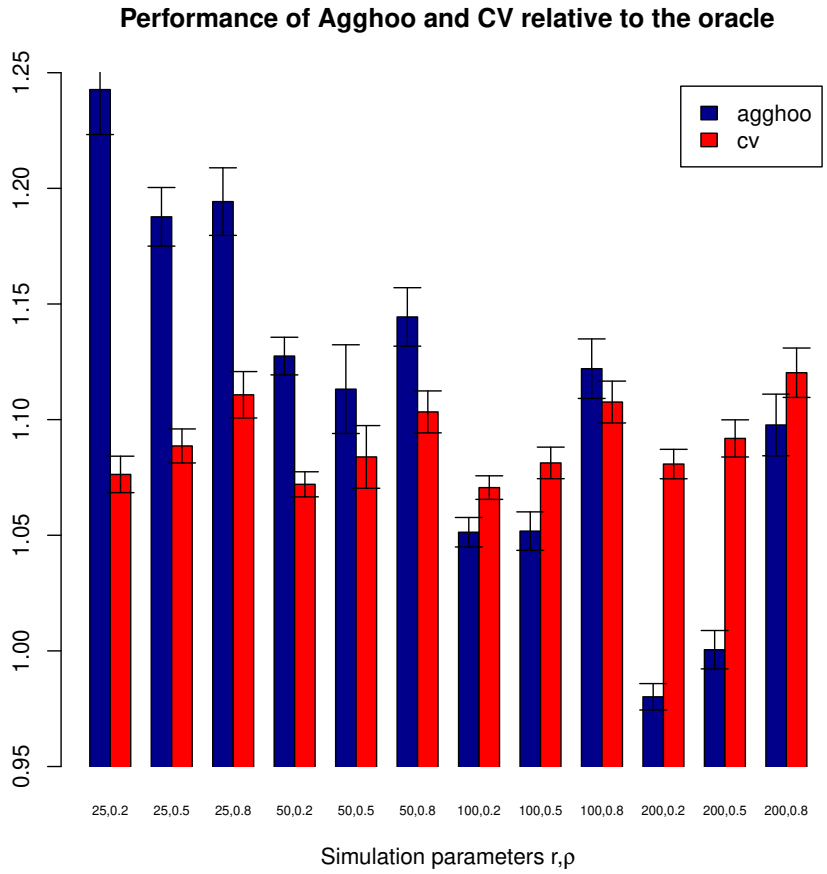

Figure 4.3: Relative risk in experimental setup 3 (section 4.4.3)

			r = 150		r = 60		r = 24	
	method	V	15	1	15	1	15	1
1	grid agghoo	1	2.2	2.7	3.0	2.7	0.5	5.6
2	grid agghoo	2	2.5	2.1	3.1	1.4	1.0	7.9
3	grid agghoo	5	2.5	6.8	3.5	0.6	0.6	11.9
4	grid agghoo	10	0.7	7.2	3.7	1.1	4.5	16.7
5	grid cv	1	1.0	3.9	1.6	0.1	1.2	1.5
6	grid cv	2	0.8	5.0	2.6	0.5	1.4	1.1
7	grid cv	5	1.4	2.8	1.5	0.8	0.5	3.7
8	grid cv	10	2.0	2.6	2.9	1.1	1.6	5.9
9	grid agcv	1	1.0	3.9	1.6	0.1	1.2	1.5
10	grid agcv	2	0.3	2.0	1.4	1.9	0.3	0.8
11	grid agcv	5	0.3	2.2	0.5	0.7	0.5	1.1
12	grid agcv	10	0.5	0.4	0.0	0.3	0.8	1.0
13	0−norm agghoo	1	1.3	4.1	2.0	0.3	0.5	5.6
14	0−norm agghoo	2	3.0	1.4	3.2	1.3	1.9	9.2
15	0−norm agghoo	5	4.0	6.7	5.1	3.3	4.0	13.7
16	0−norm agghoo	10	4.6	7.3	7.0	3.7	5.2	18.5
17	0−norm cv	1	4.3	9.4	4.3	1.1	2.0	3.9
18	0−norm cv	2	1.9	7.2	1.8	4.4	4.8	2.7
19	0−norm cv	5	2.7	5.3	2.4	3.3	1.5	0.7
20	0−norm cv	10	6.1	4.6	5.4	3.5	0.6	0.1
21	0−norm agcv	1	4.3	9.4	4.3	1.1	2.0	3.9
22	0−norm agcv	2	1.9	5.8	2.4	4.5	5.9	3.5
23	0−norm agcv	5	2.1	1.9	1.0	4.0	5.7	3.7
24	0−norm agcv	10	4.5	1.0	3.3	3.6	7.3	3.9

Table 4.1: $\hat{G}(\mathcal{A}, \tau, V)$ for sub-optimal $\tau \in \{0.8, 0.9\}$ and various distributions. Colours show optimal τ_*: blue for $\tau_* = 0.8$, black for 0.9, red when $\tau_* \notin \{0.8, 0.9\}$.

4.5 Conclusion

Aggregated hold-out (Agghoo) satisfies an oracle inequality (Theorem 4.3.2) in sparse linear regression with the huber loss. This oracle inequality is asymptotically optimal in the non-parametric case where the intrinsic dimension tends to $+\infty$ with the sample size n, provided that an $L^\infty(X) - L^2(X)$ norm inequality holds on the set of sparse linear predictors, where X is the random vector of covariates. When X is a vector of independent Bernoulli variables, this condition amounts to restricting the zero-norm of the coefficients to be less than a constant times $\frac{n}{\log n}$. Theorem 4.3.2 also applies to adaptive piecewise constant regression, yielding an oracle inequality in that setting (Proposition 4.3.5).

When Monte-Carlo subsampling is used (Definition 4.3.6), Agghoo has two parameters, τ and V. Theoretically, it is shown that Agghoo's performance always improves when V grows for a fixed τ. Simulations show a large improvement from $V = 1$ to $V = 5$ in some cases, but diminishing returns for $V > 5$. With respect to τ, simulations show that $\tau = 0.8$ or $\tau = 0.9$ is optimal or near optimal in most cases. In particular, a default choice of $V = 10$, $\tau = 0.8$ seems reasonable.

Compared to cross-validation with the same number of splits V, simulations show that Agghoo performs better when the intrinsic dimension r is large enough ($r = 150$ in section 4.4.1, $r = 50$ in section 4.4.2 and $r = 100$ in 4.4.3) for $n = 100$ observations and $d = 1000$ covariates. Correlations between predictive and non-predictive covariates, which increase the number of covariates correlated with the response Y, clearly favour Agghoo relative to CV and the oracle, whereas the effect of correlations between predictive covariates is ambiguous.

4.6 Proof of Theorem 4.3.2

The idea is to apply Theorem 3.7.3 of Chapter 3 using suitable functions $(\hat{w}_{i,j})_{(i,j)\in\{1;2\}^2}$. Fix a dataset D_{n_t}, $K \in \{1, \dots, n_t\}$ and for any $k \in [|1; K|]^2$, let $\hat{t}_k = \mathcal{A}_k(D_{n_t}) : x \to \hat{q}_k(D_{n_t}) + \langle \hat{\theta}_k(D_{n_t}), x\rangle$. More precisely, to apply Theorem 3.7.3 of Chapter 3, one must show inequalities of the form $H(w_1, w_2, (\hat{t}_k)_{1\leqslant k\leqslant K})$: for all $r \geqslant 2$,

$$\mathbb{E}\Big(\big|\phi_c(\hat{t}_k(X) - Y) - \phi_c(\hat{t}_l(X) - Y) - c_l^k\big|^r\Big) \leqslant k!\Big[w_1\big(\sqrt{\ell(s, \hat{t}_k)}\big) + w_1\big(\sqrt{\ell(s, \hat{t}_l)}\big)\Big]^2 \times \Big[w_2\big(\sqrt{\ell(s, \hat{t}_k)}\big) + w_2\big(\sqrt{\ell(s, \hat{t}_l)}\big)\Big]^{r-2}, \tag{4.12}$$

where w_1, w_2 are non-decreasing functions. Since ϕ_c is Lipschitz, it is enough to control $\big\|\hat{t}_k - \hat{t}_l\big\|_{L^\infty(X)}$ and $\big\|\hat{t}_k - \hat{t}_l\big\|_{L^2(X)}$ by functions of $\ell(s, \hat{t}_k)$ and $\ell(s, \hat{t}_l)$.

4.6.1 Controlling the supremum norm $\left\|\hat{t}_k - \hat{t}_l\right\|_{L^\infty(X)}$

First, let us bound the supremum norm by the L^2 norm.

Claim 4.6.0.1 *For any $k \in \{1, \dots, K\}$, recall that $\hat{t}_k = \mathcal{A}_k(D_{n_t})$. Then:*

$$\forall (k,l) \in \{1,\dots,K\}^2, \left\|\hat{t}_k - \hat{t}_l\right\|_{L^\infty(X)} \leqslant \sqrt{2}\kappa(K) \left\|\hat{t}_k - \hat{t}_l\right\|_{L^2(X)} \text{ a.s. .}$$

Proof Let X be independent from D_n and observe that for any k,

$$\hat{t}_k(X) = \hat{b}_k + \hat{\theta}_k^T(X - EX),$$

where $\hat{b}_k = \hat{q}_k + \hat{\theta}_k^T EX$ (using the notations of hypothesis 4.2.1). Hence,

$$\left\|\hat{t}_k(X) - \hat{t}_l(X)\right\|_{L^\infty} \leqslant |\hat{b}_k - \hat{b}_l| + \left\|(\hat{\theta}_k - \hat{\theta}_l)^T(X - EX)\right\|_{L^\infty}.$$

By hypothesis 4.2.1, $\left\|\hat{\theta}_k\right\|_0 = k$. Thus, if $K \geqslant \max(k,l)$, $\left\|\hat{\theta}_k - \hat{\theta}_l\right\|_0 \leqslant k + l \leqslant 2K$. The definition of κ (equation (4.5)) implies that

$$\begin{aligned}
\left\|\hat{t}_k(X) - \hat{t}_l(X)\right\|_{L^\infty} &\leqslant |\hat{b}_k - \hat{b}_l| + \kappa(K) \left\|(\hat{\theta}_k - \hat{\theta}_l)^T(X - EX)\right\|_{L^2} \\
&\leqslant \kappa(K) \left[|\hat{b}_k - \hat{b}_l| + \left\|(\hat{\theta}_k - \hat{\theta}_l)^T(X - EX)\right\|_{L^2}\right] (\kappa(K) \geqslant 1 \text{ by definition}) \\
&\leqslant \sqrt{2}\kappa(K)\sqrt{|\hat{b}_k - \hat{b}_l|^2 + \left\|(\hat{\theta}_k - \hat{\theta}_l)^T(X - EX)\right\|_{L^2}^2} \\
&= \sqrt{2}\kappa(K) \left\|\hat{t}_k(X) - \hat{t}_l(X)\right\|_{L^2}.
\end{aligned}$$

■

A uniform bound on the supremum norm is also required.

Definition 4.6.1 *Let*

$$\hat{\beta} = \max_{(k,l) \in [|1;n_t|]^2} \left\|\hat{t}_k - \hat{t}_l\right\|_{L^\infty(X)}$$

$\mathbb{E}[\hat{\beta}]$ can be bounded as follows.

Claim 4.6.1.1 *The $\hat{\beta}$ of definition 4.6.1 is such that*

$$\mathbb{E}[\hat{\beta}] \leqslant 8LRn^\alpha$$

Proof Let $(k,l) \in [|1;n_t|]^2$. Defining $\widetilde{X}_i = X_i - \frac{1}{n}\sum_{i=1}^n X_i$ and changing variables in hypothesis 4.2.1 from (q,θ) to $\left(b = q+ < \theta, \frac{1}{n}\sum_{i=1}^n X_i >, \theta\right)$, we can rewrite $\hat{t}_k$ as

$$\hat{t}_k(x) = \hat{b}_k(D_n) + \hat{\theta}_k(D_n)^T \left(x - \frac{1}{n}\sum_{i=1}^n X_i \right)$$

where

$$\hat{b}_k(D_n) \in \underset{b \in \hat{Q}'\left(D_n, \hat{\theta}_k(D_n)\right)}{\operatorname{argmin}} |b|$$

$$\hat{Q}'(D_n,\theta) = \underset{b\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n \phi_c\left(Y_i - b - \theta^T \widetilde{X}_i\right).$$

Therefore, differentiating with respect to b,

$$\frac{1}{n}\sum_{i=1}^n \phi_c'\big(Y_i - \hat{b}_k - \hat{\theta}_k^T \widetilde{X}_i\big) = 0.$$

Assume by contradiction that

$$\exists b > 0, \forall i \in [|1;n_t|], \hat{b}_k + b + \hat{\theta}_k^T \widetilde{X}_i \leqslant \hat{b}_l + \hat{\theta}_l^T \widetilde{X}_i. \tag{4.13}$$

Let b be such that (4.13) holds. Then by monotony of ϕ_c', for all ε in $[0;\frac{b}{2}]$,

$$\begin{aligned}
0 &= \frac{1}{n}\sum_{i=1}^n \phi_c'\big(Y_i - \hat{b}_k - \hat{\theta}_k^T \widetilde{X}_i\big) \\
&\geqslant \frac{1}{n}\sum_{i=1}^n \phi_c'\big(Y_i - \hat{b}_k - \varepsilon - \hat{\theta}_k^T \widetilde{X}_i\big) \\
&\geqslant \frac{1}{n}\sum_{i=1}^n \phi_c'\big(Y_i - \hat{b}_k - \frac{b}{2} - \hat{\theta}_k^T \widetilde{X}_i\big) \\
&\geqslant \frac{1}{n}\sum_{i=1}^n \phi_c'\big(Y_i - \hat{b}_l + \frac{b}{2} - \hat{\theta}_l^T \widetilde{X}_i\big) \\
&\geqslant \frac{1}{n}\sum_{i=1}^n \phi_c'\big(Y_i - \hat{b}_l + \varepsilon - \hat{\theta}_l^T \widetilde{X}_i\big) \\
&\geqslant \frac{1}{n}\sum_{i=1}^n \phi_c'\big(Y_i - \hat{b}_l - \hat{\theta}_l^T \widetilde{X}_i\big) \\
&= 0.
\end{aligned}$$

It follows that

$$\forall \varepsilon \in [0; \frac{b}{2}], \frac{1}{n}\sum_{i=1}^{n} \phi_c'\big(Y_i - \hat{b}_k - \varepsilon - \hat{\theta}_k^T \widetilde{X}_i\big) = \frac{1}{n}\sum_{i=1}^{n} \phi_c'\big(Y_i - \hat{b}_l + \varepsilon - \hat{\theta}_l^T \widetilde{X}_i\big) = 0. \quad (4.14)$$

By integration, this implies that for all $\varepsilon \in [0; \frac{b}{2}]$,

$$(\hat{b}_k + \varepsilon) \in \hat{Q}'\left(D_n, \hat{\theta}_k(D_n)\right), \quad (4.15)$$

$$(\hat{b}_l - \varepsilon) \in \hat{Q}'\left(D_n, \hat{\theta}_l(D_n)\right). \quad (4.16)$$

If $\hat{b}_l > 0$, then for small enough ε, (4.16) contradicts the minimality of $|\hat{b}_l|$. On the other hand, if $\hat{b}_l \leqslant 0$, then averaging (4.13) over $i \in \{1, \dots, n\}$ yields

$$\hat{b}_k + b \leqslant \hat{b}_l \leqslant 0.$$

Then for $\varepsilon \in [0; \frac{b}{2}]$, (4.15) contradicts the minimality of $|\hat{b}_k|$. Thus, (4.13) leads to a contradiction. Let i be such that $\hat{b}_k + \hat{\theta}_k^T \widetilde{X}_i \geqslant \hat{b}_l + \hat{\theta}_l^T \widetilde{X}_i$. Then

$$\hat{b}_l - \hat{b}_k \leqslant \big(\hat{\theta}_k - \hat{\theta}_l\big)^T \widetilde{X}_i \leqslant \max_{i=1,\dots,n_t} \big|\big(\hat{\theta}_k - \hat{\theta}_l\big)^T \widetilde{X}_i\big|.$$

Exchanging k and l yields

$$|\hat{b}_l - \hat{b}_k| \leqslant \max_{1 \leqslant i \leqslant n_t} \big|\big(\hat{\theta}_k - \hat{\theta}_l\big)^T \widetilde{X}_i\big|.$$

Therefore, for any k, l,

$$\begin{aligned}
\big\|\hat{t}_k - \hat{t}_l\big\|_{L^\infty}(X) &\leqslant |\hat{b}_l - \hat{b}_k| + \sup_{x \in \mathrm{supp}(X)} \big|(\hat{\theta}_k - \hat{\theta}_l)^T (x - \frac{1}{n}\sum_{i=1}^{n} X_i)\big| \\
&\leqslant 2 \sup_{x \in \mathrm{supp}(X)} \big|(\hat{\theta}_k - \hat{\theta}_l)^T (x - \frac{1}{n}\sum_{i=1}^{n} X_i)\big| \\
&\leqslant 2\big\|\hat{\theta}_k - \hat{\theta}_l\big\|_1 \sup_{(x,y) \in \mathrm{supp}(X)^2} \|x - y\|_\infty \\
&\leqslant 4\big\|\hat{\theta}_k - \hat{\theta}_l\big\|_1 \sup_{x \in \mathrm{supp}(X)} \|x - EX\|_\infty \\
&\leqslant 8 \sup_{1 \leqslant k \leqslant n_t} \left\|\hat{\theta}_k\right\| \sup_{x \in \mathrm{supp}(X)} \|x - EX\|_\infty .
\end{aligned}$$

Thus, by definition 4.6.1, $\hat{\beta} \leqslant 8 \sup_{1 \leqslant k \leqslant n_t} \left\|\hat{\theta}_k\right\| \sup_{x \in \mathrm{supp}(\bar{X})} \|x\|_\infty$.

Hence, by hypothesis 4.2.1,

$$\mathbb{E}[\hat{\beta}] \leqslant 8LRn^\alpha.$$

■

4.6.2 Proving hypotheses $H\left(\hat{w}_{i,1}, \hat{w}_{i,2}, (\hat{t}_k)_{1\leqslant k\leqslant K}\right)$

The following lemma will be useful.

Lemma 4.6.2 *Let r, s, x be positive real numbers. Let*

$$I_{r,s}(x) = \{v \in \mathbb{R}_+ : v \leqslant (r \vee s\sqrt{v})x\}$$

and $h_{r,s}(x) = (\sqrt{r}x) \vee sx^2$. Then for all $x, y \geqslant 0$,

$$\sup I_{r,s}(x+y) \leqslant (h_{r,s}(\sqrt{x}) + h_{r,s}(\sqrt{y}))^2.$$

Proof Let $v \in I_{r,s}(x)$. Then if $s\sqrt{v} \leqslant r$, by definition of $I_{r,s}(x)$, $v \leqslant rx$. Otherwise $s\sqrt{v} > r$. Therefore by definition of $I_{r,s}(x)$,

$$\begin{aligned} v &\leqslant s\sqrt{v}x \\ v &\leqslant s^2x^2. \end{aligned}$$

In all cases, $v \leqslant (rx) \vee s^2x^2$. Therefore,

$$\sqrt{\sup I_{r,s}(x+y)} \leqslant \left(\sqrt{r(x+y)}\right) \vee s(x+y) \leqslant \left(\sqrt{rx}\right) \vee sx + (\sqrt{ry}) \vee sy,$$

using the elementary inequalities:

$$\begin{aligned} \forall (x, y, a, b) \in \mathbb{R}_+^3, & \\ \sqrt{x+y} &\leqslant \sqrt{x} + \sqrt{y} \\ (x+y) \vee (a+b) &\leqslant x \vee a + y \vee b \end{aligned}$$

■

We now relate the L^2 norm to the excess risk in the following Proposition.

Proposition 4.6.3 *Let $(X, Y) \in \mathcal{X} \times \mathbb{R}$ be random variables. Let ϕ_c be the Huber loss with parameter $c > 0$. Assume that there exists $\eta > 0$ such that almost everywhere,*

$$\mathbb{P}\left(|Y - s(X)| \leqslant \frac{c}{2}|X\right) \geqslant \eta.$$

Then for any measurable functions $(f_1, f_2) : \mathcal{X} \to \mathbb{R}^2$,

$$c^2 \|f_1 - f_2\|_{L^2(X)}^2 \leqslant \frac{4c}{\eta}\left[c \vee 2\|f_1 - f_2\|_{L^\infty(X)}\right][\ell(s, f_1) + \ell(s, f_2)]. \tag{4.17}$$

Proof Recall that

$$\phi_c(x) = \frac{x^2}{2}\mathbb{I}_{|x|\leqslant c} + c\big(|x| - \frac{c}{2}\big)\mathbb{I}_{|x|>c}.$$

In the rest of the proof, for any $x \in \mathcal{X}$, let $\ell_X(u) = \mathbb{E}\left[\phi_c(Y-u) - \phi_c(Y)|X = x\right]$. Let $s : x \mapsto \operatorname{argmin}_{u\in\mathbb{R}} \ell_X(u)$; s is a risk minimizer. Then $\phi_c'(x) = \operatorname{sgn}(x)(|x| \wedge c)$ and $\phi_c''(x) = \mathbb{I}_{|x|\leqslant c}$. By differentiating under the expectation, for any u such that $|u - s(x)| \leqslant \frac{c}{2}$,

$$\begin{aligned}
\ell_x''(u) &= \partial_u^2 \mathbb{E}\left[\phi_c(Y-u) - \phi_c(Y)|X = x\right] \\
&= \mathbb{E}\left[\phi_c''(Y-u)|X = x\right] \\
&= \mathbb{P}\left[|Y-u| \leqslant c|X = x\right] \\
&\geqslant \mathbb{P}\left[|Y - s(x)| + |u - s(x)| \leqslant c|X = x\right] \\
&\geqslant \mathbb{P}\left[|Y - s(x)| \leqslant \frac{c}{2}|X = x\right] \\
&\geqslant \eta
\end{aligned}$$

Since $s(x)$ is a local minimum, it follows that, for any $u \in \left[s(x) - \frac{c}{2}; s(x) + \frac{c}{2}\right]$,

$$\begin{aligned}
|\ell_x'(u)| &= \left|\ell_x'(u) - \ell_x'(s(x))\right| \\
&= \left|\int_{s(x)}^{u} \ell_x''(t)dt\right| \\
&\geqslant \eta|u - s(x)|
\end{aligned}$$

By lemma 3.9.2 of Chapter 3, it follows that for any $(u, v) \in \mathbb{R}^2$,

$$(u-v)^2 \leqslant \left(\frac{4}{\eta} \vee \left(\frac{8}{c\eta}|u-v|\right)\right)\left(\ell_X(u) + \ell_X(v) - 2\ell_X(s(x))\right). \tag{4.18}$$

Now using equation (4.18) with $u = f_1(X)$, $v = f_2(X)$, $x = X$ and taking expectations, we have:

$$c^2\left\|f_1 - f_2\right\|_{L^2(X)}^2 \leqslant \frac{4c}{\eta}\left[c \vee 2\left\|f_1 - f_2\right\|_{L^\infty(X)}\right]\left[\ell(s, f_1) + \ell(s, f_2)\right]. \tag{4.19}$$

■

We are now ready to obtain functions $(\hat{w}_{i,j})_{(i,j)\in\{1;2\}^2}$ such that $H\left(\hat{w}_{i,1}, \hat{w}_{i,2}, (\hat{t}_k)_{1\leqslant k\leqslant K}\right)$ holds. In the following, fix $K \in [|1; n_t|]$ and write $\kappa = \kappa(K)$ for short.

By Proposition 4.6.3, for all $(k, l) \in [|1; K|]^2$,

$$c^2\left\|\hat{t}_k - \hat{t}_l\right\|_{L^2(X)}^2 \leqslant \frac{4c}{\eta}\left[c \vee \left(2\left\|\hat{t}_k - \hat{t}_l\right\|_{L^\infty(X)}\right)\right]\left[\ell(s, \hat{t}_k) + \ell(s, \hat{t}_l)\right]. \tag{4.20}$$

Hence, by claim 4.6.0.1, for all $(k,l) \in [|1;K|]^2$,

$$c^2 \left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^2(X)} \leqslant \frac{4c}{\eta}\left[c \vee (2\sqrt{2}\kappa \left\|\hat{t}_k - \hat{t}_l\right\|_{L^2(X)})\right]\left[\ell(s,\hat{t}_k) + \ell(s,\hat{t}_l)\right] \tag{4.21}$$

By lemma 4.6.2 with $v = c^2 \left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^2(X)}$, $r = \frac{4c^2}{\eta}$ and $s = \frac{8\sqrt{2}\kappa}{\eta}$,

$$c^2 \left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^2(X)} \leqslant \left(\hat{w}_A(\sqrt{\ell(s,\hat{t}_k)}) + \hat{w}_A(\sqrt{\ell(s,\hat{t}_l)})\right)^2, \tag{4.22}$$

where

$$\hat{w}_A(x) = \left(\frac{2c}{\sqrt{\eta}}x\right) \vee \frac{8\sqrt{2}\kappa}{\eta}x^2. \tag{4.23}$$

Now,

$$\begin{aligned} c^2 \left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^\infty(X)} &\leqslant 2c^2\kappa^2 \left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^2(X)} \\ &\leqslant \frac{8c\kappa^2}{\eta}\left[c \vee 2\left\|\hat{t}_k - \hat{t}_l\right\|_{L^\infty(X)}\right]\left[\ell(s,\hat{t}_k) + \ell(s,\hat{t}_l)\right] \text{ by Proposition 4.6.3.} \end{aligned}$$

By lemma 4.6.2 with $v = c^2 \left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^\infty(X)}$, $s = \frac{16\kappa^2}{\eta}$, $r = \frac{8c^2\kappa^2}{\eta}$,

$$c^2 \left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^\infty(X)} \leqslant \left(\hat{w}_{2,2}(\sqrt{\ell(s,\hat{t}_k)}) + \hat{w}_{2,2}(\sqrt{\ell(s,\hat{t}_k)})\right)^2 \tag{4.24}$$

where

$$\hat{w}_{2,2}(x) = \frac{2c\sqrt{2}\kappa}{\sqrt{\eta}}x \vee \frac{16\kappa^2}{\eta}x^2 = \sqrt{2}\kappa\hat{w}_A. \tag{4.25}$$

Because the Huber loss ϕ_c is $c-$Lipschitz,

$$\forall u, v \in \mathbb{R}^d, |\phi_c(Y-u) - \phi_c(Y-v)| \leqslant c|u-v|$$

Therefore by (4.22) and (4.24),

$$\begin{aligned} \mathbb{E}\left[\left(\phi_c(Y - \hat{t}_k(X)) - \phi_c(Y - \hat{t}_l(X))\right)^k\right] &\leqslant \left(c^2\left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^2(X)}\right)\left(c\left\|\hat{t}_k - \hat{t}_l\right\|_{L^\infty(X)}\right)^{k-2} \\ &\leqslant \left(\hat{w}_A(\sqrt{\ell(s,\hat{t}_k)}) + \hat{w}_A(\sqrt{\ell(s,\hat{t}_l)})\right)^2 \\ &\quad \times \left(\sqrt{2}\kappa\hat{w}_A(\sqrt{\ell(s,\hat{t}_k)}) + \sqrt{2}\kappa\hat{w}_A(\sqrt{\ell(s,\hat{t}_l)})\right)^{k-2}, \end{aligned}$$

which proves $H\left(\hat{w}_A, \sqrt{2}\kappa\hat{w}_A, (\hat{t}_k)_{1\leqslant k\leqslant K}\right)$. Now going back to equation (4.17), by Definition 4.6.1,

$$\begin{aligned} c^2\left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^2(X)} &\leqslant \frac{4c}{\eta}\left[c \vee 2\hat{\beta}\right]\left[\ell(s,\hat{t}_k) + \ell(s,\hat{t}_l)\right] \\ &\leqslant \left(\hat{w}_B(\sqrt{\ell(s,\hat{t}_k)}) + \hat{w}_B(\sqrt{\ell(s,\hat{t}_l)})\right)^2 \end{aligned} \tag{4.26}$$

where

$$\hat{w}_B(x) = \left(\frac{2c}{\sqrt{\eta}} \vee 2\sqrt{\frac{2c\hat{\beta}}{\eta}}\right) x. \tag{4.27}$$

Moreover,

$$\begin{aligned} c^2\left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^\infty(X)} &\leqslant 2\kappa^2 c^2\left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^2(X)} \\ &\leqslant 2\kappa^2\left(\hat{w}_B(\sqrt{\ell(s,\hat{t}_k)}) + \hat{w}_B(\sqrt{\ell(s,\hat{t}_l)})\right)^2 \text{ by (4.26)} \end{aligned}$$

Therefore, by (4.26),

$$\begin{aligned} P|\gamma(\hat{t}_k) - \gamma(\hat{t}_l)|^k &\leqslant \mathbb{E}\left[\left(\phi_c(Y - \hat{t}_k(X)) - \phi_c(Y - \hat{t}_l(X))\right)^2\right]\left\|\phi_c(Y - \hat{t}_k(X)) - \phi_c(Y - \hat{t}_l(X))\right\|^{k-2}_\infty \\ &\leqslant \left(c^2\left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^2(X)}\right)\left(c\left\|\hat{t}_k - \hat{t}_l\right\|_{L^\infty(X)}\right)^{k-2} \\ &\leqslant \left(\hat{w}_B(\sqrt{\ell(s,\hat{t}_k)}) + \hat{w}_B(\sqrt{\ell(s,\hat{t}_l)})\right)^2 \\ &\quad \times \left(\sqrt{2}\kappa\hat{w}_B(\sqrt{\ell(s,\hat{t}_k)}) + \sqrt{2}\kappa\hat{w}_B(\sqrt{\ell(s,\hat{t}_l)})\right)^{k-2}, \end{aligned}$$

which proves $H\left(\hat{w}_B, \sqrt{2}\kappa\hat{w}_B, (\hat{t}_k)_{1\leqslant k\leqslant K}\right)$.

4.6.3 Conclusion of the proof

We have proved that $H\left(\hat{w}_B, \sqrt{2}\kappa\hat{w}_B, (\hat{t}_k)_{1\leqslant k\leqslant K}\right)$ and $H\left(\hat{w}_A, \sqrt{2}\kappa\hat{w}_A, (\hat{t}_k)_{1\leqslant k\leqslant K}\right)$ hold, where $\hat{w}_B$ is given by equation (4.27) and $\hat{w}_A$ is defined by equation (4.23). It remains to apply Theorem 3.7.3 of Chapter 3 and to express the remainder term as a simple function of $c, n_v, n_t, \kappa, L, R, K$ and α. We recall here the definition of the operator δ used in the statement of that theorem.

Definition 4.6.4 *For any function $h : \mathbb{R}_+ \mapsto \mathbb{R}_+$ and any $\xi > 0$, let*

$$\delta(h, \xi) = \inf\{x \in \mathbb{R}_+ : \forall u \geqslant x, h(u) \leqslant \xi u^2\}.$$

The following lemma will facilitate the computation of $\delta(\hat{w}_A, \cdot)$.

Lemma 4.6.5 *Let $r > 0, s > 0$ and $h_{r,s}(x) = (\sqrt{r}x) \vee sx^2$. Then $\delta(h_{r,s}, \xi) < \infty$ if and only if $\xi \geqslant s$ and then $\delta(h_{r,s}, \xi) = \frac{\sqrt{r}}{\xi}$.*

Proof To find $\delta(h_{r,s}, \xi)$, notice that given the definition of $\delta(h_{r,s}, \xi)$, the condition $s \leqslant \xi$ is obviously necessary for the infimum to be finite. Assume now that $\xi \geqslant s$. For any $u \geqslant \frac{\sqrt{r}}{\xi}$, then $\xi u^2 \geqslant \sqrt{r}u$ as well as $\xi u^2 \geqslant su^2$ (since we assumed $\xi \geqslant s$), therefore $\xi u^2 \geqslant h_{r,s}(u)$. Thus by definition, $\delta(h_{r,s}, \xi) \leqslant \frac{\sqrt{r}}{\xi}$ (in particular, $\delta(h_{r,s}, \xi)$ is finite). Furthermore, by definition of $\delta(h_{r,s}, \xi)$, $\sqrt{r}\delta(h_{r,s}, \xi) \leqslant \xi\delta(h_{r,s}, \xi)^2$, that is $\delta(h_{r,s}, \xi) \geqslant \frac{\sqrt{r}}{\xi}$. ■

The following claim can now be proved.

Claim 4.6.5.1 *If $K \in [|1; n_t|]$ and $b > 1$ are such that*

$$\kappa(K) \leqslant \frac{\eta}{8}\sqrt{\frac{n_v}{8b\log K}}, \tag{4.28}$$

then applying Agghoo to the collection $(\mathcal{A}_k)_{1\leqslant k\leqslant K}$ yields the following oracle inequality.

$$(1-\theta)\mathbb{E}[\ell(s, \widehat{f}_{\mathcal{T}}^{\text{ag}})] \leqslant (1+\theta)\mathbb{E}[\min_{1\leqslant k\leqslant K} \ell(s, \hat{t}_k)] + 24\theta b\frac{c\log K}{\eta n_v}\left[c + \frac{c + 2LRn_t^\alpha}{K^{\theta^2 b-1}}\right].$$

Proof Theorem 3.7.3 of Chapter 3 applies with $\hat{w}_{1,1} = \hat{w}_B, \hat{w}_{1,2} = \sqrt{2}\kappa\hat{w}_B, \hat{w}_{2,1} = \hat{w}_A, \hat{w}_{2,2} = \sqrt{2}\kappa\hat{w}_A$, $x = (\theta^2 b - 1)\log K$ and it remains to bound the remainder terms $(R_{2,i})_{1\leqslant i\leqslant 4}$. Now assume that equation (4.28) holds.

Bound on $R_{2,1}(\theta) = \sqrt{2}\theta\mathbb{E}\left[\delta^2\left(\hat{w}_A, \frac{\theta}{2}\sqrt{\frac{n_v}{\theta^2 b\log K}}\right)\right]$

By (4.23), we can apply lemma 4.6.5 with $s = \frac{8\sqrt{2}\kappa}{\eta}$, $r = \frac{4c^2}{\eta}$ and $\xi = \frac{1}{2}\sqrt{\frac{n_v}{b\log K}}$. By (4.28),

$$s = \frac{8}{\eta}\sqrt{2}\kappa \leqslant \sqrt{\frac{n_v}{4b\log K}} = \xi.$$

It follows by lemma 4.6.5 that

$$\delta\left(\hat{w}_A, \sqrt{\frac{n_v}{4b\log K}}\right) = \frac{2c}{\sqrt{\eta}}\sqrt{\frac{4b\log K}{n_v}}.$$

Hence,

$$R_{2,1}(\theta) \leqslant \sqrt{2}\theta\frac{4c^2}{\eta}\frac{4b\log K}{n_v} \leqslant 23\theta b\frac{c^2\log K}{\eta n_v} \tag{4.29}$$

Bound on $R_{2,2}(\theta) = \frac{\theta^2}{2}\mathbb{E}\left[\delta^2\left(\sqrt{2}\kappa\hat{w}_A, \frac{\theta^2}{4}\frac{n_v}{\theta^2 b\log K}\right)\right]$

By (4.23), we can apply lemma 4.6.5 with $s = \frac{16\kappa^2}{\eta}$, $r = \frac{8c^2\kappa^2}{\eta}$ and $\xi = \frac{n_v}{4b\log K}$. By (4.28) and since $\eta \leqslant 1$,

$$s = \frac{16\kappa^2}{\eta} \leqslant \frac{16}{\eta}\frac{\eta^2}{64}\frac{n_v}{4b\log K} \leqslant \frac{n_v}{4b\log K} = \xi.$$

Therefore

$$\begin{aligned}\delta\left(\sqrt{2}\kappa\hat{w}_A, \frac{\theta^2}{4}\frac{n_v}{\theta^2 b\log K}\right) &\leqslant \frac{2c\sqrt{2}\kappa}{\sqrt{\eta}}\frac{4b\log K}{n_v} \text{ by lemma 4.6.5}\\ &\leqslant \frac{c}{2}\sqrt{\frac{\eta b\log K}{n_v}} \text{ by (4.28).}\end{aligned}$$

Hence, since $\theta, \eta \in [0;1]$,

$$R_{2,2}(\theta) \leqslant \frac{\theta^2}{2}\frac{c^2}{4}\frac{\eta b\log K}{n_v} \leqslant \frac{\theta b}{8}\frac{c^2\log K}{n_v}. \tag{4.30}$$

Bound on $R_{2,3}(\theta) = \frac{1}{K^{\theta^2 b-1}}\left(\theta + \frac{2\left[1+\log(K)\right]}{\theta}\right)\mathbb{E}\left[\delta^2\big(\hat{w}_B, \sqrt{n_v}\big)\right]$

By (4.27), $x \to \frac{\hat{w}_B(x)}{x}$ is constant and, in particular, non-increasing. Therefore, $\delta\big(\hat{w}_B, \sqrt{n_v}\big)$ is the unique nonnegative solution to the equation

$$\hat{w}_B(x) = \sqrt{n_v}x^2 \iff \left(\frac{2c}{\sqrt{\eta}} \vee 2\sqrt{\frac{2c\hat{\beta}}{\eta}}\right)x = \sqrt{n_v}x^2.$$

It follows that

$$\delta\big(\hat{w}_B, \sqrt{n_v}\big) = \left(\frac{2c}{\sqrt{\eta}} \vee 2\sqrt{\frac{2c\hat{\beta}}{\eta}}\right)\frac{1}{\sqrt{n_v}}. \tag{4.31}$$

and

$$\begin{aligned}\delta^2\big(\hat{w}_B, \sqrt{n_v}\big) &= \left(\frac{4c^2}{\eta} \vee \frac{8c\hat{\beta}}{\eta}\right)\frac{1}{n_v}\\ &= 4\frac{c(c \vee 2\hat{\beta})}{\eta n_v}.\end{aligned} \tag{4.32}$$

As $n_t \geqslant 3$, we can assume that $K \geqslant 3$, hence $\log K \geqslant 1$ and

$$\theta + \frac{2(1+\log K)}{\theta} \leqslant \frac{5\log K}{\theta}.$$

Since $\theta \geqslant \frac{1}{\sqrt{b}}$, $\frac{1}{\theta} \leqslant \theta b$ and therefore

$$\theta + \frac{2(1+\log K)}{\theta} \leqslant 5\theta b \log K. \tag{4.33}$$

By equations (4.32) and (4.33),

$$R_{2,3}(\theta) \leqslant 20\theta b \frac{c \log K}{\eta n_v} \frac{c + 2\mathbb{E}[\hat{\beta}]}{K^{\theta^2 b - 1}} \tag{4.34}$$

Bound on $R_{2,4}(\theta) = \frac{1}{K^{\theta^2 b-1}} \left(\theta + \frac{2(1+\log K)+\log^2 K}{\theta}\right) \mathbb{E}\left[\delta^2\left(\sqrt{2}\kappa\hat{w}_B, n_v\right)\right]$

$\delta^2\left(\sqrt{2}\kappa\hat{w}_B, \sqrt{n_v}\right)$ is the unique nonnegative solution to the equation

$$\sqrt{2}\kappa\hat{w}_B(x) = \sqrt{n_v}x^2 \iff \left(\frac{2c}{\sqrt{\eta}} \vee 2\sqrt{\frac{2c\hat{\beta}}{\eta}}\right) \sqrt{2}\kappa x = n_v x^2,$$

which yields

$$\begin{aligned}
\delta\left(\sqrt{2}\kappa\hat{w}_B, n_v\right) &= \left(\frac{2c}{\sqrt{\eta}} \vee 2\sqrt{\frac{2c\hat{\beta}}{\eta}}\right) \frac{\sqrt{2}\kappa}{n_v} \\
&\leqslant \frac{1}{8}\sqrt{\frac{\eta}{b n_v \log K}} \left(c \vee \sqrt{2c\hat{\beta}}\right) \text{ by (4.28) .}
\end{aligned}$$

Since $\eta \leqslant 1$ and $b \geqslant 1$, it follows that

$$\delta^2\left(\sqrt{2}\kappa\hat{w}_B, n_v\right) \leqslant \frac{\eta}{64} \frac{c^2 \vee 2c\hat{\beta}}{b n_v \log K} \tag{4.35}$$

$$\tag{4.36}$$

By equation (4.33), and since $\frac{1}{\theta} \leqslant \theta b$,

$$\begin{aligned}
\theta + \frac{2(1+\log K) + \log^2 K}{\theta} &\leqslant 5\theta b \log K + \theta b \log^2 K \\
&\leqslant 6\theta b \log^2 K \text{ since } K \geqslant 3.
\end{aligned}$$

Therefore, since $b \geqslant 1$ and $\eta \in [0; 1]$,

$$R_{2,4}(\theta) \leqslant \frac{6\theta\eta}{64} \frac{c \log K}{n_v} \frac{c + 2\mathbb{E}[\hat{\beta}]}{K^{\theta^2 b - 1}} \leqslant \frac{3\theta b}{32} \frac{c \log K}{\eta n_v} \frac{c + 2\mathbb{E}[\hat{\beta}]}{K^{\theta^2 b - 1}}. \tag{4.37}$$

Conclusion Summing up equations (4.29), (4.30), (4.34) and (4.37), Theorem 3.7.3 of Chapter 3 implies that assuming equation (4.28) holds for K, for all $\theta \in \left[\frac{1}{\sqrt{b}}; 1\right]$,

$$(1-\theta)\mathbb{E}[\ell(s, \widehat{f}^{\text{ag}}_{\mathcal{T}})] \leqslant (1+\theta)\mathbb{E}[\min_{1\leqslant k\leqslant K} \ell(s, \hat{t}_k)] + 24\theta b \frac{c \log K}{\eta n_v} \left[c + \frac{c + 2\mathbb{E}[\hat{\beta}]}{K^{\theta^2 b - 1}}\right]. \quad (4.38)$$

It follows by claim 4.6.1.1 that

$$(1-\theta)\mathbb{E}[\ell(s, \widehat{f}^{\text{ag}}_{\mathcal{T}})] \leqslant (1+\theta)\mathbb{E}[\min_{1\leqslant k\leqslant K} \ell(s, \hat{t}_k)] + 24\theta b \frac{c \log K}{\eta n_v} \left[c + \frac{c + 16LRn_t^{\alpha}}{K^{\theta^2 b - 1}}\right].$$

■

Let K satisfy assumption (4.6) from Theorem 4.3.2. Take now $b = b_0 \frac{\log n_t}{\log K}$. Then equation (4.28) holds for this value of b. Moreover, since $K \leqslant n_t$, $b_0 \geqslant b$ and therefore $\theta \in \left[\frac{1}{\sqrt{b_0}}; 1\right] \implies \theta \in \left[\frac{1}{\sqrt{b}}; 1\right]$. Thus, claim 4.6.5.1 implies that for any $\theta \in \left[\frac{1}{\sqrt{b_0}}; 1\right]$,

$$(1-\theta)\mathbb{E}[\ell(s, \widehat{f}^{\text{ag}}_{\mathcal{T}})] \leqslant (1+\theta)\mathbb{E}[\min_{1\leqslant k\leqslant K} \ell(s, \hat{t}_k)] + 24\theta b_0 \frac{c \log n_t}{\eta n_v} \left[c + \frac{cK}{n_t^{\theta^2 b_0}} + \frac{16KLR}{n_t^{\theta^2 b_0 - \alpha}}\right].$$

This proves Theorem 4.3.2.

4.7 Applications of Theorem 4.3.2

4.7.1 Proof of corollary 4.3.3

Let $\bar{x} \in \text{supp}(X - EX)$. Let $\theta \in \mathbb{R}^p$ be such that $\|\theta\|_0 \leqslant 2K$ and let $I = \{i : \theta_i \neq 0\}$. For all $i \in [|1; p|]$, $\bar{x}_i \in [-1; 1]$, hence

$$\begin{aligned} \left|\sum_{i=1}^{p} \theta_i \bar{x}_i\right| &= \left|\sum_{i\in I} \theta_i \bar{x}_i\right| \\ &\leqslant \sqrt{\sum_{i\in I} \bar{x}_i^2} \sqrt{\sum_{i\in I} \theta_i^2} \\ &\leqslant \sqrt{2K} \sqrt{\sum_{i\in I} \theta_i^2}. \end{aligned} \quad (4.39)$$

On the other hand, let $\bar{X}_i = X_i - EX_i$. Then by independence of the X_i,

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{p}\theta_i\bar{X}_i\right)^2\right] &= \sum_{i=1}^{p}\theta_i^2\mathrm{Var}(X_i)\\ &= \sum_{i=1}^{p}\theta_i^2 p_i(1-p_i)\\ &\geqslant \min_{1\leqslant i\leqslant p}\{p_i(1-p_i)\}\times\sum_{i\in I}\theta_i^2. \end{aligned} \tag{4.40}$$

Combining inequalities (4.39) and (4.40) yields

$$\sup_{\bar{x}\in\mathrm{supp}(X-EX)}\left|\sum_{i=1}^{p}\theta_i\bar{x}_i\right| \leqslant \sqrt{\frac{2K}{\min_{1\leqslant i\leqslant p}p_i(1-p_i)}}\sqrt{\mathbb{E}\left[\langle\theta,\bar{X}\rangle^2\right]}.$$

4.7.2 Proof of Proposition 4.3.5

We associate to the original piecewise constant regression problem an ordinary sparse regression problem. Define the linear operator

$$\begin{aligned} S : \mathbb{R}^d &\to \mathbb{R}^d\\ (x_j)_{1\leqslant j\leqslant d} &\mapsto \left(\sum_{i=1}^{j}x_i\right)_{1\leqslant j\leqslant d}\end{aligned}$$

and:

$$\begin{aligned}\Delta : \mathbb{R}^d &\to \mathbb{R}^{d-1}\\ (x_j)_{1\leqslant j\leqslant d} &\mapsto \left((x_j - x_{j-1})_{j\geqslant 2}\right).\end{aligned}$$

Fo any $k \in [|1;d|]$, let also $E_k = \cup_{j=k}^{d} I_j$. Then t_u has k jumps if and only if $\|\Delta(u)\|_0 = k$, moreover we have the representation:

$$t_u = u_1\mathbb{I} + \sum_{j=2}^{d}\Delta(u)_j\mathbb{I}_{E_j}. \tag{4.41}$$

Equivalently,

$$t_{S(u)} = \sum_{j=1}^{d}u_j\mathbb{I}_{E_j} \tag{4.42}$$

It follows that the original jump detection problem is equivalent to a sparse regression problem with covariate vector $X_i = \left(\mathbb{I}_{E_j}(U_i)\right)_{2\leqslant j\leqslant d}$, where the non-penalized

intercept corresponds to the component u_1 of the original problem. The set of $X-$measurable functions coïncides with the set of linear functions on the finite set $\mathrm{supp}(X)$, hence a Bayes estimator is $s : x \mapsto u_{*,1} + \langle \Delta(u_*), x \rangle)$, where $u_* = \mathrm{argmin}_{u \in \mathbb{R}^d} \mathbb{E}\left[\phi_c(Y - t_u(X)\right]$. Let $\hat{\theta}_k = \Delta(\hat{u}_k)$ be the sparse estimator associated to $\hat{u}_k$. This allows to apply Theorem 4.3.2, provided that $(\hat{\theta}_k)_{1 \leqslant k \leqslant d}$ satisfies also the two last points of hypothesis 4.2.1. Let us now prove this. Write $\hat{\theta}_k, \hat{u}_k$ for $\theta_k(\hat{D}_{n_t}), \hat{u}_k(D_{n_t})$ to simplify notation. Using the notation of Definition 4.3.4, we have:

$$\left\|\hat{\theta}_k\right\|_1 = \|\Delta(\hat{u}_k)\|_1 = \sum_{r=2}^{k} |\hat{u}_{k,j_r(\hat{u}_k)} - \hat{u}_{k,j_{r-1}(\hat{u}_k)}| \leqslant 2 \sum_{r=1}^{k} |\hat{u}_{k,j_r(\hat{u}_k)}|.$$

Hence, by the triangular inequality,

$$\left\|\hat{\theta}_k\right\|_1 \leqslant 2 \sum_{r=0}^{k} \min_{i : X_i \in A_r(\hat{u}_k)} |Y_i| + |Y_i - \hat{u}_{j_r(\hat{u}_k)}|.$$

Using the inequality $|x| \leqslant c + \frac{1}{c}\phi_c(x)$ yields

$$\begin{aligned}\left\|\hat{\theta}_k\right\|_1 &\leqslant 2 \sum_{r=0}^{k} \min_{i : X_i \in A_r(\hat{u}_k)} |Y_i| + c + \frac{1}{c}\phi_c\left(Y_i - \hat{u}_{j_r(\hat{u}_k)}\right) \\ &\leqslant 2 \sum_{r=0}^{k} c + \sum_{i : X_i \in A_r(\hat{u}_k)} |Y_i| + \frac{1}{c}\phi_c\left(Y_i - \hat{u}_{j_r(\hat{u}_k)}\right).\end{aligned}$$

Hence, by Definition 4.3.4,

$$\begin{aligned}\left\|\hat{\theta}_k\right\|_1 &\leqslant 2 \sum_{r=0}^{k} c + \sum_{i : X_i \in A_r(\hat{u}_k)} |Y_i| + \frac{1}{c}\phi_c\left(Y_i\right) \\ &\leqslant 2kc + 4 \sum_{i=1}^{n_t} |Y_i|.\end{aligned}$$

It follows that

$$\mathbb{E}\left[\sup_{1 \leqslant k \leqslant d-1} \left\|\hat{\theta}_k\right\|_1\right] \leqslant 2dc + 4n_t \mathbb{E}[|Y|], \tag{4.43}$$

so since $d \leqslant n_v \leqslant n_t$, the second point of hypothesis 4.2.1 is satisfied with $\alpha = 1$ and $L = 2c + 4\mathbb{E}[|Y|]$. Let $u \in \mathbb{R}^d$. For any $y \in \mathbb{R}$, by convexity of ϕ_c,

$$\begin{aligned} y \in \operatorname*{argmin}_{q\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} \phi_c(Y_i - t_u(U_i) - q) &\iff \frac{1}{n}\sum_{i=1}^{n} \phi_c'\left(Y_i - y - t_u(U_i)\right) = 0 \\ &\iff \frac{1}{n}\sum_{i=1}^{n} \phi_c'\left(Y_i - y - u_1 - \sum_{j=2}^{d} \Delta(u)_j \mathbb{I}_{E_j}(U_i)\right) = 0 \\ &\iff u_1 + y \in \operatorname*{argmin}_{x} \frac{1}{n}\sum_{i=1}^{n} \phi_c\left(Y_i - x - \langle \Delta(u), X\rangle\right). \end{aligned}$$

Hence, using the notations of hypothesis 4.2.1, for any $u \in \mathbb{R}^d$,

$$\hat{Q}\big((X_i, Y_i)_{1\leqslant i\leqslant n}, \Delta(u)\big) = u_1 + \operatorname*{argmin}_{q\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} \phi_c(Y_i - t_u(U_i) - q). \tag{4.44}$$

By Definition 4.3.4, $\hat{u}_k \in \hat{C}_t(D_n, \hat{u}_k)$, therefore

$$\begin{aligned} \sum_{i=1}^{n} \phi_c'(Y_i - t_{\hat{u}_k}(U_i)) &= \sum_{r=1}^{k} \sum_{i:X_i\in A_r(\hat{u}_k)} \phi_c'\left(Y_i - \hat{u}_{k,j_r(\hat{u}_k)}\right) \\ &= 0, \end{aligned}$$

therefore $0 \in \operatorname{argmin}_{q\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} \phi_c(Y_i - t_{\hat{u}_k}(U_i) - q)$ by convexity of ϕ_c. Hence, by equation (4.44), $\hat{u}_{k,1} \in \hat{Q}\big((X_i, Y_i)_{1\leqslant i\leqslant n}, \hat{\theta}_k\big)$. Let now $u_1' \in \hat{Q}\big((X_i, Y_i)_{1\leqslant i\leqslant n}, \hat{\theta}_k\big)$, and let $y = u_1' - \hat{u}_{k,1}$. Assume by contradiction that $(\hat{u}_{k,j} + y)_{1\leqslant j\leqslant d} \notin \hat{C}_t(D_n, \hat{u}_k)$. Then there exists $l \in [|1; k|]$ such that

$$\sum_{i:X_i\in A_l(\hat{u}_k)} \phi_c\left(Y_i - \hat{u}_{k,j_l(\hat{u}_k)} - y\right) > \sum_{i:X_i\in A_l(\hat{u}_k)} \phi_c\left(Y_i - \hat{u}_{k,j_l(\hat{u}_k)}\right),$$

while for all $r \neq l$, since by assumption $\hat{u}_k \in \hat{C}_t(D_n, \hat{u}_k)$,

$$\sum_{i:X_i\in A_r(\hat{u}_k)} \phi_c\left(Y_i - \hat{u}_{k,j_r(\hat{u}_k)} - y\right) \geqslant \sum_{i:X_i\in A_r(\hat{u}_k)} \phi_c\left(Y_i - \hat{u}_{k,j_r(\hat{u}_k)}\right).$$

It follows that:

$$\begin{aligned}\sum_{i=1}^{n}\phi_c(Y_i - t_{\hat{u}_k}(U_i) - y) &= \sum_{r=1}^{k}\sum_{i:X_i\in A_r(\hat{u}_k)}\phi_c\left(Y_i - \hat{u}_{k,j_r(\hat{u}_k)} - y\right)\\ &> \sum_{r=1}^{k}\sum_{i:X_i\in A_r(\hat{u}_k)}\phi_c\left(Y_i - \hat{u}_{k,j_r(\hat{u}_k)}\right)\\ &= \sum_{i=1}^{d}\phi_c(Y_i - t_{\hat{u}_k}(U_i)).\end{aligned}$$

Yet by equation (4.44), $y \in \operatorname{argmin}_{q\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}\phi_c(Y_i - t_{\hat{u}_k}(U_i) - q)$, which yields a contradiction. Therefore, $(\hat{u}_{k,j} + y)_{1\leqslant j\leqslant d} \in \hat{C}_t(D_n, \hat{u}_k)$.

Thus,

$$\begin{aligned}\left|u'_1 + \langle\hat{\theta}_k, \frac{1}{n}\sum_{i=1}^{n}X_i\rangle\right| &= \left|\hat{u}_{k,1} + y + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=2}^{d}\Delta(\hat{u}_k)_j\mathbb{I}_{E_j}(U_i)\right|\\ &= \left|y + \frac{1}{n}\sum_{i=1}^{n}t_{\hat{u}_k}(U_i)\right|\\ &= \left|\sum_{j=1}^{d}P_n(I_j)\left[\hat{u}_{k,j} + y\right]\right|\\ &\geqslant \left|\sum_{j=1}^{d}P_n(I_j)\hat{u}_{k,j}\right| \text{ by Definition 4.3.4, equation (4.8)}\\ &= \left|\frac{1}{n}\sum_{i=1}^{n}t_{\hat{u}_k}(U_i)\right|\\ &= \left|\hat{u}_{k,1} + \langle\hat{\theta}_k, \frac{1}{n}\sum_{i=1}^{n}X_i\rangle\right|.\end{aligned}$$

This shows that the linear regressor $D_n \to x \to \hat{u}_{k,1} + \langle\Delta(\hat{u}_k), x\rangle$ satisfies also the last point of hypothesis 4.2.1.

It remains to check the assumptions of Theorem 4.3.2. We first remark that

$R \leqslant \sup_{x\in\mathbb{R}} \max_j |\mathbb{I}_{E_j}(x) - P(E_j)| \leqslant 1$. We now bound $\kappa(K)$. For any $\theta \in \mathbb{R}^d$,

$$\begin{aligned}\sum_{j=k}^{d} \theta_k \bar{X}_k &= \sum_{k=1}^{d} \theta_k \left(\mathbb{I}_{E_k}(U) - P(E_k)\right) \\ &= \sum_{k=1}^{d} \theta_k \sum_{j=1}^{d} \mathbb{I}_{j\geqslant k} \left(\mathbb{I}_{I_j}(U) - P(I_j)\right) \\ &= \sum_{j=1}^{d} \left(\sum_{k=1}^{j} \theta_k\right) \left(\mathbb{I}_{I_j}(U) - P(I_j)\right)\end{aligned}$$

Thus, for any $\theta \in \mathbb{R}^d$, $\langle \bar{X}, \theta\rangle$ is in span $\left(\mathbb{I}_{I_j}(U)\right)_{1\leqslant j\leqslant d}$ (constant functions also belong to this space). Now, for any $Z = \sum_{j=1}^{d} z_j \mathbb{I}_{I_j}(U) \in$ span $\left(\mathbb{I}_{I_j}(U)\right)_{1\leqslant j\leqslant d}$, by assumption (4.10) of Proposition 4.3.5,

$$\|Z\|_\infty \leqslant \max_{1\leqslant j\leqslant d} |z_j| \leqslant \sqrt{\sum_{j=1}^{d} z_j^2} \leqslant \frac{1}{\sqrt{\min_{1\leqslant j\leqslant d} P(I_j)}} \sqrt{\sum_{j=1}^{d} z_j^2 P(I_j)} \leqslant \sqrt{\frac{\eta^2 n_v}{1536 \log^2 n_t}} \mathbb{E}[Z^2]^{\frac{1}{2}}.$$

This yields $\kappa(d) \leqslant \sqrt{\frac{\eta^2 n_v}{1536 \log^2 n_t}}$. Let $b_0 = 3 \log n_t$, then $\kappa(d) \leqslant \frac{\eta}{8 \log n_t} \sqrt{\frac{n_v}{24}} = \frac{\eta}{8} \sqrt{\frac{n_v}{8 b_0 \log n_t}}$. Finally, applying Theorem 4.3.2 with $K = d$, $R = 1$, $L = 2c + 4\mathbb{E}[|Y|]$, $\alpha = 1$ (by equation (4.43)), $b_0 = 3 \log n_t$ and $\theta = \frac{1}{\sqrt{\log n_t}}$ yields equation (4.11), proving Proposition 4.3.5.

Chapter 5

A detailed analysis of Agghoo: asymptotic approximation of the hold-out risk estimator

5.1 Introduction

In the previous chapters, we have shown oracle inequalities which certify that the hold-out procedure performs, asymptotically, as well as the best estimators in the given collection. However, these were only upper bounds and we did not have access to the actual order of magnitude of the gap between the risk of the hold-out estimator and the oracle. Moreover, these results made no difference between hold-out and its aggregated version, hence they cannot explain why Agghoo sometimes perform much better than the hold-out or even cross-validation in practice.

In this chapter and the next, we will conduct a detailed study of hold-out and aggregation of hold-outs in a particular setting where it is possible to be more accurate. This setting is that of L^2 density estimation, with the collection of estimators given by *empirical orthogonal projection* on a trigonometric basis, indexed by the number of basis functions used; see Section 5.2.

The basic unit of hold-out aggregation is the individual hold-out estimator it aggregates. The starting point for a detailed study of hold-out aggregation is therefore a detailed study of the individual hold-out estimator. The hold-out procedure itself consists of three steps: risk estimation, selection of a parameter and calculation of the final estimator. The analysis of the hold-out procedure therefore starts with a detailed study of the hold-out risk estimator as a random process. The purpose of this chapter is to show that the hold-out process, properly renormalized, can be asymptotically approximated by a simpler continuous process.

As in any asymptotic study, the choice of the scale to normalize the hold-out

process is crucial to get an interesting limit. From the point of view of studying the hold-out procedure, the point is to obtain an approximation of the risk estimator in the neighborhood of the optimal parameter k_* (which the hold-out seeks to estimate), at the scale of $\hat{k} - k_*$, where $\hat{k}$ denotes the parameter chosen by hold-out, i.e the minimizer of the hold-out *risk estimator*. Since the analysis of the risk-estimator precedes that of its minimizer $\hat{k}$, the correct scaling will have to be guessed, based on heuristic arguments. The results of the next chapter will prove that this guess is correct.

At the relevant scale, we show that the hold-out process behaves as the sum of a convex function f_n and a Brownian motion changed in time W_{g_n}. The approximating process depends on n (this is not a limit), but several inequalities on f_n and g_n show that the approximating process does not become trivial when n tends to $+\infty$. This process is independent from the training data constituting the "training sample" of the hold-out. The interest of this result from the point of view of studying the hold-out procedure is that it allows to make use of the abundant theory available on Brownian motion in order to study the parameter selection step. It is thus an indispensable prerequisite for demonstrating oracle inequalities for the hold-out and for hold-out aggregation, which is the subject of the next chapter.

5.2 L^2 density estimation

Let $s \in L^2([0;1])$ be a probability density function. Given a sample $X_1, \ldots, X_n$ drawn according to the density s, the L^2 density estimation problem consists in constructing an estimator $\hat{s}_n$ that approaches s in terms of the L^2 norm.

Although it is not obvious at first glance (this is not true for the other L^p norms), this non-parametric density estimation problem can be reformulated as a risk minimization problem, with a contrast function: $\gamma(t,x) = \|t\|^2 - 2t(x)$, which yields the *risk* $\mathbb{E}[\gamma(t,X)] = \|t\|^2 - 2\int s(x)t(x)dx = \|t-s\|^2 - \|s\|^2$. It follows that s is indeed the minimizer of the risk corresponding to the γ contrast function, and furthermore

$$\ell(s,t) = \|t-s\|^2 .$$

This problem therefore falls within the theoretical framework of Chapter 3, section 3.2 Therefore, it is possible to use hold-out or aggregated hold-out to select among a collection of estimators.

Here we will consider as a family of non-parametric estimators the *empirical orthogonal series* estimators [41, Section 3.1] on a trigonometric basis. To ease the presentation, we consider only cosine functions, which is equivalent to assuming that s is symmetrical with respect to $\frac{1}{2}$. This restriction is of no fundamental

importance — it is reasonable to conjecture that the results remain valid with the complete trigonometric basis.

For every $j \in \mathbb{N}^*$, let $\psi_j : x \mapsto \sqrt{2}\cos(2\pi jx)$ and let $\psi_0 : x \mapsto 1$. The collection $(\psi_j)_{j\in\mathbb{N}}$ is an orthonormal basis of the subset of $L^2([0;1])$ of functions symmetrical with respect to $\frac{1}{2}$.

Let $D_n = (X_1, ..., X_n)$ be a sample. For any $n \in \mathbb{N}$ and any $T \subset \{1, \ldots, n\}$, we will denote, for any real valued measurable function t,

$$P_n^T(t) = \frac{1}{|T|}\sum_{i\in T} t(X_i).$$

Consider the estimators defined as follows.

Definition 5.2.1 *For all $k \in \mathbb{N}$ and all $T \subset \{1, \ldots, n\}$,*

$$\hat{s}_k^T = \sum_{j=0}^{k} P_n^T(\psi_j)\psi_j,$$

where $\psi_0 = 1$ and for all $j \geqslant 1$, $\psi_j(x) = \sqrt{2}\cos(2\pi jx)$.

The estimators $\hat{s}_k^T$ are empirical risk minimizers on the *models*

$$E_k = \left\{ \sum_{j=0}^{k} v_j\psi_j : v \in \mathbb{R}^{k+1} \right\}.$$

The problem of parameter choice k is therefore a problem of *model selection* within the model collection $(E_k)_{k\geqslant 0}$. Here, the models are nested, meaning $E_k \subset E_{k'}$ for every $k \leqslant k'$.

5.3 Risk estimation for the hold-out

The larger k is, the better the approximation of s by the functions of E_k, but the more difficult it is to estimate the best approximation to s within E_k. The choice of k is therefore subject to a bias-variance trade-off which, if properly carried out, allows adaptation to the smoothness of s, simultaneously reaching the minimax risk on Lipschitz spaces of periodic functions [41, Chapter 7].

Since the risk, except for a constant, is expressed as the expectation of a contrast function

$$P\gamma(\hat{s}_k^T)] := E_X\left[\gamma(\hat{s}_k^T, X)\right] = \left\|\hat{s}_k^T - s\right\|^2 - \|s\|^2,$$

it can be estimated by hold-out as in the previous chapters.

This is the subject of the following definition.

Definition 5.3.1 *Let D_n be an i.i.d sample drawn from the distribution $s(x)dx$. Let $n_t \in \{1, \dots, n-1\}$. Let $T \subset \{1..n\}$ be a subset with cardinality $|T| = n_t$. Then, for all $k \in \mathbb{N}$, we define the hold-out estimator of the risk of $\hat{s}_k$ with training sample indices T by*

$$HO_T(k) = \left\|\hat{s}_k^T\right\|^2 - 2P_n^{T^c}(\hat{s}_k^T).$$

The hold-out risk estimator depends on the choice of a subset T of $\{1, \dots, n\}$, but its distribution depends only on the cardinality of that subsample. The precise choice of a subset T of cardinality n_t will therefore play no role in the sequel. We will therefore denote by T any subset of $\{1, \dots, n\}$ of cardinality n_t.

$\mathrm{HO}_T(\cdot)$ is indeed an estimator since the norm $\|\cdot\|$ is computed with respect to a known dominating measure (in this case the Lebesgue measure) and so does not depend on the distribution of X. Moreover,

$$\mathrm{HO}_T(k) = \left\|\hat{s}_k^T - s\right\|^2 - 2(P_n^{T^c} - P)(\hat{s}_k^T) :$$

the hold-out risk estimator can be expressed as the sum of the excess risk and a centered empirical process.

As a result, $\mathrm{HO}_T(k)$ is an unbiased, consistent estimator of $\left\|\hat{s}_k^T - s\right\|^2$, which approximates $\left\|\hat{s}_k^T - s\right\|^2$ within an error of order $\frac{1}{\sqrt{n-n_t}}$, by a variance calculation. Since the purpose of the hold-out procedure is to select the parameter k, we want to understand the behavior of $\mathrm{HO}_T(\cdot)$ around its optimum $\hat{k}$ and where that optimum lies with high probability. The consistency and unbiasedness of $\mathrm{HO}_T(k)$ suggest that $\hat{k}$ should be "close" to $\mathrm{argmin}_{k\in\mathbb{N}} \left\|\hat{s}_k^T - s\right\|^2$, the true optimal parameter.

Under a few conditions, it is possible to give a simple, deterministic approximant for this optimal parameter. More precisely, let $n_t(n)$ be a sequence of integers such that, for all $n \in \mathbb{N}$, $1 \leqslant n_t(n) \leqslant n$, and define $n_v(n) = n - n_t(n)$. In the following, we shall denote $n_t = n_t(n)$ and $n_v = n_v(n)$ for a generic value of n. Whenever n, n_v, n_t appear in the same expression, it will be understood that $n_t = n_t(n)$ and $n_v = n_v(n) = n - n_t(n)$.

For any $j \in \mathbb{N}$, let $\theta_j = \langle s, \psi_j \rangle$ denote the Fourier coefficients of s on the cosine basis. Suppose that the squared Fourier coefficients θ_j^2 form a non-increasing sequence. The expected L^2 risk can be approximated as follows (see also claim 5.4.3.1):

$$\left\|\hat{s}_k^T - s\right\|^2 \sim \frac{k}{n_t} + \sum_{j=k+1}^{+\infty} \theta_j^2.$$

This leads to the following definition.

Definition 5.3.2 *For all $n \in \mathbb{N}$, let*

$$k_*(n) = \max\left\{k \in \mathbb{N} : \theta_k^2 \geqslant \frac{1}{n}\right\}$$
$$\textit{and} \qquad \mathrm{or}(n) = \inf_{k\in\mathbb{N}}\left\{\sum_{j=k+1}^{+\infty} \theta_j^2 + \frac{k}{n}\right\}.$$

Equivalently,

$$k_*(n) = \max \operatorname*{argmin}_{k\in\mathbb{N}}\left\{\sum_{j=k+1}^{+\infty} \theta_j^2 + \frac{k}{n}\right\}$$
$$\textit{and} \qquad \mathrm{or}(n) = \sum_{j=k_*(n)+1}^{+\infty} \theta_j^2 + \frac{k_*(n)}{n}.$$

$k_*(n_t)$ and $\mathrm{or}(n_t)$ are thus, approximately, the minimizer and the minimum in k of the L^2 risk of the estimators $\hat{s}_k^T$, which explains the name $\mathrm{or}(n_t)$ (oracle). Thus, it is to be expected that the minimizer of the hold-out risk estimator lies close to $k_*(n_t) = k_*(n_t)$. For this reason, $k_*(n_t)$ will be the most relevant value of k_* in the following, and we will often omit the argument n_t, with the understanding that $k_* = k_*(n_t)$.

Assuming to simplify that k_* minimizes the L^2 risk,

$$\mathrm{HO}_T(k) - \mathrm{HO}_T(k_*(n_t)) = \left\|\hat{s}_k^T - s\right\|^2 - \left\|\hat{s}_{k_*(n_t)}^T - s\right\|^2 - 2(P_n^{T^c} - P)(\hat{s}_k^T - \hat{s}_{k_*(n_t)}^T) \tag{5.1}$$

is the sum of a non-negative term (the excess risk) and a centered empirical process. Both tend to 0 as k tends to k_*. The relevant scale at which to study the hold-out procedure, which minimizes $\mathrm{HO}_T(k)$, is the scale of the fluctuations $\hat{k} - k_*(n_t)$ of the argmin $\hat{k}$ of $\mathrm{HO}_T(k)$. Since the asymptotic study of $\mathrm{HO}_T(\cdot)$ precedes that of $\hat{k}$, the correct scaling must be *guessed* in this chapter, and the correctness of this guess will follow from the results of the following chapter. The guess is made based on the following heuristic:

The correct scaling Δ for $|\hat{k} - k_*|$ is such that the excess risk $\left\|\hat{s}_{k_*\pm\Delta}^T - s\right\|^2 - \left\|\hat{s}_{k_*}^T - s\right\|^2$ and the centered empirical process $2(P_n^{T^c} - P)(\hat{s}_{k_*\pm\Delta}^T - \hat{s}_{k_*}^T)$ have the same order of magnitude.

The justification for this heuristic is that in order for the inequality $\mathrm{HO}_T(k) \leqslant \mathrm{HO}_T(k_*)$ to hold, as it does for $\hat{k}$, $2\left|(P_n^{T^c} - P)(\hat{s}_k^T - \hat{s}_{k_*}^T)\right|$ must be greater than $\left\|\hat{s}_k^T - s\right\|^2 - \left\|\hat{s}_{k_*(n_t)}^T - s\right\|^2$. Consider the following generic scaled and centered

hold-out process:

$$\begin{aligned}\frac{1}{\mathfrak{e}}\left(\mathrm{HO}_T\left(k_*+\alpha\Delta\right)-\mathrm{HO}_T\left(k_*(n_t)\right)\right) = \frac{1}{\mathfrak{e}}\left(\left\|\hat{s}^T_{k_*+\alpha\Delta}-s\right\|^2-\left\|\hat{s}^T_{k_*(n_t)}-s\right\|^2\right)\\ -\frac{2}{\mathfrak{e}}(P_n^{T^c}-P)(\hat{s}^T_{k_*+\alpha\Delta}-\hat{s}^T_{k_*(n_t)}),\end{aligned} \tag{5.2}$$

where $\alpha \in \{\frac{k-k_*}{\Delta} : k \in \mathbb{N}\}$ and $\Delta, \mathfrak{e}$ are values which may depend on s, n and n_t. The relative size of the excess risk and the empirical process in (5.2) depends on Δ. If the excess risk is much larger than the empirical process in equation (5.2), then the scaled process is asymptotically deterministic, conditionally on D_n^T (the excess risk depends only on D_n^T). In contrast, if the centered empirical process is dominant in equation (5.2), then the scaled hold-out is asymptotically a centered process. Thus, choosing Δ based on the heuristic means rescaling the hold-out process such that it is neither deterministic (conditionally on D_n^T) nor zero-mean, or in other words, so that its bias is of the same order of magnitude as its standard deviation. $\mathfrak{e}$ should then be chosen so that bias and standard deviation both remain of order 1, in order to avoid divergence of the scaled process or convergence to 0 (which would be uninformative). The appropriate choice of $\Delta, \mathfrak{e}$ is given in the following Definition.

Definition 5.3.3 *For all $n \in \mathbb{N}$, let*

$$\begin{aligned}\Delta_d(s,n_t,n) &= \max\left\{l \in \mathbb{N} : \theta^2_{k_*(n_t)+l} \geqslant \left[1-\sqrt{\frac{n_t}{n-n_t}}\frac{1}{\sqrt{l}}\right]\frac{1}{n_t}\right\}\\ \Delta_g(s,n_t,n) &= \min\left\{l \in \{0,\dots,k_*(n_t)\} : \theta^2_{k_*(n_t)-l} \geqslant \left[1+\sqrt{\frac{n_t}{n-n_t}}\frac{1}{\sqrt{l}}\right]\frac{1}{n_t}\right\}\\ \Delta(s,n_t,n) &= \max\left(\Delta_d(s,n_t,n), \Delta_g(s,n_t,n)\right)\\ \mathcal{E}(s,n_t,n) &= \frac{\Delta(s,n_t,n)}{n_t}.\\ \mathfrak{e}(s,n_t,n) &= \sqrt{\frac{\mathcal{E}(s,n_t,n)}{n-n_t}}.\end{aligned}$$

Definition 5.3.3 also introduces the quantity $\mathcal{E}(s,n_t,n)$. This quantity appears often in the proofs, so it is helpful to have notation for it; it will play a much more significant role in the study of Aggregated hold-out, in the following chapter. It can be interpreted as the order of magnitude of the fluctuations in the variance term,

$$\mathbb{E}\left[\left\|\hat{s}^T_k - s\right\|^2\right] - \mathbb{E}\left[\left\|\hat{s}^T_{k_*} - s\right\|^2\right],$$

and the bias term,

$$\left\|P(\hat{s}_k^T) - s\right\|^2 - \left\|P(\hat{s}_{k_*}^T) - s\right\|^2 = -\text{sign}(k - k_*) \sum_{j=k_* \wedge k+1}^{k \vee k_*} \theta_j^2,$$

of the estimators $\hat{s}_k^T$, for $k - k_*$ "of order" Δ (in a sense to be made precise later).

As the sequence $n_t(n)$ and the density s are considered to be fixed once and for all, the notation $\Delta(s, n_t, n), \mathcal{E}(s, n_t, n), \mathfrak{e}(s, n_t, n)$ will frequently be replaced by the abbreviations $\Delta, \mathcal{E}, \mathfrak{e}$.

Definition 5.3.3 does not make clear how large $\Delta, \mathcal{E}$ and $\mathfrak{e}$ are. Their order of magnitude may depend on the sequence $(\theta_j)_{j\in\mathbb{N}}$ of Fourier coefficients of s as well as on $n_t(n)$. However, the following inequalities always hold.

Lemma 5.3.4 *For any density s such that the sequence $\theta_j^2 = \langle s, \psi_j\rangle^2$ is non-increasing,*

$$\Delta \geqslant \frac{n_t}{n - n_t} \tag{5.3}$$

$$\mathcal{E} \geqslant \frac{1}{n - n_t} \tag{5.4}$$

$$\mathfrak{e} \geqslant \frac{1}{n - n_t} \tag{5.5}$$

$$\mathfrak{e} \leqslant \mathcal{E} \tag{5.6}$$

$$\mathcal{E} \leqslant 2\text{or}(n_t) + \frac{1}{n - n_t}. \tag{5.7}$$

This lemma is proved in section 5.4.1. The following two examples show that in extreme cases, lemma 5.3.4 may be optimal, at least up to constants.

Two examples

- Let $n_t(n)$ and u_n be two integer sequences, such that $\frac{n_t(n)}{n} \to 1$, $u_n \to +\infty$ and $u_n \leqslant \frac{\sqrt{n}}{2}$ for all n. Assume also that $\frac{n-n_t}{n_t} = o(\frac{u_{n_t}}{n_t})$.. Let for all $j \in \mathbb{N}$

$$\theta_{j,n}^2 = \begin{cases} 1 \text{ if } j = 0 \\ \frac{1}{n_t} \text{ if } 1 \leqslant j \leqslant u_{n_t} \\ \frac{1}{2^j n_t} \text{ if } j \geqslant u_{n_t} + 1, \end{cases} \tag{5.8}$$

  corresponding for example to the pdf $s_n = 1 + \sum_{j=1}^{u_{n_t}} \sqrt{\frac{1}{n_t}}\psi_j$. Remark that equation (5.8) implies that $k_*(n_t) = u_{n_t}$. Then as $n \to +\infty$, $\mathcal{E}(s_n, n_t, n) \sim \frac{k_*(n_t)}{n_t} \sim \text{or}(n_t)$ and $\frac{n-n_t}{n_t} = o(\text{or}(n_t))$, so $\mathfrak{e}(s_n, n_t, n) = o(\mathcal{E}(s_n, n_t, n))$.

- For all $j \in \mathbb{N}$, let $\theta_j = \frac{1}{3^j}$. Let $n_t(n)$ be a sequence of integers such that $\frac{n_t(n)}{n} \to 1$. Then by Lemma 6.6.2, $\Delta \geqslant \frac{n_t}{n-n_t}$, but as $9^{-\frac{n_t}{n-n_t}} = o\left(1 - \sqrt{\frac{1}{1+\frac{n-n_t}{n_t}}}\right)$, it follows that $\Delta(s, n_t, n) \sim \frac{n_t}{n-n_t}$, hence

$$\mathcal{E}(s, n_t, n) \sim \frac{n_t}{(n-n_t)n_t} \sim \frac{1}{n-n_t}.$$

  As a result, $\mathcal{E}(s, n_t, n) \sim \mathfrak{e}(s, n_t, n)$, and this for *any* sequence $n_t(n)$ such that $n_t(n) \sim n$.

Now that $\Delta, \mathfrak{e}$ are defined, the hold-out process can be rescaled as in equation (5.3.5). More precisely, the rescaled hold-out process is given by Definition 5.3.5 below.

Definition 5.3.5 *For all $j \in [-k_*; +\infty[\cap\mathbb{Z}$, let*

$$\hat{R}^{ho}\left(\frac{j}{\Delta}\right) = \frac{1}{\mathfrak{e}}\left(HO_T\left(k_* + j\right) - HO_T\left(k_*\right)\right),$$

in other words (by definition 5.3.1)

$$\hat{R}^{ho}\left(\frac{j}{\Delta}\right) = \frac{1}{\mathfrak{e}}\left(\left\|\hat{s}^T_{k_*+j} - s\right\|^2 - \left\|\hat{s}^T_{k_*} - s\right\|^2\right) - \frac{2}{\mathfrak{e}}\left(P_n^{T^c} - P\right)\left(\hat{s}^T_{k_*+j} - \hat{s}^T_{k_*}\right).$$

The $\hat{R}^{ho}$ function is extended by linear interpolation to all $\alpha \in \left[\frac{-k_*(n_t)}{\Delta}; +\infty\right[$.

The extension of $\hat{R}^{ho}$ by linear interpolation simplifies its approximation by a continuous process. Notice that any minimizer of $\hat{R}^{ho}$ on the grid $\frac{1}{\Delta}\left([-k_*(n_t); +\infty[\cap\mathbb{Z}\right)$ remains a minimizer of $\hat{R}^{ho}$ on the interval $\left[\frac{-k_*(n_t)}{\Delta}; +\infty\right[$. In particular, this applies to the hold-out parameter obtained by minimisation of the hold-out risk estimator.

The process $\hat{R}^{ho}$ can be expressed as the sum of the standardized excess risk $\frac{1}{\mathfrak{e}}\left(\left\|\hat{s}^T_{k_*+j} - s\right\|^2 - \left\|\hat{s}^T_{k_*} - s\right\|^2\right)$ and a centered empirical process: $\frac{2}{\mathfrak{e}}\left(P_n^{T^c} - P\right)\left(\hat{s}^T_{k_*+j} - \hat{s}^T_{k_*}\right)$. Though the excess risk is a priori random (it depends on D_n^T), the proof will show that it concentrates around a deterministic function f_n, depending on n, which is given by definition 5.3.6 below.

Definition 5.3.6 *For all $k \in \mathbb{N}$, let $R(k) = \sum_{j=k+1}^{+\infty} \theta_j^2$. Extend R to $\mathbb{R}_+$ by linear interpolation:*

$$\forall x \in \mathbb{R}_+, R(x) = (1 + \lfloor x \rfloor - x)R(\lfloor x \rfloor) + (x - \lfloor x \rfloor)R(\lfloor x \rfloor + 1).$$

$f_n :]-\frac{k_*(n_t)}{\Delta}; +\infty[\to \mathbb{R}_+$ *is now defined by:*

$$f_n(\alpha) = \frac{1}{\mathfrak{e}} \left(R(k_*(n_t) + \alpha\Delta) - R(k_*(n_t)) + \frac{\alpha\Delta}{n_t} \right). \tag{5.9}$$

Thus, for all $k \in \mathbb{N}$, $k \neq k_*(n_t)$,

$$\mathfrak{e} f_n \left(\frac{k - k_*(n_t)}{\Delta} \right) = \sum_{j=k\wedge k_*(n_t)+1}^{k\vee k_*(n_t)} \left| \theta_j^2 - \frac{1}{n_t} \right|. \tag{5.10}$$

It is clear by equation (5.10) that f_n reaches its minimum at 0, moreover the assumption that the sequence $(\theta_j^2)_{j\in\mathbb{N}}$ is non-increasing implies that f_n is convex. In particular, f_n is non-increasing on $]-\frac{k_*(n_t)}{\Delta}; 0]$ and non-decreasing on $[0; +\infty[$. Moreover, the definition of Δ and $\mathfrak{e}$ (Definition 5.3.3) implies the following bounds on the increments of f_n:

Lemma 5.3.7 *For any* $\alpha_1, \alpha_2 \in \mathbb{R}$ *such that* $\alpha_1\alpha_2 \geqslant 0$ *and* $|\alpha_2| \geqslant |\alpha_1| \geqslant 1$,

$$f_n(\alpha_2) - f_n(\alpha_1) \geqslant |\alpha_2| - |\alpha_1|.$$

In particular, since $f_n(0) = 0$, *for all* $\alpha \in \mathbb{R}$,

$$f_n(\alpha) \geqslant (|\alpha| - 1)_+.$$

Moreover, using the notation from Definition 5.3.3,

- *If* $\Delta = \Delta_d$, *then for any* $\alpha_1, \alpha_2 \in [0; 1]$ *such that* $\alpha_1 \leqslant \alpha_2$, $f_n(\alpha_2) - f_n(\alpha_1) \leqslant \alpha_2 - \alpha_1$.
- *If* $\Delta = \Delta_g$, *then for any* $\alpha_1, \alpha_2 \in [-1; 0]$ *such that* $\alpha_1 \leqslant \alpha_2$, $f_n(\alpha_1) - f_n(\alpha_2) \leqslant \alpha_2 - \alpha_1$.

This lemma is proved in section 5.4.1. It guarantees that f_n remains in a sense of "finite order" and "non zero" as $n \to +\infty$, which means that f_n remains uniformly bounded on $[-1; 0]$ or on $[0; 1]$, and is lower-bounded on $\mathbb{R}$ by the non-zero function $(|x| - 1)_+$.

The following theorem shows that the process $\hat{R}^{ho}(\cdot)$ can be approximated in a neighbourhood of 0 by the sum of f_n and a time-changed Brownian motion.

Theorem 5.3.8 *Assume that the pdf* s *is such that the sequence* $(\theta_j^2)_{j\in\mathbb{N}}$ *is non-increasing, where* $\theta_j = \langle s, \psi_j \rangle$ *are the Fourier coefficients of* s. *Assume that the sequence* $(\theta_j^2)_{j\in\mathbb{N}}$ *and the numbers* n_t, n *satisfy the following conditions.*

H1. *There exists constants* $c_1 \geqslant 0$ *and* $\delta_1 \geqslant 0$ *such that for all* $k \in \mathbb{N}$, $\sum_{j=k+1}^{+\infty} \theta_j^2 \leqslant \frac{c_1}{k^{2+\delta_1}}$.

H2. *There exists constants* $c_2 \geqslant 0$, $\rho_1 \geqslant 0$ *such that for all* $k \in \mathbb{N}$, $\sum_{j=k+1}^{+\infty} \theta_j^2 \geqslant \frac{c_2}{k^{\rho_1}}$

H3. *There exists constants* $c_3 > 0$, $\delta_2 > 0$ *such that for all* $k \geqslant 1$,

$$\theta_{k+k^{\delta_2}}^2 \geqslant c_3 \theta_{k-k^{\delta_2}}^2.$$

H4. *There exists a constant* $\delta_3 > 0$ *such that* $n - n_t \leqslant n^{1-\delta_3}$.

H5. *There exists a constant* $\delta_4 > 0$ *such that* $n_v = n - n_t \geqslant n^{\frac{2}{3}+\delta_4}$.

Let $T \subset \{1 \dots n\}$ *be a subset of cardinality* n_t *and let* $k_* = k_*(n_t)$. *For all* $x > 0$, *let*

$$a_x = \min\{\tfrac{j}{\Delta} : j \in \{-k_*(n_t), \dots, 0\}, f_n(\tfrac{j}{\Delta}) \leqslant x\}$$
$$b_x = \max\{\tfrac{j}{\Delta} : j \in \mathbb{N}, f_n(\tfrac{j}{\Delta}) \leqslant x\}.$$

Then, there exists a non-decreasing function $g_n : [-\frac{k_*(n_t)}{\Delta}; +\infty[\to \mathbb{R}$ *and for any* $x > 0$, *there exists a two-sided Brownian motion* $(W_t)_{t\in[a_x;b_x]}$ *independent from* D_n^T *such that, with probability greater than* $1 - \frac{3}{n^2}$,

$$\mathbb{E}\left[\sup_{u\in[a_x;b_x]} \left|\hat{R}^{ho}(u) - (f_n(u) - W_{g_n(u)})\right| \,\middle|\, D_n^T\right] \leqslant \kappa_0 (1+x)^{\frac{3}{2}} n^{-u_1}, \tag{5.11}$$

where $u_1 > 0$ *and* $\kappa_0 \geqslant 0$ *are two constants which depend only on* $\delta_1, \delta_2, \delta_4, \rho_1$ *and* c_1, c_2, δ_1, c_3, *respectively. Moreover,* g_n *and* W *can be chosen so as to satisfy the following conditions.*

1. $g_n(0) = 0$, $W_0 = 0$,
2. g_n *increases on its domain* $\left[\frac{-k_*}{\Delta}; +\infty\right[$
3. $\sup_{\alpha\in[a_x;b_x]} |g_n(\alpha)| \leqslant 20 \|s\|_\infty (1+x)$
4. $\forall(\alpha_1, \alpha_2) \in \left[\frac{-k_*}{\Delta}; +\infty\right[^2$, $|g_n(\alpha_1) - g_n(\alpha_2)| \geqslant 4\|s\|^2 |\alpha_1 - \alpha_2|$.

5. *For all* $(\alpha_1, \alpha_2) \in \left[\frac{-k_*}{\Delta}; +\infty\right[^2$ *such that* $\alpha_1 < \alpha_2 < 0$ *or* $0 < \alpha_1 < \alpha_2$,

$$g_n(\alpha_2) - g_n(\alpha_1) \leqslant -\frac{8\|s\|_\infty}{(n-n_t)\mathfrak{e}}[f_n(\alpha_2) - f_n(\alpha_1)] + \left(8\|s\|_\infty + 4\|s\|^2\right)[\alpha_2 - \alpha_1]. \tag{5.12}$$

In particular, as $(n-n_t)\mathfrak{e} \geqslant 1$ *by lemma 5.3.4 and* f_n *is non-decreasing on* $\mathbb{R}_+$,

- *If* $(\alpha_1, \alpha_2) \in \mathbb{R}_+^2$, $|g_n(\alpha_2) - g_n(\alpha_1)| \leqslant \left(8\|s\|_\infty + 4\|s\|^2\right)|\alpha_2 - \alpha_1|$
- *If* $(\alpha_1, \alpha_2) \in \left[\frac{-k_*}{\Delta}; 0\right]^2$,

$$|g_n(\alpha_2) - g_n(\alpha_1)| \leqslant 8\|s\|_\infty |f_n(\alpha_2) - f_n(\alpha_1)| + \left(8\|s\|_\infty + 4\|s\|^2\right)|\alpha_2 - \alpha_1|.$$

This theorem is proved in section 5.4. It shows that the renormalized hold-out process $\hat{R}^{ho}$ can be approximated by the sum of a convex function f_n and a Brownian motion changed in time W_{g_n}. f_n and g_n depend on n_t and n, but not on the data (they are deterministic functions), while W depends on the data only through the test sample $D_n^{T^c}$. In particular, in this asymptotic, $\hat{R}^{ho}$ doesn't depend on D_n^T, the training data.

Convergence occurs on an interval $[a_x; b_x]$, where $a_x \leqslant 0$, $b_x \geqslant 0$. Remark first that for all $x \geqslant 1$, $\max(-a_x, b_x) \geqslant 1$ since by lemma 5.3.7, we either have $0 \leqslant f_n \leqslant 1$ on $[0; 1]$ (when $\Delta = \Delta_d$) or $0 \leqslant f_n \leqslant 1$ on $[-1; 0]$ (when $\Delta = \Delta_g$). In particular, the length of the interval $[a_x; b_x]$ is lower-bounded by 1 for all $x \geqslant 1$, equation (5.11) is thus non-trivial. Figure 5.1 gives an illustration of the situation for $x = 25$ and

$$f_n : \alpha \mapsto \begin{cases} e^{-\alpha} - 1 \text{ if } \alpha \leqslant 0 \\ \frac{8}{10}\alpha + \frac{8}{30}\alpha^3 \text{ if } \alpha \geqslant 0 \end{cases}$$

$$g_n : \alpha \mapsto \begin{cases} 7.8\alpha \text{ if } \alpha \geqslant 0 \\ 7.8\alpha - 3f_n(\alpha) \text{ if } \alpha \leqslant 0 \end{cases}$$

(which satisfy the properties of lemma 5.3.7 and Theorem 5.3.8 when $\|s\|^2 \leqslant 1.2$ and $\|s\|_\infty \leqslant 1.5$).

Figure 5.1 suggests that for large x, $\sqrt{g_n}$ and hence W_{g_n} become negligible compared to f_n outside the interval $[a_x, b_x]$. This intuition can be theoretically justified: if $\alpha \in \frac{1}{\Delta}\mathbb{Z}$ does not belong to $[a_x; b_x]$, then $f_n(\alpha) \geqslant x$ by definition of a_x, b_x, while on the other hand, by equation (5.12), there exists a constant κ,

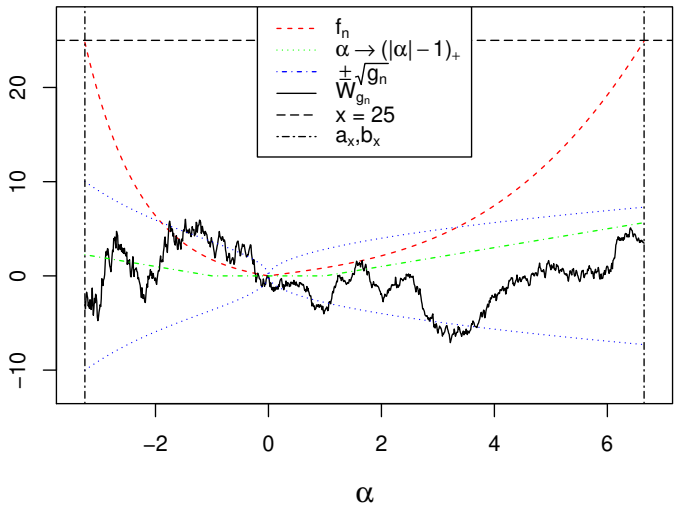

Figure 5.1: A plot of f_n, W_{g_n} on $[a_x; b_x]$, for $x = 25$, $g_n : \alpha \mapsto 7.8\alpha - 3f_n(\alpha)\mathbb{I}_{\alpha<0}$.

depending only on $\|s\|_\infty, \|s\|^2$, such that

$$\begin{aligned}\frac{\sqrt{\mathrm{Var}(W_{g_n(\alpha)})}}{f_n(\alpha)} &\leqslant \frac{\sqrt{g_n(\alpha)}}{f_n(\alpha)} \\ &\leqslant \frac{\sqrt{\kappa f_n(\alpha)}}{f_n(\alpha)} + \frac{\sqrt{\kappa|\alpha|}}{f_n(\alpha)}.\end{aligned}$$

By lemma 5.3.7, $f_n(\alpha) \geqslant (|\alpha| - 1)_+$ therefore $|\alpha| \leqslant 2f_n(\alpha)$ whenever $f_n(\alpha) \geqslant 1$, i.e for $\alpha \notin [a_1; b_1]$. This yields:

$$\forall x \geqslant 1,\ \forall \alpha \in \frac{1}{\Delta}\mathbb{Z}\backslash[a_x; b_x]\,, \qquad \frac{\sqrt{\mathrm{Var}(W_{g_n(\alpha)})}}{f_n(\alpha)} \leqslant \sqrt{\frac{\kappa}{x}} + \sqrt{\frac{2\kappa}{x}} = \frac{\sqrt{\kappa}(1+\sqrt{2})}{\sqrt{x}}.$$

Hence, for sufficiently large x, the random term W_{g_n} becomes negligible relative to the deterministic f_n outside the interval $[a_x; b_x]$. The proof of Theorem 6.6.5 in the following chapter shows that a relation similar to $\mathrm{Var}(W_{g_n}) \leqslant \kappa f_n$ – a *margin condition* – holds for the hold-out process $\hat{R}^{ho}$. Thus, $\hat{R}^{ho}$ becomes equivalent to the rescaled excess risk, hence to f_n (by the results of Section 5.4.2), when f_n is large enough. As a result, Theorem 6.6.5 of Chapter 6.6.5 shows that the minimizer of $\hat{R}^{ho}$ belongs to an interval $[a_{x_n}; b_{x_n}]$ with high probability, where x_n

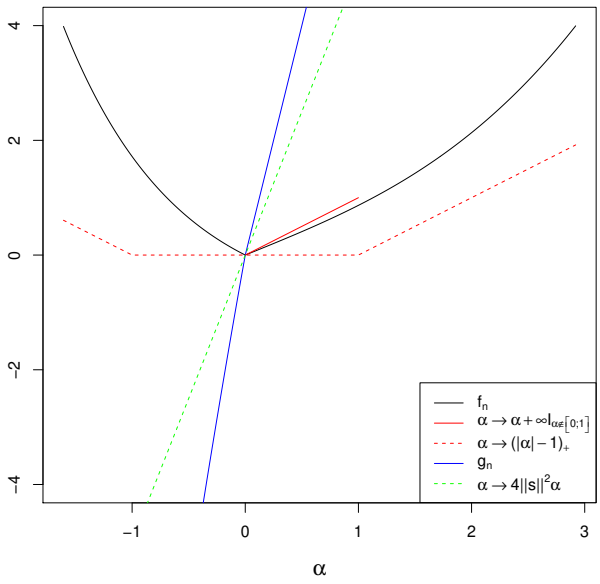

Figure 5.2: A plot of f_n, g_n on $[a_6; b_6]$ with upper and lower bounds, for $\|s\|^2 = 1.2$.

is of order $\log^2 n$. This justifies restricting the study of $\hat{R}^{ho}$ to intervals $[a_x; b_x]$ for $x > 0$. Note that Theorem 5.3.8 is non-asymptotic, so it can be applied with a sequence $x_n \to +\infty$.

Since $f_n(\alpha) \geqslant (|\alpha| - 1)_+$, the interval $[a_x; b_x]$ has length $\mathcal{O}(x)$, which is equivalent to an interval of length $\approx \Delta$ for the original parameter $k = k_* + \alpha\Delta$ of the hold-out risk estimator $\mathrm{HO}_T(k)$.

Furthermore, as illustrated on Figure 5.2, f_n is uniformly upper-bounded either on $[-1; 0]$ or on $[0; 1]$, while by equation (4), $|g_n|$ is lower-bounded on $[-1; -0.2] \cup [0, 2; 1]$ by a strictly positive constant independent from n.

It follows that the random processe $f_n - W_{g_n}$ remains random and bounded as $n \to +\infty$, at least on an interval of fixed length. This justifies the scaling used in the definition of $\hat{R}^{ho}$.

Theorem 5.3.8 is valid under some assumptions on the Fourier coefficients θ_j of s. The main one is that the sequence θ_j^2 should be non-increasing: this hypothesis is equivalent to the convexity of f_n, which approximates the excess risk. It guarantees that the set of "almost optimal" parameters k (i.e the level sets of the risk function) are intervals. This avoids situations where the hold-out "jumps" between two widely separated regions. However, it seems likely that these desir-

able effects of the hypothesis can be retained under weaker conditions. As the process $\hat{R}^{ho}(\alpha)$ consists of sums $\sum_{j=1}^{k_*+\alpha\Delta}$, it is probably sufficient to replace the hypotheses on the individual coefficients θ_j^2 with hypotheses bearing on local averages $\bar{\theta}_j^2 = \frac{1}{2m_n}\sum_{r=j-m_n}^{j+m_n}\theta_r^2$, at some scale $m_n << \Delta$. Depending on the scale m_n, the hypothesis that a smoothed sequence $\bar{\theta}_j^2$ is non-decreasing may be quite plausible, considering the fact that the Fourier coefficients tend to 0 at a prescribed rate for sufficiently smooth functions s.

The other assumptions on s (Hypotheses (H1), (H2), (H3)) basically mean that the Fourier coefficients of s on the cosine basis decrease polynomially. For example, they are satisfied if there are two strictly positive constants μ, L and a constant $\beta > 1$, such that

$$\forall k \in \mathbb{N}, \mu k^{-2\beta} \leqslant \sum_{j=k+1}^{+\infty} \theta_j^2 \leqslant L k^{-2\beta}.$$

More precisely, hypotheses H2 and H1 require that the remainder of the Fourier series lies between two polynomial sequences $\frac{c_2}{k^{\rho_1}}$ and $\frac{c_1}{k^{2+\delta_1}}$. The upper bound corresponds to a smoothness assumption on s of the type $s \in H^\beta$, $\beta > 1$, where H^β is a Sobolev space. It implies in particular that $(k_*(n_t) \leqslant n_t)$ is satisfied for all sufficiently large n_t (thus also for all large enough n). The lower bound arises from the fact that if the θ_j^2 decrease too fast, $k_*(n_t)$ grows to infinity at a very slow rate, and can even remain bounded if s is a trigonometric polynomial, which implies that the Brownian process is not a satisfactory approximation to the empirical process $(P_n^{T^c} - P)(\hat{s}_k^T - \hat{s}_{k_*}^T)$ for $k = \mathcal{O}(k_*)$.

Hypothesis H3 means that the sequence θ_j^2 cannot decrease too abruptly, excluding in particular a locally exponential decrease such as $\theta_{k_n+j}^2 = 2^{-j}w_n$, for $j \in \{1, \ldots, \varepsilon \log k_n\}$ and $k_n \to +\infty$. It is satisfied by polynomially decreasing sequences, $\theta_j^2 = \kappa j^{-\beta}$, with $\delta_2 = 1$, but also by sequences $\theta_j^2 = \kappa \exp(-j^\alpha)$, as long as $\alpha < 1$. Locally, θ_j^2 can thus decrease much faster than its global rate of convergence (which is polynomial).

Hypotheses H4 and H5 are of a different kind since they do not bear on s, but on the parameter n_t which is chosen by the statistician. One should therefore check that these assumptions are compatible with practical applications of the hold-out. The oracle inequalities of [5] show that the risk of the hold-out in model selection for L^2 density estimation is of order $\frac{n}{n_t}\mathrm{or}(n) + \frac{\log(n-n_t)}{n-n_t}$. If $\mathrm{or}(n_t)$ decreases in n_t with rate $\frac{1}{n_t^\alpha}$ $(\alpha < 1)$, which is the case under assumption H1, $n - n_t$ can be chosen within the interval $[\frac{1}{2}n^{\frac{1+\alpha}{2}}; \frac{n}{2}]$ —so that assumptions H4 and H5 are satisfied— without changing the order of magnitude of the risk.

5.4 Proofs

In this section, the term constant means a function of $\|s\|_\infty, \|s\|^2$ and the constants $c_1, c_2, \delta_1, \delta_2, \delta_3, \delta_4, \rho_1$ which appear in the hypotheses of Theorem 5.3.8. Note that by hypothesis (H1) $\|\theta\|_{\ell^1}, \|s\|_\infty, \|s\|^2$ are finite and can be bounded by functions of c_1, δ_1. The letter u will denote strictly positive constants that only depend on $\delta_1, \delta_2, \delta_3, \delta_4, \rho_1$ (they will generally appear as exponents of $\frac{1}{n}$). The letter κ denotes a non-negative constant. The notation $n_v = n - n_t$ will also be used frequently.

5.4.1 Preliminary results

The results of this section are independent from the rest. They will be used in the rest of the proof of Theorem 5.3.8, as well as in the Appendix. Let's start by proving some basic properties of a_x, b_x and f_n that will be used repeatedly in the main proofs.

Proof of lemma 5.3.4

- By definition and non-negativity of θ_j^2, $\sqrt{\frac{n_t}{n-n_t}}\frac{1}{\sqrt{\Delta_d}} \leqslant 1$, therefore $\Delta \geqslant \Delta_d \geqslant \frac{n_t}{n-n_t}$.
- $\mathcal{E} = \frac{\Delta}{n_t} \geqslant \frac{1}{n-n_t}$.
- $\mathfrak{e} = \sqrt{\frac{\mathcal{E}}{n-n_t}} \geqslant \sqrt{\frac{1}{(n-n_t)^2}} = \frac{1}{n-n_t}$.
- $\frac{\mathcal{E}}{\mathfrak{e}} = \mathcal{E}\sqrt{\frac{n-n_t}{\mathcal{E}}} = \sqrt{(n-n_t)\mathcal{E}} \geqslant 1$.
- By definition, $\Delta_g \leqslant k_*$. Thus $\frac{\Delta_g}{n_t} \leqslant \frac{k_*}{n_t} \leqslant \mathrm{or}(n_t)$. Moreover,

$$\Delta_d\left[1 - \sqrt{\frac{n_t}{n-n_t}}\frac{1}{\sqrt{\Delta_d}}\right]\frac{1}{n_t} \leqslant \sum_{j=k_*+1}^{k_*+\Delta_d}\theta_j^2 \leqslant \sum_{j=k_*+1}^{+\infty}\theta_j^2 \leqslant \mathrm{or}(n_t).$$

  Thus

$$\begin{aligned} n_t\mathrm{or}(n_t) &\geqslant \Delta_d - \sqrt{\frac{n_t}{n-n_t}}\sqrt{\Delta_d} \\ &\geqslant \Delta_d - \frac{1}{2}\frac{n_t}{n-n_t} - \frac{1}{2}\Delta_d \\ &\geqslant \frac{1}{2}\Delta_d - \frac{1}{2}\frac{n_t}{n-n_t}. \end{aligned}$$

It follows that

$$\Delta_d \leqslant 2n_t \mathrm{or}(n_t) + \frac{n_t}{n - n_t},$$

so since $\frac{n_t}{n-n_t}\frac{1}{n_t} = \frac{1}{n-n_t}$,

$$\frac{\Delta_d}{n_t} \leqslant 2\mathrm{or}(n_t) + \frac{1}{n - n_t},$$

which proves the result.

Proof of lemma 5.3.7

f_n is continuous and piecewise linear by definition 5.3.6. f_n is convex because the sequence θ_j^2 is non-increasing by assumption. Let $j \in \mathbb{Z}$ and $\alpha \in \left]\frac{j}{\Delta}; \frac{j+1}{\Delta}\right[$ be two numbers. By definition, f_n is linear on the interval $\left]\frac{j}{\Delta}; \frac{j+1}{\Delta}\right[$, in particular f_n is differentiable on this interval and

$$\begin{aligned} f_n'(\alpha) &= \Delta\left[f_n\left(\tfrac{j+1}{\Delta}\right) - f_n\left(\tfrac{j}{\Delta}\right)\right] \\ &= \frac{\Delta}{\mathfrak{e}}\left[\frac{1}{n_t} - \theta^2_{k_*+j+1}\right]. \end{aligned} \tag{5.13}$$

Because the sequence θ_j^2 is non-increasing, it follows from the definition of $k_*(n_t)$ that f_n is increasing on $\left]\frac{j}{\Delta}; \frac{j+1}{\Delta}\right[$ if $j \geqslant 0$ and non-decreasing if $j < 0$. This implies that f_n reaches its minimum at $k_*(n_t)$. If $\alpha \geqslant 1$, then $j = \lfloor \alpha\Delta \rfloor \geqslant \Delta$, therefore by definition of $\Delta_d \leqslant \Delta$,

$$\begin{aligned} f_n'(\alpha) &\geqslant \frac{\Delta}{\mathfrak{e}}\left[\frac{1}{n_t} - \theta^2_{k_*+\Delta+1}\right] \\ &\geqslant \frac{\Delta}{\mathfrak{e}}\sqrt{\frac{n_t}{n-n_t}}\frac{1}{\sqrt{\Delta}}\frac{1}{n_t} \\ &= \Delta\sqrt{\frac{(n-n_t)n_t}{\Delta}}\sqrt{\frac{n_t}{n-n_t}}\frac{1}{\sqrt{\Delta}}\frac{1}{n_t} \\ &= 1. \end{aligned}$$

In the same way, if $\alpha < -1$, then $j + 1 = \lceil \alpha\Delta \rceil \leqslant -\Delta \leqslant -\Delta_g$, so

$$\begin{aligned} f_n'(\alpha) &\leqslant \frac{\Delta}{\mathfrak{e}}\left[\frac{1}{n_t} - \theta^2_{k_*-\Delta}\right] \\ &\leqslant -\frac{\Delta}{\mathfrak{e}}\sqrt{\frac{n_t}{n-n_t}}\frac{1}{\sqrt{\Delta}}\frac{1}{n_t} \\ &\leqslant -1. \end{aligned}$$

Furthermore,

- If $\Delta = \Delta_d$, then for all $\alpha \in [0;1]$, $j+1 = \lceil \alpha\Delta \rceil \leqslant \Delta = \Delta_d$, therefore by definition of Δ_d,

$$\begin{aligned} f_n'(\alpha) &\leqslant \frac{\Delta}{\mathfrak{e}} \left[\frac{1}{n_t} - \theta^2_{k_*+\Delta} \right] \\ &\leqslant \frac{\Delta}{\mathfrak{e}} \sqrt{\frac{n_t}{n-n_t}} \frac{1}{\sqrt{\Delta}} \frac{1}{n_t} \\ &\leqslant 1. \end{aligned}$$

- If $\Delta = \Delta_g$, then for all $\alpha \in [-1;0]$, $j = \lfloor \alpha\Delta \rfloor \geqslant -\Delta = -\Delta_g$, therefore by definition of Δ_g and since the sequence $(\theta^2_j)_{j\in\mathbb{N}}$ is non-increasing,

$$\begin{aligned} f_n'(\alpha) &\geqslant \frac{\Delta}{\mathfrak{e}} \left[\frac{1}{n_t} - \theta^2_{k_*-\Delta+1} \right] \\ &\geqslant -\frac{\Delta}{\mathfrak{e}} \sqrt{\frac{n_t}{n-n_t}} \frac{1}{\sqrt{\Delta}} \frac{1}{n_t} \\ &\geqslant -1. \end{aligned}$$

By continuity of f_n, this proves the lemma.

Properties of the interval $[a_x; b_x]$

Lemma 5.4.1 *Let a_x, b_x be as defined in Theorem5.3.8. Then for all $x > 0$,*

$$\begin{aligned} &[b_x - a_x] \leqslant 2(1+x) \\ &\sum_{k_*+a_x\Delta}^{k_*+b_x\Delta} \theta_j^2 \leqslant 4(1+x)\mathcal{E}. \end{aligned}$$

Proof Either $b_x \leqslant 1$, or $b_x > 1$ and by lemma 5.3.7, $f_n(b_x) \geqslant b_x - 1$ which implies that $b_x \leqslant f_n(b_x)+1 \leqslant x+1$. In all cases, $b_x \leqslant x+1$. In the same way, $a_x > -1-x$. Thus $b_x - a_x \leqslant 2(1+x)$. Moreover,

$$\begin{aligned} \sum_{k_*+a_x\Delta}^{k_*+b_x\Delta} \theta_j^2 &\leqslant \sum_{k_*+a_x\Delta}^{k_*+b_x\Delta} \left| \theta_j^2 - \frac{1}{n_t} \right| + \frac{(b_x - a_x)\Delta}{n_t} \\ &\leqslant \mathfrak{e}[f_n(a_x) + f_n(b_x)] + [b_x - a_x]\mathcal{E} \\ &\leqslant 2x\mathfrak{e} + 2\left[1+x\right]\mathcal{E}. \end{aligned}$$

Since $\mathfrak{e} \leqslant \mathcal{E}$ by lemma 5.3.4,

$$\sum_{k_*+a_x\Delta}^{k_*+b_x\Delta} \theta_j^2 \leqslant (4x+2)\mathcal{E}.$$

This proves lemma 5.4.1. ■

We now introduce some notation which will be used in the remainder of this chapter.

Definition 5.4.2 *Let an i.i.d sample D_n be given, with distribution P and pdf s on $[0;1]$. For all $j \in \mathbb{N}$ and any $T \subset \{1,\dots,n\}$, let*

$$\theta_j = P\psi_j = \langle s, \psi_j \rangle$$
$$\hat{\theta}_j^T = P_n^T(\psi_j).$$

This notation will be used very often in the remainder of the chapter.

The hold-out risk estimator can be expressed as the sum of two terms. Definition 5.4.3 below gives a name to each of these terms.

Definition 5.4.3 *For all $j \in [-k_*(n_t); +\infty[\cap\mathbb{Z}$, let*

$$L(\tfrac{j}{\Delta}) = \frac{1}{\mathfrak{e}}\left(\left\|\hat{s}_{k_*+j}^T - s\right\|^2 - \left\|\hat{s}_{k_*}^T - s\right\|^2\right).$$

The function L is extended to the interval $[-\frac{k_*(n_t)}{\Delta}; +\infty[$ by linear interpolation. Let Z be the random function defined for all $j \in [-\frac{k_*(n_t)}{\Delta}; +\infty[\cap\mathbb{Z}$ by

$$Z\left(\frac{j}{\Delta}\right) = \frac{2}{\mathfrak{e}}\left(P_n^{T^c} - P\right)\left(\hat{s}_{k_*+j}^T - \hat{s}_{k_*}^T\right)$$

and extended by linear interpolation to the interval $[-\frac{k_*(n_t)}{\Delta}; +\infty[$, so that for all α, $\hat{R}^{ho}(\alpha) = L(\alpha) - Z_\alpha$.

Thus, L is the rescaled excess risk, and Z is a centered empirical process. These two terms will be approximated separately.

5.4.2 Approximation of the excess risk

Let $x > 0$ be fixed for the entirety of this section. We now prove the following claim.

Claim 5.4.3.1 *Let L be the function introduced in definition 5.4.3, and f_n be given by definition 5.3.6. There exists a constant κ_1 such that, with probability greater than $1 - \frac{1}{n^2}$,*

$$\sup_{\alpha\in[a_x;b_x]} |L(\alpha) - f_n(\alpha)| \leqslant \kappa_1(1+x)[\log(2+x)^2 + \log^2 n]n^{-\min(\frac{1}{12},\frac{\delta_3}{2})}.$$

Proof Let $j \in \{a_x\Delta, \dots, b_x\Delta\}$. Since $\hat{s}_k^T = \sum_{j=1}^k P_n^T(\psi_j)\psi_j = \sum_{j=1}^k \hat{\theta}_j^T \psi_j$,

$$\left\|\hat{s}_{k_*+j}^T - s\right\|^2 - \left\|\hat{s}_{k_*}^T - s\right\|^2 = \operatorname{sgn}(j) \sum_{i=k_*+(j)_-+1}^{k_*+(j)_+} \left(\hat{\theta}_i^T - \theta_i\right)^2 - \theta_i^2.$$

It is known [5, Lemma 14] [70, Proposition 6.3] that the process

$$\sum_{i=k_*+(j)_-+1}^{k_*+(j)_+} \left(\hat{\theta}_i^T - \theta_i\right)^2 = \sum_{i=k_*+(j)_-+1}^{k_*+(j)_+} (P^T - P)(\psi_j)^2$$

concentrates around its expectation, so that

$$\sum_{i=k_*+(j)_-+1}^{k_*+(j)_+} \left(\hat{\theta}_i^T - \theta_i\right)^2 \sim \sum_{i=k_*+(j)_-+1}^{k_*+(j)_+} \frac{\operatorname{Var}(\psi_j)}{n_t}.$$

Furthermore, by lemma 5.5.1 in the appendix, $\operatorname{Var}(\psi_j) \sim 1$, therefore

$$\sum_{i=k_*+(j)_-+1}^{k_*+(j)_+} \left(\hat{\theta}_i^T - \theta_i\right)^2 \sim \frac{|j|}{n_t}.$$

More precisely, proposition 5.5.3 in the appendix and a union bound show that, with probability greater than $1 - \frac{1}{n^2}$, for any $j \in \mathbb{Z} \cap [a_x\Delta; b_x\Delta[$,

$$\left| \sum_{i=k_*+(j)_-+1}^{k_*+(j)_+} \left(\hat{\theta}_i^T - \theta_i\right)^2 - \frac{|j|}{n_t} \right| \leqslant \kappa_1 (3\log n + \log((b_x - a_x)\Delta))^2 n^{-\min(\frac{1}{12}, \frac{\delta_3}{2})} \frac{j}{\Delta} \mathfrak{e}.$$

Let $r_n = \kappa_1 (3\log n + \log((b_x - a_x)\Delta))^2 n^{-\min(\frac{1}{12}, \frac{\delta_3}{2})}$. Then for any $j \geqslant 1$,

$$\begin{aligned}
\left\|\hat{s}_{k_*+j}^T - s\right\|^2 - \left\|\hat{s}_{k_*}^T - s\right\|^2 &= -\sum_{i=k_*+1}^{k_*+j} \theta_i^2 + \sum_{i=k_*+1}^{k_*+j} \left(\hat{\theta}_i^T - \theta_i\right)^2 \\
&= -\sum_{i=k_*+1}^{k_*+j} \theta_i^2 + \frac{j}{n_t} \pm \frac{j}{\Delta} r_n \mathfrak{e} \\
&= \sum_{i=k_*+1}^{k_*+j} \left[\frac{1}{n_t} - \theta_i^2\right] \pm \frac{j}{\Delta} r_n \mathfrak{e} \\
&= \mathfrak{e} f_n\left(\frac{j}{\Delta}\right) \pm \frac{j}{\Delta} r_n \mathfrak{e}.
\end{aligned}$$

On this same event, for any $j \in \{-k_*(n_t), \dots, -1\}$,

$$\begin{aligned}
\left\|\hat{s}^T_{k_*+j} - s\right\|^2 - \left\|\hat{s}^T_{k_*} - s\right\|^2 &= \sum_{i=k_*+j+1}^{k_*} \theta_i^2 - \sum_{i=k_*+j+1}^{k_*} \left(\hat{\theta}^T_i - \theta_i\right)^2 \\
&= \sum_{i=k_*+j+1}^{k_*} \theta_i^2 - \frac{|j|}{n_t} \pm \frac{|j|}{\Delta} r_n \mathfrak{e} \\
&= \sum_{i=k_*+j+1}^{k_*} \left[\theta_j^2 - \frac{1}{n_t}\right] \pm \frac{j}{\Delta} r_n \mathfrak{e} \\
&= \mathfrak{e} f_n\left(\frac{j}{\Delta}\right) \pm \frac{j}{\Delta} r_n \mathfrak{e}.
\end{aligned}$$

Thus, since f_n and $\hat{R}^{ho}$ are linear between the points of $\frac{1}{\Delta}\mathbb{Z}$,

$$\begin{aligned}
\sup_{\alpha \in [a_x; b_x]} |L(\alpha) - f_n(\alpha)| &= \frac{1}{\mathfrak{e}} \max_{a_x \Delta \leqslant j \leqslant b_x \Delta} \left| \left\|\hat{s}^T_{k_*+j} - s\right\|^2 - \left\|\hat{s}^T_{k_*} - s\right\|^2 - \mathfrak{e} f_n\left(\tfrac{j}{\Delta}\right) \right| \\
&\leqslant \max(|a_x|, |b_x|) r_n.
\end{aligned}$$

By lemma 5.4.1, $\max(|a_x|, |b_x|) \leqslant b_x - a_x \leqslant 2(1+x)$ so

$$\begin{aligned}
\max(|a_x|, |b_x|) r_n &\leqslant 2(1+x)\kappa_1 (3 \log n + \log(2(1+x)) + \log \Delta)^2 n^{-\min(\frac{1}{12}, \frac{\delta_3}{2})} \\
&\leqslant \kappa (1+x)[\log(2+x)^2 + \log^2 n] n^{-\min(\frac{1}{12}, \frac{\delta_3}{2})}
\end{aligned}$$

for some constant κ, since by lemma 5.3.4 and hypothesis (H5) of Theorem 5.3.8,

$$\begin{aligned}
\Delta &= n_t \mathcal{E} \\
&\leqslant 2 n_t o r(n_t) + \frac{n_t}{n - n_t} \\
&\leqslant 2(\|s\|^2 - 1) n_t + n_t^{\frac{1}{3}}.
\end{aligned}$$

This proves claim 5.4.3.1. ■

We will now seek to approximate the process Z given by definition 5.4.3.

5.4.3 Strong approximation of the hold-out process

Let us start by showing that the empirical process Z (definition 5.4.3) can be approximated by a gaussian process, uniformly on $[a_x; b_x]$. This is the purpose of the following result, which will be proven in this section.

Claim 5.4.3.2 *Let Z be the process given by definition 5.4.3. There exists a gaussian process $(Z^1_\alpha)_{\alpha\in[a_x;b_x]}$ with the same variance-covariance function as Z: for any $(\alpha_1,\alpha_2)\in[a_x;b_x]^2$, $\mathrm{Cov}(Z^1_{\alpha_1},Z^1_{\alpha_2})=\mathrm{Cov}(Z_{\alpha_1},Z_{\alpha_2})$ and such that for all $n\geqslant 1$, for all $x>0$, with probability greater than $1-e^{-y}$,*

$$\mathbb{E}\left[\sup_{\alpha\in[a_x;b_x]}|Z_\alpha-Z^1_\alpha|\Big|D_n^T\right]\leqslant\kappa_5(c_1)(1+y)(1+x)^{\frac{3}{2}}n^{-\frac{\delta_4}{3}}.$$

Furthermore, Z^1 can be expressed as $Z^1=H(Z,\nu)$, with ν a uniform random variable independent from D_n and H a measurable function on $C([0;1],\mathbb{R})$.

Let $n_v=|T^c|=n-|T|=n-n_t$. Let $F:x\to\int_0^x s(t)dt$ be the cumulative distribution function of the given X_i. Let $F_{T^c}:x\to\frac{1}{n_v}\sum_{i\notin T}\mathbb{I}_{X_i\leqslant x}$ be the empirical cumulative distribution function of the sample $D_n^{T^c}$. By the Komlos-Major-Tusnady approximation theorem [62, Theorem 3], there exist a universal constant C and a standard Brownian bridge process B_{T^c} such that for all $y>0$, with probability greater than $1-e^{-y}$, $\left\|B_{T^c}\circ F-\sqrt{n_v}(F_{T^c}-F)\right\|_\infty\leqslant\frac{C(\log n_v+y)}{\sqrt{n_v}}$ (remark that since F is continuous, $F(X_i)\sim\mathcal{U}([0;1])$, which means that the result for general F follows from the result for the uniform distribution). Furthermore, B_{T^c} can always be realized as a measurable function of $D_n^{T^c}$ and an auxiliary, uniformly distributed random variable ν: $B_{T^c}=H(D_n^{T^c},\nu)$, with ν independant from D_n. Let B^{T^c} be obtained in this way. From $B_{T^c}\circ F$, one can define an operator on the Sobolev space $W^1(\mathbb{R})$:

Definition 5.4.4 *For any function f such that $f'\in L^1([0;1])$, let*

$$G_{T^c}(f)=-\int_0^1 f'(x)B_{T^c}(F(x))dx.$$

G_{T^c} "approximates" the empirical process $\sqrt{n_v}(P_n^{T^c}-P)$ on the space W^1. Lemma 5.4.5 below gives a bound on the error made with this approximation.

Lemma 5.4.5 *For any function f such that $f'\in L^1([0;1])$,*

$$\left|G_{T^c}(f)-\sqrt{n_v}(P_n^{T^c}-P)(f)\right|\leqslant\left\|B_{T^c}-\sqrt{n_v}(F_{T^c}-F)\right\|_\infty\left\|f'\right\|_{L^1}.$$

Furthermore, for all functions f,g such that $f',g'\in L^1([0;1])$,

$$\mathrm{Cov}(G_{T^c}(f),G_{T^c}(g))=P[fg]-P[f]P[g]=\mathrm{Cov}\left(\sqrt{n_v}(P_n^{T^c}-P)(f),\sqrt{n_v}(P_n^{T^c}-P)(g)\right).$$

Proof Let f be a function such that $f' \in L^1([0;1])$. Then

$$\begin{aligned}(P_n^{T^c} - P)(f) &= \int f d(P_n^{T^c} - P) \\ &= \int [f - f(0)] d(P_n^{T^c} - P) \\ &= \int_0^1 \int_0^1 \mathbb{I}_{t \leqslant x} f'(t) dt \; d(F_{T^c} - F)(x) \\ &= \int_0^1 f'(t)(P_n^{T^c} - P)(]t; +\infty)) \\ &= -\int_0^1 f'(t)(F_{T^c} - F)(t) dt. \end{aligned} \tag{5.14}$$

Il follows that for all functions f such that $f' \in L^1([0;1])$,

$$\begin{aligned}\left|G_{T^c}(f) - \sqrt{n_v}(P_n^{T^c} - P)(f)\right| &= \left|\int_0^1 f'(t)\left[\sqrt{n_v}(F_{T^c} - F) - B_{T^c} \circ F\right](t) dt\right| \\ &\leqslant \|f'\|_{L^1([0;1])} \left\|B_{T^c} \circ F - \sqrt{n_v}(F_{T^c} - F)\right\|_\infty .\end{aligned}$$

By definition, it is clear that $\mathbb{E}[G_{T^c}(f)] = 0$. Thus,

$$\begin{aligned}\mathrm{Cov}\left(G_{T^c}(f), G_{T^c}(g)\right) &= \mathbb{E}\left[G_{T^c}(f) G_{T^c}(g)\right] \\ &= \mathbb{E}\left[\int_0^1 \int_0^1 f'(u) g'(v) B_{T^c}(F(u)) B_{T^c}(F(v))\right] \\ &= \int_0^1 \int_0^1 f'(u) g'(v) [F(u) \wedge F(v)][1 - F(u) \vee F(v)] du dv \\ &= \int_0^1 \int_0^1 f'(u) g'(v) \left(\mathbb{E}[\mathbb{I}_{X \leqslant u} \mathbb{I}_{X \leqslant v}] - \mathbb{E}[\mathbb{I}_{X \leqslant u}] \mathbb{E}[\mathbb{I}_{X \leqslant v}]\right) \\ &= n_v \int_0^1 \int_0^1 f'(u) g'(v) \mathbb{E}\left[(F_{T^c} - F)(u)(F_{T^c} - F)(v)\right] \\ &= \mathrm{Cov}\left(\sqrt{n_v}(P_n^{T^c} - P)(f), \sqrt{n_v}(P_n^{T^c} - P)(g)\right) \text{ by equation (5.14)}\end{aligned}$$

■

Let the process Z^1 be defined for all $j \in \{a_x\Delta, \ldots, b_x\Delta\}$ by

$$Z^1\left(\frac{j}{\Delta}\right) = \frac{2}{\sqrt{n_v}\mathfrak{e}} G_{T^c}\left(\hat{s}^T_{k_*+j} - \hat{s}^T_{k_*}\right).$$

Z^1 is extended to the interval $[a_x; b_x]$ by linear interpolation, as for Z. By lemma 5.4.5, the variance-covariance function of Z^1 conïncides with that of Z at the

points $\frac{j}{\Delta}, j \in \mathbb{Z} \cap [a_x\Delta; b_x\Delta]$, and this property extends by bilinearity to the whole interval $[a_x; b_x]$. Furthermore,

$$\begin{aligned}
\sup_{a_x \leqslant \alpha \leqslant b_x} |Z^1_\alpha - Z_\alpha| &\leqslant \max_{j \in \in \mathbb{Z} \cap [a_x; b_x]} \left| Z^1\left(\frac{j}{\Delta}\right) - Z\left(\frac{j}{\Delta}\right)\right| \\
&\leqslant \frac{1}{\sqrt{n_v}} \left\| B_{T^c} \circ F - \sqrt{n_v}(F_{n_v} - F)\right\|_\infty \times \max_{a_x\Delta \leqslant j \leqslant b_x\Delta} \left\| \sum_{i=k_*+1}^{k_*+j} i\hat{\theta}^T_i \sin(i\cdot)\right\|_1 \\
&\leqslant \frac{1}{\sqrt{n_v}} \left\| B_{T^c} \circ F - \sqrt{n_v}(F_{n_v} - F)\right\|_\infty \times \sqrt{\sum_{i=k_*+a_x\Delta+1}^{k_*+b_x\Delta} i^2(\hat{\theta}^T_i)^2}
\end{aligned}$$

By construction, the process $B_{T^c} \circ F - \sqrt{n_v}(F_{n_v} - F)$ is independent from D^T_n. As a result,

$$\begin{aligned}
\mathbb{E}\left[\sup_{a_x \leqslant \alpha \leqslant b_x} |Z^1_\alpha - Z_\alpha| | D^T_n\right] &\leqslant \frac{2}{\mathfrak{e}\sqrt{n_v}} \mathbb{E}\left[\left\| B_{T^c} \circ F - \sqrt{n_v}(F_{n_v} - F)\right\|_\infty\right] \\
&\quad \times (k_* + b_x\Delta) \sqrt{\sum_{i=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \left(\hat{\theta}^T_j\right)^2} \\
&\leqslant \frac{2C \log n_v}{\mathfrak{e} n_v} \times (k_* + b_x\Delta) \sqrt{\sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} 2\theta_j^2 + 2\left(\hat{\theta}^T_j - \theta_j\right)^2}.
\end{aligned} \tag{5.15}$$

By proposition 5.5.3, there exists an event $E_1(y)$ of probability greater than $1 - e^{-y}$ such that, for all $D^T_n \in E_1(y)$,

$$\sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \left(\hat{\theta}^T_j - \theta_j\right)^2 \leqslant [b_x - a_x]\frac{\Delta}{n_t} + \kappa_1(b_x - a_x)[\log n + y]^2 n^{-\min(\frac{1}{12}, \frac{\delta_3}{2})} \mathfrak{e}(n),$$

therefore by lemma 5.4.1 and equation (5.15), for all $D^T_n \in E_1(y)$,

$$\begin{aligned}
\mathbb{E}\left[\sup_{a_x \leqslant \alpha \leqslant b_n} |Z^1_\alpha - Z_\alpha| \middle| D^T_n\right] &\leqslant (k_* + b_x\Delta) 2\sqrt{1+x} \frac{C \log n_v}{n_v} \\
&\quad \times \left[2\sqrt{\mathcal{E}} + \sqrt{2\kappa_1}(\log n + y) n^{-\min(\frac{1}{12}, \frac{\delta_3}{2})} \sqrt{\mathfrak{e}}\right].
\end{aligned}$$

Since $\mathfrak{e} \leqslant \mathcal{E}$ and $n^{-\min(\frac{1}{12}, \frac{\delta_3}{2})} \log n \to 0$, there exists therefore a constant κ such

that for all $D_n^T \in E_1(y)$:

$$\mathbb{E}\left[\sup_{a_x\leqslant\alpha\leqslant b_n}|Z_\alpha^1 - Z_\alpha|\Big|D_n^T\right] \leqslant \kappa\frac{\log n_v}{n_v}\times(k_* + b_x\Delta)(1+y)\sqrt{(1+x)\mathcal{E}}$$
$$\leqslant \kappa\frac{\log n_v}{\sqrt{n_v}}\times(k_* + b_x\Delta)(1+y)\sqrt{(1+x)}\mathfrak{e}. \qquad (5.16)$$

By lemma 5.3.4, $\mathcal{E}\leqslant 2\mathrm{or}(n_t)+\frac{1}{n_v}$ therefore $\Delta\leqslant 2n_t\mathrm{or}(n_t)+\frac{n_t}{n_v}$ and by definition of *or*, $k_*(n_t) = n_t\frac{k_*}{n_t}\leqslant n_t\mathrm{or}(n_t)$ therefore $\frac{k_*+b_x\Delta}{\sqrt{n_v}}\leqslant(2b_x+1)\frac{n_t\mathrm{or}(n_t)}{\sqrt{n_v}}+\frac{b_x n_t}{n_v\sqrt{n_v}}$. By hypothesis H5 of Theorem 5.3.8, $n_v\geqslant n^{\frac{2}{3}+\delta_4}$, so

$$\frac{k_*+b_x\Delta}{\sqrt{n_v}}\leqslant(2b_x+1)\frac{n_t\mathrm{or}(n_t)}{n^{\frac{1}{3}+\frac{\delta_4}{2}}}+\frac{b_x}{n^{\frac{3\delta_4}{2}}}.$$

Moreover, by hypothesis H1 of Theorem 5.3.8, $\sum_{j=k+1}^{+\infty}\theta_j^2\leqslant\frac{c_1}{k^{2+\delta_1}}$, therefore

$$\mathrm{or}(n_t)\leqslant\min_{k\in\mathbb{N}^*}\frac{c_1}{k^{2+\delta_1}}+\frac{k}{n_t}\leqslant 2\inf_{x\geqslant 1}\frac{c_1}{x^{2+\delta_1}}+\frac{x}{n_t}\leqslant 3\frac{c_1^{\frac{1}{3+\delta_1}}}{n_t^{\frac{2+\delta_1}{3+\delta_1}}},$$

whence (since $c_1\geqslant 1$) $n_t\mathrm{or}(n_t)\leqslant 3(c_1n_t)^{\frac{1}{3+\delta_1}}$. It follows that:

$$\log n\frac{k_*+b_x\Delta}{\sqrt{n_v}}\leqslant 3(2b_x+1)c_1^{\frac{1}{3+\delta_1}}\log n n^{-\frac{\delta_4}{2}}+\frac{b_x\log n}{n_t^{\frac{3\delta_4}{2}}}. \qquad (5.17)$$

Since $\frac{\log n}{n^{\frac{\delta_4}{2}}} = o\left(n^{-\frac{\delta_4}{3}}\right)$, by equations (5.16), (5.17) and lemma 5.4.1, there exists a constant $\kappa(c_1)$ such that for any n, with probability greater than $1-e^{-y}$,

$$\mathbb{E}\left[\sup_{a_x\leqslant\alpha\leqslant b_n}|Z_\alpha^1 - Z_\alpha|\Big|D_n^T\right]\leqslant\kappa(1+y)(1+x)^{\frac{3}{2}}n^{-\frac{\delta_4}{3}}.$$

5.4.4 Approximation of the covariance function

We will now seek to approximate the process Z^1 given by claim 5.4.3.2 by a time-changed Wiener process. To this end, we first approximate the variance-covariance function of Z^1 (which is the same as that of Z).

Claim 5.4.5.1 *There exists a function g_n satisfying the hypotheses of Theorem 5.3.8 and a constant $u_5>0$ such that, for all $x>0$, with probability greater than $1-\frac{1}{n^2}$,*

$$\max_{(j_1,j_2)\in\{0,\dots,b_x\Delta\}^2}\left|\mathrm{Cov}\left(Z\left(\tfrac{j_1}{\Delta}\right),Z\left(\tfrac{j_2}{\Delta}\right)|D_n^T\right)-\min\left(g_n\left(\tfrac{j_1}{\Delta}\right),g_n\left(\tfrac{j_2}{\Delta}\right)\right)\right|\leqslant\kappa_6(1+x)^2\log^2(n)n^{-u_5}$$
$$\max_{(j_1,j_2)\in\{a_x\Delta,\dots,0\}^2}\left|\mathrm{Cov}\left(Z\left(\tfrac{j_1}{\Delta}\right),Z\left(\tfrac{j_2}{\Delta}\right)|D_n^T\right)-\max\left(g_n\left(\tfrac{j_1}{\Delta}\right),g_n\left(\tfrac{j_2}{\Delta}\right)\right)\right|\leqslant\kappa_6(1+x)^2\log^2(n)n^{-u_5}$$
$$\max_{(j_1,j_2)\in\{a_x\Delta,\dots,0\}\times\{0,\dots,b_x\Delta\}}\left|\mathrm{Cov}\left(Z\left(\tfrac{j_1}{\Delta}\right),Z\left(\tfrac{j_2}{\Delta}\right)|D_n^T\right)\right|\leqslant\kappa_6(1+x)^2\log^2(n)n^{-u_5}.$$

We introduce the following definition.

Definition 5.4.6 *Let $(W_t)_{t\in\mathbb{R}}$ be a two-sided Wiener process such that $W_0 = 0$. For any function $g : I \to \mathbb{R}$, where I is an interval containing 0, let $K(g) : I^2 \to \mathbb{R}$ be defined for any $(s,t) \in I^2$ by*

$$K(g)(s,t) = \begin{cases} g(s \wedge t) \text{ if } (s,t) \in (I \cap \mathbb{R}_+)^2 \\ -g(s \vee t) \text{ if } (s,t) \in (I \cap \mathbb{R}_-)^2 \\ 0 \text{ else .} \end{cases} \tag{5.18}$$

For all $j \in \mathbb{Z} \cap [a_x\Delta; b_x\Delta]$, by definition 5.4.3 of Z:

$$Z\left(\frac{j}{\Delta}\right) = \frac{2}{\mathfrak{e}}\left(P_n^{T^c} - P\right)\left(\hat{s}^T_{k_*+j} - \hat{s}^T_{k_*}\right) = \begin{cases} 0 \text{ if } j = 0, \\ \frac{2}{\mathfrak{e}}\left(P_n^{T^c} - P\right)\sum_{i=k_*+1}^{k_*+j} \hat{\theta}_i^T \psi_i \text{ if } j > 0, \\ \frac{2}{\mathfrak{e}}\left(P_n^{T^c} - P\right)\sum_{i=k_*+j+1}^{k_*} \hat{\theta}_i^T \psi_i \text{ if } j < 0 \end{cases} \tag{5.19}$$

In other words, for all $j \in \mathbb{Z} \cap [a_x\Delta; b_x\Delta]$,

$$Z\left(\frac{j}{\Delta}\right) = \operatorname{sgn}(j)\frac{2}{\mathfrak{e}} \sum_{i=k_*-(j)_-+1}^{k_*+(j)_+} \hat{\theta}_i^T \left(P_n^{T^c} - P\right)(\psi_i).$$

Let $n_v = |T^c| = n - n_t$. Thus, for any $(j_1, j_2) \in \{a_x\Delta, \ldots, b_x\Delta\}^2$ and any variable X with distribution $s(x)dx$,

$$\operatorname{Cov}\left(Z\left(\tfrac{j_1}{\Delta}\right), Z\left(\tfrac{j_2}{\Delta}\right) | D_n^T\right) = \operatorname{sgn}(j_1)\operatorname{sgn}(j_2)\frac{4}{n_v(\mathfrak{e})^2} \sum_{i_1=k_*-(j_1)_-+1}^{k_*+(j_1)_+} \sum_{i_2=k_*-(j_2)_-+1}^{k_*+(j_2)_+} \hat{\theta}_{i_1}^T \hat{\theta}_{i_2}^T \times \operatorname{Cov}(\psi_{i_1}(X), \psi_{i_2}(X)).$$

Let us now introduce the following definition.

Definition 5.4.7 *Let $(m_1, m_2, m_3) \in \mathbb{N}^3$ be three integers. Let $m_{(1)} \leqslant m_{(2)} \leqslant m_{(3)}$ be their non-decreasing rearrangement. Define $E_{m_1,m_2,m_3} = 0$ if $m_{(1)} = m_{(2)}$ or $m_{(2)} = m_{(3)}$ and*

$$E_{m_1,m_2,m_3} = \sum_{j_1=m_{(1)}+1}^{m_{(2)}} \sum_{j_2=m_{(2)}+1}^{m_{(3)}} \hat{\theta}_{j_1}^T \hat{\theta}_{j_2}^T \operatorname{Cov}(\psi_{j_1}(X), \operatorname{Cov}(\psi_{j_2}(X)) \tag{5.20}$$

if $m_{(1)} < m_{(2)} < m_{(3)}$.

Let $n_v = n - n_t$. The covariance can be broken down as follows: If $0 < j_1 \leqslant j_2$, conditionally on D_n^T.

$$\operatorname{Cov}\left(Z\left(\frac{j_1}{\Delta}\right), Z\left(\frac{j_2}{\Delta}\right)|D_n^T\right) = \operatorname{Var}\left(Z\left(\frac{j_1}{\Delta}\right)\right) + \frac{4}{n_v \mathfrak{e}^2} \sum_{i_1=k_*+1}^{k_*+j_1} \sum_{i_2=k_*+j_1+1}^{k_*+j_2} \hat{\theta}_{i_1}^T \hat{\theta}_{i_2}^T \operatorname{Cov}(\psi_{i_1}(X), \psi_{i_2}(X)).$$

If $j_1 \leqslant j_2 < 0$, symetrically,

$$\operatorname{Cov}\left(Z\left(\frac{j_1}{\Delta}\right), Z\left(\frac{j_2}{\Delta}\right)|D_n^T\right) = \operatorname{Var}\left(Z\left(\frac{j_2}{\Delta}\right)\right) + \frac{4}{n_v \mathfrak{e}^2} \sum_{i_2=k_*+j_2+1}^{k_*} \sum_{i_1=k_*+j_1+1}^{k_*+j_2} \hat{\theta}_{i_1}^T \hat{\theta}_{i_2}^T \operatorname{Cov}(\psi_{i_1}(X), \psi_{i_2}(X)).$$

Finally, if $j_1 < 0 < j_2$,

$$\operatorname{Cov}\left(Z\left(\frac{j_1}{\Delta}\right), Z\left(\frac{j_2}{\Delta}\right)|D_n^T\right) = \frac{-4}{n_v \mathfrak{e}^2} \sum_{i_1=k_*+j_1+1}^{k_*} \sum_{i_2=k_*+1}^{k_*+j_2} \hat{\theta}_{i_1}^T \hat{\theta}_{i_2}^T \operatorname{Cov}(\psi_{i_1}(X), \psi_{i_2}(X)).$$

It follows from the previous equations that for any $(k_1, k_2) \in \mathbb{N}^2$,

$$\operatorname{Cov}\left(Z\left(\frac{k_1-k_*}{\Delta}\right), Z\left(\frac{k_2-k_*}{\Delta}\right)\right) = \begin{cases} \operatorname{Var}(Z\left(\frac{k_1-k_*}{\Delta}\right)) + 4\frac{E_{k_*,k_1,k_2}}{n_v \mathfrak{e}^2} \text{ if } k_* < k_1 \leqslant k_2 \\ 4\frac{E_{k_*,k_1,k_2}}{n_v \mathfrak{e}^2} \text{ if } k_1 < k_* < k_2 \\ \operatorname{Var}(Z\left(\frac{k_2}{\Delta}\right)) + 4\frac{E_{k_*,k_1,k_2}}{n_v \mathfrak{e}^2} \text{ if } k_1 \leqslant k_2 < k_* \end{cases}. \tag{5.21}$$

Let $I, J \subset \{a_x\Delta, \ldots, b_x\Delta\}$. Assuming concentration around the expectation yields

$$\sum_{i\in I}\sum_{j\in J} \hat{\theta}_i^T \hat{\theta}_j^T \operatorname{Cov}(\psi_i(X), \psi_j(X)) \sim \sum_{i\in I}\sum_{j\in J} \theta_i\theta_j \operatorname{Cov}(\psi_i(X), \psi_j(X)) + \frac{1}{n_t}\sum_{i\in I}\sum_{j\in J} \operatorname{Cov}(\psi_i(X), \psi_j(X))^2.$$

Moreover, for any $(i_1, i_2) \in \mathbb{N}^2$,

$$\psi_{i_1}(X)\psi_{i_2}(X) = 2\cos(2i_1\pi X)\cos(2i_2\pi X) = \cos\left(2(i_1+i_2)\pi X\right) + \cos(2(i_1-i_2)\pi X)$$

and by definition, for all $i \in \mathbb{N}^*$, $\psi_i = \sqrt{2}\cos(2i\pi X)$, while $\psi_0 = 1 = \cos(0\pi x)$. As a result, $\psi_{i_1}(X)\psi_{i_2}(X) = \frac{\psi_{i_1+i_2}(X)+\psi_{|i_1-i_2|}(X)}{\sqrt{2}}$ if $i_1 \neq i_2$ and

$$\operatorname{Cov}(\psi_{i_1}(X), \psi_{i_2}(X)) = \frac{\theta_{i_1+i_2}}{\sqrt{2}} + \left(\frac{1-\delta_{i_1,i_2}}{\sqrt{2}} + \delta_{i_1,i_2}\right)\theta_{|i_2-i_1|} - \theta_{i_1}\theta_{i_2}.$$

By assumption, the sequence $|\theta_k|$ tends to 0 with a polynomial rate of convergence, hence for sequences $i_1 \sim i_2$ tending to $+\infty$, $\theta_{|i_1-i_2|}$ dominates $\theta_{i_1}\theta_{i_2}$ and $\theta_{i_1+i_2}$. Heuristically, it can thus be expected that

$$\sum_{i\in I}\sum_{j\in J}\hat{\theta}_i^T\hat{\theta}_j^T\operatorname{Cov}(\psi_i(X),\psi_j(X)) \sim \sum_{i\in I}\sum_{j\in J}\theta_i\theta_j\left(\frac{1-\delta_{i,j}}{\sqrt{2}}+\delta_{i,j}\right)\theta_{|j-i|} + \sum_{i\in I}\sum_{j\in J}\left(\frac{1-\delta_{i,j}}{\sqrt{2}}+\delta_{i,j}\right)^2\theta_{|j-i|}^2.$$

This leads to the following proposition, the rigourous proof of which can be found in the appendix (proposition 5.5.6).

Proposition 5.4.8 *Let P be the probability measure with pdf s on $[0;1]$, let $\theta_j = \langle s,\psi_j\rangle = P(\psi_j)$ and assume that the coefficients θ_j satisfy the hypotheses of Theorem 5.3.8. Let $\hat{\theta}_j^T = P^T\psi_j$. Let $I_k^1, I_k^2 \subset \{k_*+a_x\Delta,\dots,k_*+b_x\Delta\}$ be two intervals. Then the statistics*

$$U_{I_k^1,I_k^2} = \sum_{i\in I_k^1}\sum_{j\in I_k^2}\hat{\theta}_i^T\hat{\theta}_j^T[P(\psi_i\psi_j)-P\psi_iP\psi_j]$$

can be approximated in the following way: there exists two constants κ_4 and $u_3>0$ such that, with probability greater than $1-e^{-y}$,

$$U_{I_k^1,I_k^2} = \frac{1}{2}\frac{|I_k^1\cap I_k^2|}{n_t} + \left(1-\frac{1}{\sqrt{2}}\right)\sum_{i\in I_k^1\cap I_k^2}\theta_i^2 + \frac{1}{\sqrt{2}}\sum_{i\in I_k^1}\sum_{j\in I_k^2}\theta_i\theta_j\theta_{|i-j|} + \frac{1}{2n_t}\sum_{i\in I_k^1}\sum_{j\in I_k^2}\theta_{|i-j|}^2 \pm \kappa_4(y+\log n)^2(1+x)n^{-u_3}\mathcal{E}.$$

It is now possible to show that the terms E_{k_*,k_1,k_2} which appear in equation (5.21) are negligible compared to $\mathcal{E}$. That is the point of the following claim.

Claim 5.4.8.1 *Under the assumptions of Theorem 5.3.8, there exists constants $\kappa_7 \geqslant 0$ and $u_4>0$ such that for all $n\in\mathbb{N}$, $x>0$ and $(m_1,m_2,m_3)\in\{a_x\Delta,\dots,b_x\Delta\}^3$ such that $m_1<m_2<m_3$,*

$$\sum_{j_1=m_1+1}^{m_2}\sum_{j_2=m_2+1}^{m_3}\theta_{j_1}\theta_{j_2}\theta_{|j_1-j_2|} \leqslant \kappa_7(1+x)^2n^{-u_4}\mathcal{E} \tag{5.22}$$

$$\frac{1}{n_t}\sum_{j_1=m_1+1}^{m_2}\sum_{j_2=m_2+1}^{m_3}\theta_{|j_1-j_2|}^2 \leqslant \kappa_7(1+x)^2n^{-u_4}\mathcal{E}. \tag{5.23}$$

and moreover, for all $x>0$, with probability greater than $1-e^{-y}$, for any integers $(m_1,m_2,m_3)\in\{a_x\Delta,\dots,b_x\Delta\}^3$,

$$|E_{m_1,m_2,m_3}| \leqslant \kappa_7(1+x)^2(y+\log n)^2n^{-u_4}\mathcal{E}. \tag{5.24}$$

Proof Assume without loss of generality that $m_1 < m_2 < m_3$. We start by proving equation (5.22). First, changing variables from j_1, j_2 to $i = j_1, r = j_2 - j_1$ yields

$$\begin{aligned}
\sum_{j_1=m_1+1}^{m_2}\sum_{j_2=m_2+1}^{m_3}\theta_{j_1}\theta_{j_2}\theta_{|j_1-j_2|} &= \sum_{r\in\mathbb{N}}\theta_r \sum_{i=m_2+1-r}^{m_2}\mathbb{I}_{i\geqslant m_1+1}\mathbb{I}_{i+r\leqslant m_3}\theta_i\theta_{i+r} \\
&\leqslant \frac{1}{2}\sum_{r\leqslant r_0}|\theta_r| \sum_{i=(m_2+1-r)\vee(m_1+1)}^{m_2\wedge(m_3-r)} \theta_i^2+\theta_{i+r}^2 \\
&\quad + \frac{1}{2}\sum_{r>r_0}|\theta_r|\left[\sum_{j_1=m_1+1}^{m_2}\theta_{j_1}^2 + \sum_{j_2=m_2+1}^{m_3}\theta_{j_2}^2\right] \\
&\leqslant \frac{\|\theta\|_{\ell^1}}{2}\max_{1\leqslant r\leqslant r_0}\sum_{i=(m_2+1-r)\vee(m_1+1)}^{m_2\wedge(m_3-r)}\theta_i^2+\theta_{i+r}^2 \\
&\quad + \frac{1}{2}\sum_{r>r_0}|\theta_r|\sum_{i=a_x\Delta+1}^{b_x\Delta}\theta_{k_*+i}^2.
\end{aligned}$$

By claim 5.5.5.1 in appendix, for any $k \in \{m_1+1,\dots,m_3\} \subset \{a_x\Delta+1,\dots,b_x\Delta\}$, $\theta_k^2 \leqslant \kappa_3(1+x)^2 n^{-u_2}\mathfrak{e}$, therefore

$$\sum_{j_1=m_1+1}^{m_2}\sum_{j_2=m_2+1}^{m_3}\theta_{j_1}\theta_{j_2}\theta_{|j_1-j_2|} \leqslant r_0\frac{\|\theta\|_{\ell^1}}{2}\kappa_3(1+x)^2n^{-u_2}\mathfrak{e} + \frac{1}{2}\sum_{r>r_0}|\theta_r|\sum_{i=a_x\Delta+1}^{b_x\Delta}\theta_{k_*+i}^2.$$

By hypothesis H1 of Theorem 5.3.8,

$$\sum_{j\geqslant r_0+1}|\theta_j| \leqslant \sum_{j=r_0+1}^{+\infty}\sqrt{\sum_{i=j}^{+\infty}\theta_i^2} \leqslant \sum_{j=r_0+1}^{+\infty}\frac{c_1}{(j-1)^{1+\frac{\delta_1}{2}}} \leqslant \frac{2c_1}{\delta_1}r_0^{\frac{\delta_1}{2}}.$$

Thus, by lemma 5.4.1,

$$\sum_{j_1=m_1+1}^{m_2}\sum_{j_2=m_2+1}^{m_3}\theta_{j_1}\theta_{j_2}\theta_{|j_1-j_2|} \leqslant \frac{r_0\|\theta\|_{\ell^1}}{2}\kappa_3(1+x)^2n^{-u_2}\mathcal{E} + \frac{2c_1}{\delta_1}r_0^{-\frac{\delta_1}{2}}4(1+x)\mathcal{E}. \quad (5.25)$$

Let $r_0 = \lceil n^{\frac{2u_2}{2+\delta_1}}\rceil \leqslant 2n^{\frac{2u_2}{2+\delta_1}}$ and $u = \frac{\delta_1(u_2)}{2+\delta_1} > 0$. For all $n \geqslant 2$,

$$\sum_{j_1=m_1+1}^{m_2}\sum_{j_2=m_2+1}^{m_3}\theta_{j_1}\theta_{j_2}\theta_{|j_1-j_2|} \leqslant \left[\|\theta\|_{\ell^1}\kappa_3 + 8\frac{c_1}{\delta_1}\right](1+x)^2n^{-u}\mathcal{E}, \quad (5.26)$$

which proves equation (5.22).

Moreover,

$$\begin{aligned}\sum_{j_1=m_1+1}^{m_2}\sum_{j_2=m_2+1}^{m_3}\theta^2_{|j_1-j_2|} &= \sum_{r\in\mathbb{N}}\theta_r^2\left|\{j_1 : (m_1+1\leqslant j_1\leqslant m_2)\wedge(m_2+1\leqslant j_1+r\leqslant m_3)\}\right| \\ &\leqslant \sum_{r\in\mathbb{N}}\theta_r^2\left[(m_3-m_1)\wedge r\right] \\ &\leqslant r_0\sum_{r=0}^{r_0}\theta_r^2 + (m_3-m_1)\sum_{r>r_0}\theta_r^2 \\ &\leqslant r_0\left\|s\right\|^2 + (b_x-a_x)\Delta\sum_{r>r_0}\theta_r^2 \\ &\leqslant r_0\left\|s\right\|^2 + 2(1+x)\Delta\frac{c_1}{r_0^{2+\delta_1}},\end{aligned}$$

by hypothesis H1 of Theorem 5.3.8 and lemma 5.4.1. Let now $r_0 = \left\lceil \Delta^{\frac{1}{3}}\right\rceil$. Since $\Delta \geqslant 1$, it follows that:

$$\begin{aligned}\frac{1}{n_t}\sum_{j_1=m_1+1}^{m_2}\sum_{j_2=m_2+1}^{m_3}\theta^2_{|j_1-j_2|} &\leqslant \frac{\Delta^{\frac{1}{3}}+1}{n_t}\left\|s\right\|^2 + 2c_1(1+x)\frac{\Delta}{n_t}\left(\Delta\right)^{\frac{-2}{3}} \\ &\leqslant \left[2\frac{\left\|s\right\|^2}{\left(\Delta\right)^{\frac{2}{3}}}\mathcal{E} + 2c_1(1+x)\mathcal{E}(\Delta)^{-\frac{2}{3}}\right] \\ &\leqslant \left[2\left\|s\right\|^2 + 2c_1(1+x)\right]\frac{\mathcal{E}}{\left(\Delta\right)^{\frac{2}{3}}}.\end{aligned}$$

On the other hand, $\Delta \geqslant \frac{n_t}{n-n_t} \geqslant n^{\delta_3}$ by hypothesis H4 of Theorem 5.3.8. There exists therefore $\kappa(c_1, \left\|s\right\|^2)$ such that, for any n,

$$\frac{1}{n_t}\sum_{j_1=m_1+1}^{m_2}\sum_{j_2=m_2+1}^{m_3}\theta^2_{|j_1-j_2|} \leqslant \kappa(1+x)n^{-\frac{2\delta_3}{3}}\mathcal{E}, \tag{5.27}$$

which proves equation (5.23). Since for all $x>0$, $(b_x-a_x)\leqslant 2(1+x)\Delta\leqslant\kappa(1+x)n$, by proposition 5.4.8 and a union bound, there exists an event A of probability greater than $1-e^{-y}$ and a constant κ such that, if $a_x\leqslant m_1<m_2<m_3\leqslant b_x$, then

$$\begin{aligned}\left|E_{m_1,m_2,m_3}\right| = \frac{1}{\sqrt{2}}\sum_{j_1=m_1+1}^{m_2}\sum_{j_2=m_2+1}^{m_3}\theta_{j_1}\theta_{j_2}\theta_{|j_1-j_2|} + \frac{1}{2n_t}\sum_{j_1=m_1+1}^{m_2}\sum_{j_2=m_2+1}^{m_3}\theta^2_{|j_1-j_2|} \\ + \kappa(y+\log(2+x)+\log n)^2(1+x)n^{-u_3}\mathcal{E}.\end{aligned} \tag{5.28}$$

From equations (5.28), (5.26) and (5.27), equation (5.24) follows with $u_4 = \min\left(u_3, \frac{2\delta_3}{3}, \frac{\delta_1 u_2}{2+\delta_1}\right)$. ■

Let then $g_n^0 : \left[\frac{-k_*}{\Delta}; +\infty\right[\to \mathbb{R}$ be defined first for all $\alpha \in \left\{\frac{j}{\Delta} : j \in \mathbb{N} - k_*\right\}$ by

$$\forall \alpha \in \left\{\frac{j}{\Delta} : (j+k_*) \in \mathbb{N}\right\}, g_n^0(\alpha) = \mathrm{sgn}(\alpha)\mathrm{Var}(Z_\alpha), \tag{5.29}$$

then for all $\alpha \in \left[\frac{-k_*}{\Delta}; +\infty\right[$ by linear interpolation (hence in general, $g_n^0(\alpha) \neq \mathrm{Var}(Z_\alpha)$). Let $K(g_n^0)$ be given by definition 5.4.6, then by equation (5.21) and claim 5.4.8.1, with probability greater than $1 - e^{-y}$, for any $x > 0$,

$$\begin{aligned}
&\max_{(j_1,j_2)\in\{a_x\Delta,\dots,b_x\Delta\}^2} \left|\mathrm{Cov}\left(Z\left(\frac{j_1}{\Delta}\right), Z\left(\frac{j_2}{\Delta}\right)\right) - K\left(g_n^0, \frac{j_1}{\Delta}, \frac{j_2}{\Delta}\right)\right| \\
&\quad \leqslant 4\kappa_7(1+x)^2 + (y+\log n)^2 n^{-u_4}\frac{\mathcal{E}}{n_v \mathfrak{e}^2} \\
&\quad \leqslant 4\kappa_7(1+x)^2(y+\log n)^2 n^{-u_4}.
\end{aligned} \tag{5.30}$$

Moreover, for any $j \in \mathbb{Z} \cap [a_x\Delta; b_x\Delta]$, by definition of Z,

$$\begin{aligned}
\mathrm{sgn}(j) g_n^0\left(\frac{j}{\Delta}\right) = \mathrm{Var}\left(Z\left(\frac{j}{\Delta}\right)\right) = \frac{4}{n_v\mathfrak{e}^2} \sum_{i_1=k_*-(j)_-+1}^{k_*+(j)_+} \sum_{i_2=k_*-(j)_-+1}^{k_*+(j)_+} \hat{\theta}_{i_1}^T \hat{\theta}_{i_2}^T \Big[\frac{\theta_{i_1+i_2}}{\sqrt{2}} \\
+ \Big(\frac{1-\delta_{i_1,i_2}}{\sqrt{2}} + \delta_{i_1,i_2}\Big)\theta_{|i_1-i_2|} - \theta_{i_1}\theta_{i_2}\Big].
\end{aligned} \tag{5.31}$$

Moreover, since $\mathfrak{e}^2 = \frac{\mathcal{E}}{n_v}$,

$$\begin{aligned}
\frac{4}{n_v\mathfrak{e}^2}\frac{1}{2}\frac{j}{n_t} &= \frac{4}{n_v\mathfrak{e}^2}\frac{1}{2}\frac{j}{\Delta}\frac{\Delta}{n_t} \\
&= \frac{\mathcal{E}}{n_v\mathfrak{e}^2} 2\frac{j}{\Delta} \\
&= 2\frac{j}{\Delta} \ .
\end{aligned}$$

Thus, by proposition 5.4.8, with probability greater than $1 - e^{-y}$,

$$\begin{aligned}\operatorname{sgn}(j) g_n^0\left(\frac{j}{\Delta}\right) = 2\frac{|j|}{\Delta} &+ \frac{4}{n_v \mathfrak{e}^2}\left(1 - \frac{1}{\sqrt{2}}\right) \sum_{i=k_*-(j)_-+1}^{k_*+(j)_+} \theta_i^2 \\ &+ \frac{4}{n_v \mathfrak{e}^2} \sum_{i_1=k_*-(j)_-+1}^{k_*+(j)_+} \sum_{i_2=k_*-(j)_-+1}^{k_*+(j)_+} \theta_{i_1}\theta_{i_2} \frac{\theta_{|i_1-i_2|}}{\sqrt{2}} \\ &+ \frac{4}{n_v \mathfrak{e}^2}\frac{1}{2n_t} \sum_{i_1=k_*-(j)_-+1}^{k_*+(j)_+} \sum_{i_2=k_*-(j)_-+1}^{k_*+(j)_+} \theta_{|i_1-i_2|}^2 \\ &\pm 4\kappa_4 (y + \log n)^2 (1+x) n^{-u_3}. \end{aligned} \tag{5.32}$$

Let g_n^1 be defined for all $\alpha = \frac{j}{\Delta}, j \in \mathbb{Z} \cap [-k_*(n_t); +\infty)$ by

$$\begin{aligned}\operatorname{sgn}(j) g_n^1\left(\frac{j}{\Delta}\right) &= \frac{4}{n_v \mathfrak{e}^2} \sum_{i_1=k_*-(j)_-+1}^{k_*+(j)_+} \sum_{i_2=k_*-(j)_-+1}^{k_*+(j)_+} \theta_{i_1}\theta_{i_2} \frac{\theta_{|i_1-i_2|}}{\sqrt{2}} \\ &\quad + \frac{4}{n_v \mathfrak{e}^2}\left(1 - \frac{1}{\sqrt{2}}\right) \sum_{i=k_*-(j)_-+1}^{k_*+(j)_+} \theta_i^2 \\ &= \frac{4}{n_v \mathfrak{e}^2} \sum_{i_1=k_*-(j)_-+1}^{k_*+(j)_+} \sum_{i_2=k_*-(j)_-+1}^{k_*+(j)_+} \theta_{i_1}\theta_{i_2} \left(\frac{1 - \delta_{i_1,i_2}}{\sqrt{2}} + \delta_{i_1,i_2}\right) \theta_{|i_1-i_2|}, \end{aligned} \tag{5.33}$$

and for all $\alpha \in \left[-\frac{k_*}{\Delta}; +\infty\right[$ by linear interpolation.

We will now apply lemma 5.5.8 to g_n^1. Let $x > 0$ and $(k_1, k_2) \in \{k_* + a_x\Delta, \ldots, k_* + b_x\Delta\}^2$ be such that $k_1 < k_2$. Thus:

- If $k_* \leqslant k_1$,

$$\begin{aligned} g_n^1\left(\frac{k_2 - k_*}{\Delta}\right) - g_n^1\left(\frac{k_1 - k_*}{\Delta}\right) &= \frac{4}{n_v \mathfrak{e}^2}\Bigg(\sum_{i=k_*+1}^{k_2} \sum_{j=k_*+1}^{k_2} \theta_i\theta_j \left(\frac{1-\delta_{i,j}}{\sqrt{2}} + \delta_{i,j}\right)\theta_{|i-j|} \\ &\quad - \sum_{i=k_*+1}^{k_1} \sum_{j=k_*+1}^{k_1} \theta_i\theta_j \left(\frac{1-\delta_{i,j}}{\sqrt{2}} + \delta_{i,j}\right)\theta_{|i-j|}\Bigg) \\ &= \frac{4}{n_v \mathfrak{e}^2}\Bigg(\sum_{i=k_1+1}^{k_2} \sum_{j=k_1+1}^{k_2} \theta_i\theta_j \left(\frac{1-\delta_{i,j}}{\sqrt{2}} + \delta_{i,j}\right)\theta_{|i-j|} \\ &\quad + 2\sum_{i=k_1+1}^{k_2} \sum_{j=k_*+1}^{k_1} \theta_i\theta_j \frac{\theta_{|i-j|}}{\sqrt{2}}\Bigg). \end{aligned}$$

By lemma 5.5.7 in appendix,

$$0 \leqslant \sum_{i=k_1+1}^{k_2} \sum_{j=k_1+1}^{k_2} \theta_i \theta_j \left(\frac{1-\delta_{i,j}}{\sqrt{2}} + \delta_{i,j} \right) \theta_{|i-j|} \leqslant \|s\|_\infty \sum_{i=k_1+1}^{k_2} \theta_i^2.$$

Thus by equation (5.22) from claim 5.4.8.1,

$$g_n^1\left(\tfrac{k_2-k_*}{\Delta}\right) - g_n^1\left(\tfrac{k_1-k_*}{\Delta}\right) \leqslant \frac{4\|s\|_\infty}{n_v \mathfrak{e}^2} \sum_{i=k_1+1}^{k_2} \theta_i^2 + \kappa_7 (1+x)^2 n^{-u_4} \frac{4\sqrt{2}}{n_v \mathfrak{e}^2} \mathcal{E}$$

$$g_n^1\left(\tfrac{k_2-k_*}{\Delta}\right) - g_n^1\left(\tfrac{k_1-k_*}{\Delta}\right) \geqslant -\kappa_7 (1+x)^2 n^{-u_3} \frac{4\sqrt{2}}{n_v \mathfrak{e}^2} \mathcal{E}.$$

- If $k_2 \leqslant k_*$,

$$\begin{aligned} g_n^1\left(\frac{k_2-k_*}{\Delta}\right) - g_n^1\left(\frac{k_1-k_*}{\Delta}\right) &= \frac{4}{n_v \mathfrak{e}^2} \Bigg(\sum_{i=k_1+1}^{k_*} \sum_{j=k_1+1}^{k_*} \theta_i \theta_j \left(\frac{1-\delta_{i,j}}{\sqrt{2}} + \delta_{i,j} \right) \theta_{|i-j|} \\ &\quad - \sum_{i=k_2+1}^{k_*} \sum_{j=k_2+1}^{k_*} \theta_i \theta_j \left(\frac{1-\delta_{i,j}}{\sqrt{2}} + \delta_{i,j} \right) \theta_{|i-j|} \Bigg) \\ &= \frac{4}{n_v \mathfrak{e}^2} \Bigg(\sum_{i=k_1+1}^{k_2} \sum_{j=k_1+1}^{k_2} \theta_i \theta_j \left(\frac{1-\delta_{i,j}}{\sqrt{2}} + \delta_{i,j} \right) \theta_{|i-j|} \\ &\quad + 2 \sum_{i=k_1+1}^{k_2} \sum_{j=k_2+1}^{k_*} \theta_i \theta_j \frac{\theta_{|i-j|}}{\sqrt{2}} \Bigg). \end{aligned}$$

In the same way, by lemma 5.5.7 and equation (5.22) from claim 5.4.8.1,

$$g_n^1\left(\tfrac{k_2-k_*}{\Delta}\right) - g_n^1\left(\tfrac{k_1-k_*}{\Delta}\right) \leqslant \frac{4\|s\|_\infty}{n_v \mathfrak{e}^2} \sum_{i=k_1+1}^{k_2} \theta_i^2 + \kappa_7 (1+x)^2 n^{-u_4} \frac{4\sqrt{2}}{n_v \mathfrak{e}^2} \mathcal{E}$$

$$g_n^1\left(\tfrac{k_2-k_*}{\Delta}\right) - g_n^1\left(\tfrac{k_1-k_*}{\Delta}\right) \geqslant -\kappa_7 (1+x)^2 n^{-u_3} \frac{4\sqrt{2}}{n_v \mathfrak{e}^2} \mathcal{E}.$$

- If $k_1 \leqslant k_2 \leqslant k_2$,

$$g_n^1\left(\tfrac{k_2-k_*}{\Delta}\right) - g_n^1\left(\tfrac{k_1-k_*}{\Delta}\right) = g_n^1\left(\tfrac{k_2-k_*}{\Delta}\right) - g_n(0) + g_n(0) - g_n^1\left(\tfrac{k_1-k_*}{\Delta}\right),$$

therefore by the two previous cases,

$$0 \leqslant g_n^1\left(\tfrac{k_2-k_*}{\Delta}\right) - g_n^1\left(\tfrac{k_1-k_*}{\Delta}\right) \leqslant \frac{4\|s\|_\infty}{n_v \mathfrak{e}^2} \sum_{i=k_1+1}^{k_2} \theta_i^2 + \kappa_7 (1+x)^2 n^{-u_4} \frac{8\sqrt{2}}{n_v \mathfrak{e}^2} \mathcal{E}.$$

By definition $\mathfrak{e}^2 = \frac{\mathcal{E}}{n_v}$ therefore for any $x > 0$ and $(j_1, j_2) \in [a_x\Delta; b_x\Delta]^2$ such that $j_1 \leqslant j_2$,

$$-4\sqrt{2}\kappa_7(1+x)^2 n^{-u_4} \leqslant g_n^1\left(\tfrac{j_2}{\Delta}\right) - g_n^1\left(\tfrac{j_1}{\Delta}\right) \leqslant \frac{4\left\|s\right\|_\infty}{n_v\mathfrak{e}^2} \sum_{i=j_1+1}^{j_2} \theta^2_{k_*+i} + 8\sqrt{2}\kappa_7(1+x)^2 n^{-u_4}. \tag{5.34}$$

Moreover, for any j_1, j_2 such that $0 < j_1 < j_2$, $\theta^2_{k_*+j_i} \leqslant \frac{1}{n_t}$ hence

$$\begin{aligned}\left\|s\right\|_\infty \sum_{i=j_1+1}^{j_2} \theta^2_{k_*+i} &\leqslant \left\|s\right\|_\infty \sum_{j=k_*+j_1+1}^{k_*+j_2} [\theta_j^2 - \frac{1}{n_t}] + \frac{j_2-j_1}{\Delta}\left\|s\right\|_\infty \frac{\Delta}{n_t} \\ &= -\left\|s\right\|_\infty \mathfrak{e}[f_n\left(\frac{j_2}{\Delta}\right) - f_n\left(\frac{j_1}{\Delta}\right)] + \frac{j_2-j_1}{\Delta}\left\|s\right\|_\infty \mathcal{E}. \end{aligned} \tag{5.35}$$

For j_1, j_2 such that $j_1 < j_2 \leqslant 0$, $\theta^2_{k_*+j_i} \geqslant \frac{1}{n_t}$ hence

$$\begin{aligned}\left\|s\right\|_\infty \sum_{i=j_1+1}^{j_2} \theta^2_{k_*+i} &\leqslant \left\|s\right\|_\infty \sum_{j=k_*+j_1+1}^{k_*+j_2} [\theta_j^2 - \frac{1}{n_t}] + \frac{j_2-j_1}{\Delta}\left\|s\right\|_\infty \frac{\Delta}{n_t} \\ &= \mathfrak{e}[f_n\left(\frac{j_1}{\Delta}\right) - f_n\left(\frac{j_2}{\Delta}\right)] + \frac{j_2-j_1}{\Delta}\left\|s\right\|_\infty \mathcal{E}. \end{aligned} \tag{5.36}$$

By equations (5.34), (5.35) and (5.35), it follows that, for any $(j_1, j_2) \in \left([a_x\Delta; b_x\Delta] \cap \mathbb{Z}\right)^2$,

$$\begin{aligned} g_n^1\left(\frac{j_2}{\Delta}\right) - g_n^1\left(\frac{j_1}{\Delta}\right) &\leqslant -4\frac{\left\|s\right\|_\infty}{n_v\mathfrak{e}}\left[f_n\left(\frac{j_2}{\Delta}\right) - f_n\left(\frac{j_1}{\Delta}\right)\right] + 4\left\|s\right\|_\infty \frac{j_2-j_1}{\Delta} + 8\sqrt{2}\kappa_7(1+x)^2 n^{-u_4} \\ &\leqslant -4\frac{\left\|s\right\|_\infty}{n_v\mathfrak{e}}\left[f_n\left(\frac{j_2}{\Delta}\right) - f_n\left(\frac{j_1}{\Delta}\right)\right] + 4\left\|s\right\|_\infty \frac{j_2-j_1}{\Delta} + 8\sqrt{2}\kappa_7(1+x)^2 n^{-u_4}. \end{aligned} \tag{5.37}$$

To extend the lower bound given by equation (5.34), notice that for any $(\alpha_1, \alpha_2) \in [a_x; b_x]^2$ such that $\alpha_1 < \alpha_2$,

- if $\lfloor \alpha_1\Delta \rfloor = \lfloor \alpha_2\Delta \rfloor \leqslant \alpha_1 < \alpha_2 \leqslant \lfloor \alpha_1\Delta \rfloor + 1$, by linearity of g_n^1 on $[\lfloor \alpha_1\Delta \rfloor; \lfloor \alpha_1\Delta \rfloor + 1]$,
$$g_n^1(\alpha_2) - g_n^1(\alpha_1) \geqslant -\left[g_n^1\left(\frac{\lfloor \alpha_1\Delta \rfloor + 1}{\Delta}\right) - g_n^1\left(\frac{\lfloor \alpha_1\Delta \rfloor}{\Delta}\right)\right]_-,$$
- otherwise, $\lfloor \alpha_1\Delta \rfloor + 1 \leqslant \lfloor \alpha_2\Delta \rfloor$, therefore by linearity of $(u, v) \mapsto g_n^1(u) - g_n^1(v)$ on $\frac{1}{\Delta}[\lfloor \alpha_1\Delta \rfloor; \lfloor \alpha_1\Delta \rfloor + 1] \times \frac{1}{\Delta}[\lfloor \alpha_2\Delta \rfloor; \lfloor \alpha_2\Delta \rfloor + 1]$,
$$\begin{aligned} g_n^1(\alpha_2) - g_n^1(\alpha_1) \geqslant \min\Big\{ & g_n^1(u) - g_n^1(v) : \\ & (u, v) \in \big\{\frac{\lfloor \alpha_2\Delta \rfloor}{\Delta}; \frac{\lfloor \alpha_2\Delta \rfloor + 1}{\Delta}\big\} \times \big\{\frac{\lfloor \alpha_1\Delta \rfloor}{\Delta}; \frac{\lfloor \alpha_1\Delta \rfloor + 1}{\Delta}\big\}\Big\}. \end{aligned}$$

In all cases,

$$g_n^1(\alpha_2)-g_n^1(\alpha_1) \geqslant -\max\left\{\left[g_n^1\left(\frac{j_2}{\Delta}\right)-g_n^1\left(\frac{j_1}{\Delta}\right)\right]_- : j_1 \leqslant j_2, (j_1,j_2)\in\{a_x\Delta,\dots,b_x\Delta\}^2\right\}. \tag{5.38}$$

Thus by equation (5.34), for any $x>0$:

$$\forall(\alpha_1,\alpha_2)\in[a_x;b_x]^2, \alpha_1<\alpha_2 \implies g_n^1(\alpha_2)-g_n^1(\alpha_1) \geqslant -4\sqrt{2}\kappa_7(1+x)^2 n^{-u_4}. \tag{5.39}$$

By the same argument applied to the function

$$\alpha \mapsto g_n^1(\alpha)+4\frac{\|s\|_\infty}{n_v\mathfrak{e}}f_n(\alpha)-4\|s\|_\infty\alpha,$$

which is piecewise linear on the partition $\left\{\left[\frac{j}{\Delta};\frac{j+1}{\Delta}\right[: j\in\{a_x\Delta,\dots,b_x\Delta\}\right\}$, equation (5.37) extends to $[a_x;b_x]$ for any $x>0$:

$$\begin{aligned}\forall(\alpha_1,\alpha_2)\in[a_x;b_x]^2, g_n^1(\alpha_2)-g_n^1(\alpha_1) \leqslant 4\frac{\|s\|_\infty}{n_v\mathfrak{e}}[f_n(\alpha_1)-f_n(\alpha_2)]+4\|s\|_\infty[\alpha_2-\alpha_1]\\ +8\sqrt{2}\kappa_7(1+x)^2n^{-u_4}.\end{aligned} \tag{5.40}$$

Let $\varepsilon_d:\alpha\mapsto\inf_{x\in\mathbb{R}_+:b_x\geqslant\alpha}8\sqrt{2}\kappa_7(1+x)^2n^{-u_4}$. The function ε_d is non-decreasing by definition, and $\varepsilon_d(0)\geqslant 8\kappa_7 n^{-u_4}>0$. Furthermore, by equations (5.39) and (5.40),

$$\begin{aligned}&\forall(\alpha_1,\alpha_2)\in\mathbb{R}_+^2, \alpha_1\leqslant\alpha_2 \implies\\ &-\varepsilon_d(\alpha_2)\leqslant g_n^1(\alpha_2)-g_n^1(\alpha_1)\leqslant 4\frac{\|s\|_\infty}{n_v\mathfrak{e}}[f_n(\alpha_1)-f_n(\alpha_2)]+4\|s\|_\infty[\alpha_2-\alpha_1]+\varepsilon_d(\alpha_2).\end{aligned}$$

In this situation, lemma 5.5.8 applies with $g=g_n^1$, $h_+=-8\|s\|_\infty f_n+8\|s\|_\infty\mathrm{Id}$ and $\varepsilon=2\varepsilon_d$. It guarantees existence of a non-decreasing function $g_{n,+}^2:\mathbb{R}_+\to\mathbb{R}_+$ such that $g_{n,+}^2(0)=0$,

$$\sup_{\alpha\in\mathbb{R}_+}\frac{|g_n^1(\alpha)-g_{n,+}^2(\alpha)|}{2\varepsilon_d(\alpha)}\leqslant 6$$

and for all α_1,α_2 such that $\alpha_1\leqslant\alpha_2$,

$$g_{n,+}^2(\alpha_2)-g_{n,+}^2(\alpha_1)\leqslant 8\frac{\|s\|_\infty}{n_v\mathfrak{e}}[f_n(\alpha_1)-f_n(\alpha_2)]+8\|s\|_\infty[\alpha_2-\alpha_1]$$

Symetrically, let $\varepsilon_g:\alpha\mapsto\inf_{x\in\mathbb{R}_+:-a_x\geqslant\alpha}8\sqrt{2}\kappa_7(1+x)^2n^{-u_4}$, defined on $\left[0;\frac{k_*(n_t)}{\Delta}\right]$. ε_g is non-decreasing by definition. Furthermore, $\varepsilon_g(0)\geqslant 8\kappa_7 n^{-u_4}>0$. By equations (5.39) and (5.40),

$$\begin{aligned}\forall(\alpha_1,\alpha_2)\in\mathbb{R}_+^2, \alpha_1\leqslant\alpha_2 \implies -\varepsilon_g(\alpha_2)&\leqslant g_n^1(-\alpha_1)-g_n^1(-\alpha_2)\\ &\leqslant 4\frac{\|s\|_\infty}{n_v\mathfrak{e}}[f_n(-\alpha_2)-f_n(-\alpha_1)]\\ &\quad+4\|s\|_\infty[\alpha_2-\alpha_1]+\varepsilon_g(\alpha_2).\end{aligned}$$

In this situation, lemma 5.5.8 applies with $g = -g_n^1(-\cdot)$, $h_+ = 8\|s\|_\infty f_n(-\cdot) + 8\|s\|_\infty \mathrm{Id}$, $\varepsilon = 2\varepsilon_g$. It guarantees existence of a function $g_{n,-}^2 : \left[0; \frac{k_*(n_t)}{\Delta}\right] \to \mathbb{R}_+$ such that $g_{n,-}^2(0) = 0$,

$$\sup_{\alpha \in \left[0; \frac{k_*(n_t)}{\Delta}\right]} \frac{|-g_n^1(-\alpha) - g_{n,-}^2(\alpha)|}{2\varepsilon_g(\alpha)} \leqslant 6$$

and for any α_1, α_2 such that $\alpha_1 \leqslant \alpha_2$,

$$g_{n,-}^2(\alpha_2) - g_{n,-}^2(\alpha_1) \leqslant 8\frac{\|s\|_\infty}{n_v \mathfrak{e}}[f_n(-\alpha_2) - f_n(-\alpha_1)] + 8\|s\|_\infty [\alpha_2 - \alpha_1].$$

Let then $g_n^2 : \alpha \mapsto g_{n,+}^2(\alpha)\mathbb{I}_{\alpha \geqslant 0} - g_{n,-}^2(-\alpha)\mathbb{I}_{\alpha<0}$ and $\varepsilon(\alpha) = \varepsilon_d(\alpha)\mathbb{I}_{\alpha\geqslant 0} + \varepsilon_g(-\alpha)\mathbb{I}_{x<0}$, which yields

$$\left\| \frac{g_n^2 - g_n^1}{2\varepsilon} \right\|_\infty \leqslant 6 \tag{5.41}$$

and

$$\forall (\alpha_1, \alpha_2) \in \mathbb{R}^2, g_n^2(\alpha_2) - g_n^2(\alpha_1) \leqslant 8\frac{\|s\|_\infty}{n_v \mathfrak{e}}[f_n(\alpha_1) - f_n(\alpha_2)] + 8\|s\|_\infty [\alpha_2 - \alpha_1]. \tag{5.42}$$

By definition of ε, for any $x > 0$ and any $\alpha \in [a_x, b_x]$, $\varepsilon(\alpha) \leqslant 8\sqrt{2}\kappa_7(1+x)^2 n^{-u_4}$, hence

$$\forall x > 0, \forall \alpha \in [a_x; b_x], |g_n^2(\alpha) - g_n^1(\alpha)| \leqslant 96\sqrt{2}\kappa_7(1+x)^2 n^{-u_4}. \tag{5.43}$$

Let then:

$$g_n : \alpha \mapsto g_n^2(\alpha) + 4\|s\|^2 \alpha. \tag{5.44}$$

Since g_n^2 is non-decreasing, $g_n(\alpha_2) - g_n(\alpha_1) \geqslant 4\|s\|^2 [\alpha_2 - \alpha_1]$, which proves equation 4 of Theorem 5.3.8. Moreover, equation (5.42) yields equation (5.12) of Theorem 5.3.8.

Let now $x > 0$ be fixed until the end of this section. By definition of g_n^1 (equation (5.33)), equations (5.32), (5.43) and since the functions g_n^0 and $\alpha \mapsto 4\|s\|^2 \alpha$ are piecewise linear on the partition $\left(\left[\frac{j}{\Delta}; \frac{j+1}{\Delta}\right[\right)_{a_x\Delta \leqslant j \leqslant b_x\Delta - 1}$, with probability greater than $1 - e^{-y}$,

$$\begin{aligned}
\left\|g_n^0 - g_n\right\|_\infty &\leqslant \left\|g_n^1 - g_n^2\right\|_\infty + \kappa_4(y + \log n)^2(1+x)n^{-u_3} \\
&\quad + \max_{a_x\Delta \leqslant j \leqslant b_x\Delta} \left| \frac{\mathrm{sgn}(j)}{n_v \mathfrak{e}^2} \frac{4}{2n_t} \sum_{i_1 = k_* + (j)_- + 1}^{k_* + (j)_+} \sum_{i_2 = k_* + (j)_- + 1}^{k_* + (j)_+} \theta^2_{|i_1 - i_2|} - \frac{(4\|s\|^2 - 2)j}{\Delta} \right| \\
&\leqslant \max_{a_x\Delta \leqslant j \leqslant b_x\Delta} \left| \frac{4}{n_v \mathfrak{e}^2} \frac{1}{2n_t} \sum_{i_1 = k_* + (j)_- + 1}^{k_* + (j)_+} \sum_{i_2 = k_* + (j)_- + 1}^{k_* + (j)_+} \theta^2_{|i_1 - i_2|} - 4\left(\|s\|^2 - \tfrac{1}{2}\right) \frac{|j|}{\Delta} \right| \\
&\quad + 96\sqrt{2}\kappa_7(1+x)^2 n^{-u_4} + \kappa_4(y + \log n)^2(1+x)n^{-u_3}.
\end{aligned} \tag{5.45}$$

It remains to bound the max. By parity in j of the sum, one can assume $0 \leqslant j \leqslant \max(|a_x|, |b_x|)\Delta$ instead of $a_x\Delta \leqslant j \leqslant b_x\Delta|]$. Let therefore $j \in \{0, \dots, \max(|a_x|, |b_x|)\Delta\}$, then

$$\begin{aligned}\frac{1}{2n_t}\sum_{i_1=k_*+1}^{k_*+j}\sum_{i_2=k_*+1}^{k_*+j}\theta^2_{|i_1-i_2|} &= \frac{|j|}{2n_t} + \frac{1}{2n_t}\sum_{r\in\mathbb{N}^*} 2\left|\{i : k_* \leqslant i \leqslant i+r \leqslant k_*+j\}\right|\theta_r^2 \\ &= \frac{|j|}{2n_t} + \frac{1}{n_t}\sum_{r=1}^{+\infty}(j-r)_+\theta_r^2. \end{aligned} \tag{5.46}$$

Furthermore, for all $r_0 \in \mathbb{N}^*$,

$$\begin{aligned}\left|\frac{1}{n_t}\sum_{r=1}^{+\infty}(j-r)_+\theta_r^2 - \frac{j}{n_t}(\|s\|^2-1)\right| &\leqslant \frac{1}{n_t}\sum_{r=1}^{+\infty}\theta_r^2\left|(j-r)_+ - j\right| \\ &\leqslant \frac{r_0}{n_t}\sum_{r=1}^{r_0}\theta_r^2 + \frac{j}{n_t}\sum_{r=r_0+1}^{+\infty}\theta_r^2 \\ &\leqslant \|s\|^2\frac{r_0}{n_t} + \max(|a_x|,|b_x|)\frac{\Delta}{n_t}\times\frac{c_1}{r_0^2},\end{aligned}$$

by hypothesis H1 of Theorem 5.3.8. By setting $r_0 = \lceil(\Delta)^{\frac{1}{3}}\rceil \leqslant 2(\Delta)^{\frac{1}{3}}$ (car $\Delta \geqslant 1$), it follows that

$$\begin{aligned}\left|\frac{1}{n_t}\sum_{r=1}^{+\infty}(j-r)_+\theta_r^2 - \frac{j}{n_t}(\|s\|^2-1)\right| &\leqslant \left[2\|s\|^2 + c_1\max(|a_x|,|b_x|)\right]\frac{(\Delta)^{\frac{1}{3}}}{n_t} \\ &\leqslant \left[2\|s\|^2 + 2c_1(1+x)\right](\Delta)^{-\frac{2}{3}}\mathcal{E} \text{ by lemma 5.4.1.}\end{aligned}$$

Let $\kappa = \kappa(c_1, \|s\|)$. Since by hypothesis H4 of Theorem 5.3.8, $\Delta \geqslant \frac{n_t}{n-n_t} \geqslant n^{\delta_3}$,

$$\left|\frac{1}{n_t}\sum_{r=1}^{+\infty}(j-r)_+\theta_r^2 - \frac{j}{n_t}(\|s\|^2-1)\right| \leqslant \kappa(1+x)n^{-\frac{2\delta_3}{3}}\mathcal{E}.$$

By equation (5.46) and since $\frac{j}{n_t} = \frac{j}{\Delta}\mathcal{E}$, for any $j \in [0; \max(|a_x|, |b_x|)\Delta]$.

$$\left|\frac{1}{2n_t}\sum_{i_1=k_*+1}^{k_*+j}\sum_{i_2=k_*+1}^{k_*+j}\theta^2_{|i_1-i_2|} - \frac{j}{\Delta}\left(\|s\|^2 - \frac{1}{2}\right)\mathcal{E}\right| \leqslant \kappa(1+x)n^{-\delta_3}\mathcal{E}.$$

By equation (5.45) and since $\frac{\mathcal{E}}{n_v\mathfrak{e}^2} = 1$, it follows that, with probability greater than $1-e^{-y}$,

$$\left\|g_n^0 - g_n\right\|_\infty \leqslant 96\sqrt{2}\kappa_7(1+x)^2n^{-u_4} + \kappa_4(y+\log n)^2(1+x)n^{-u_3} + 4\kappa(1+x)n^{-\delta_3}. \tag{5.47}$$

Let $\kappa = 96\sqrt{2}\kappa_7 + \kappa_4 + 4\kappa$ and $u_5 = \min(u_4, u_3, \delta_3)$, it then follows from definition 5.4.6 of K that with probability greater than $1 - e^{-y}$,

$$\left\|K(g_n^0) - K(g_n)\right\|_\infty \leqslant \left\|g_n^0 - g_n\right\|_\infty \leqslant \kappa(1+x)^2(y+\log n)^2 n^{-u_5}.$$

By equation (5.30), it follows that that, with probability greater than $1 - e^{-y}$, for any $(j_1, j_2) \in \{a_x\Delta, \ldots, b_x\Delta\}^2$,

$$\begin{aligned}\left|\mathrm{Cov}\left(Z\left(\frac{j_1}{\Delta}\right), Z\left(\frac{j_2}{\Delta}\right)\right) - K(g_n)\left(\frac{j_1}{\Delta}, \frac{j_2}{\Delta}\right)\right| &\leqslant 4\kappa_7(1+x)^2(y+\log n)^2 n^{-u_3} \\ &\quad + \kappa(1+x)^2(y+\log n)^2 n^{-u_5} \\ &\leqslant \kappa(1+x)^2(y+\log n)^2 n^{-u_5} \end{aligned} \tag{5.48}$$

by setting $\kappa = \kappa + 4\kappa_7$ and since $u_5 \leqslant u_3$. Claim 5.4.5.1 follows by setting $y = 2\log n$. It remains to upper bound g_n on $[a_x; b_x]$ in order to check equation 3 of Theorem 5.3.8. This is the subject of the following lemma.

Lemma 5.4.9 *For any $\alpha \in \mathbb{R}$,*

$$|g_n(\alpha)| \leqslant 20\|s\|_\infty f_n(\alpha) + 12\|s\|_\infty \leqslant \max\left(40\|s\|_\infty f_n(\alpha), 24\|s\|_\infty\right).$$

In particular, for all $x > 0$, $\max(|g_n(a_x)|, |g_n(b_x)|) \leqslant 20\|s\|_\infty (1+x)$.

Proof Since $\|s\|_\infty \geqslant \|s\|^2 \geqslant 1$ and $n_v\mathfrak{e} \leqslant 1$ by lemma 5.3.4, point 5 of Theorem 5.3.8 which we already proved implies that for any $\alpha \in \mathbb{R}$,

$$|g_n(\alpha))| \leqslant 8\|s\|_\infty f_n(\alpha) + 12\|s\|_\infty |\alpha|.$$

If $|\alpha| < 1$, then

$$|g_n(\alpha))| \leqslant 8\|s\|_\infty |f_n(\alpha)| + 12\|s\|_\infty \leqslant \max\left(16\|s\|_\infty f_n(\alpha), 24\|s\|_\infty\right),$$

else $|f_n(\alpha) - f_n(1)| \geqslant |\alpha| - 1$, therefore $|\alpha| \leqslant f_n(\alpha) + 1$, which yields

$$|g_n(\alpha))| \leqslant 20\|s\|_\infty f_n(\alpha) + 12\|s\|_\infty \leqslant \max\left(40\|s\|_\infty f_n(\alpha), 24\|s\|_\infty\right).$$

■

5.4.5 Construction of a Wiener process W such that $W \circ g_n$ approximates Z

Let E be the event of probability greater than $1-\frac{1}{n^2}$ on which the equations of claim 5.4.5.1 are satisfied. Let $x > 0$. Given $D_n^T \in E$, Z^1 is a piecewise linear gaussian process on the partition $([\frac{j}{\Delta}; \frac{j+1}{\Delta}[)_{a_x\Delta \leqslant j \leqslant b_x\Delta}$, such that for any $j \in \{a_x\Delta, \dots, b_x\Delta\}$,

$$\max_{(j_1,j_2)\in\{0,\dots,b_x\Delta\}^2} \left|\mathrm{Cov}\left(Z\left(\tfrac{j_1}{\Delta}\right), Z\left(\tfrac{j_2}{\Delta}\right)\right) - K(g_n)\left(\tfrac{j_1}{\Delta}, \tfrac{j_1}{\Delta}\right)\right| \leqslant \kappa_6(1+x)^2\log^2(n)n^{-u_5}, \tag{5.49}$$

where $K(g_n)$ is given by definition 5.4.6. Since g_n is non-decreasing, $K(g_n)(s,t) = \mathrm{Cov}(W_{g_n(s)}, W_{g_n(t)})$ for any two-sided Wiener process W on $\mathbb{R}$ such that $W_0 = 0$. In particular, $K(g_n)$ is a positive-definite function. Furthermore, by definition, $\forall(\alpha_1,\alpha_2) \in [a_x; b_x]^2$,

$$K(g_n)(\alpha_1,\alpha_1) + K(g_n)(\alpha_2,\alpha_2) - 2K(g_n)(\alpha_1,\alpha_2) = |g_n(\alpha_2) - g_n(\alpha_1)|.$$

Moreover, for all $j \in \{a_x\Delta, \dots, b_x\Delta - 1\}$, since $n_v\mathfrak{e} \leqslant 1$,

$$\begin{aligned}
|g_n(\tfrac{j+1}{\Delta}) - g_n(\tfrac{j}{\Delta})| &\leqslant 8\|s\|_\infty |f_n(\tfrac{j+1}{\Delta}) - f_n(\tfrac{j}{\Delta})| + \frac{12\|s\|_\infty}{\Delta} \text{ by equations (5.44) and (5.42)} \\
&\leqslant 8\kappa_3\|s\|_\infty (1+x)^2 n^{-u_2} + 12\|s\|_\infty \frac{n-n_t}{n_t} \text{ by claim 5.5.5.1} \\
&\leqslant 8\kappa_3\|s\|_\infty (1+x)^2 n^{-u_2} + 12\|s\|_\infty n^{-\delta_3} \text{ by hypothesis } H4.
\end{aligned}$$

Finally, by lemma 5.4.9 and since g_n is non-decreasing,

$$\sup_{\alpha\in[a_x;b_x]} K(g_n)(\alpha,\alpha) \leqslant \max(|g_n(a_x)|, |g_n(b_x)|) \leqslant 20\|s\|_\infty (1+x).$$

In this situation, proposition 5.5.9 in the appendix (applied to $Y = Z^1$, $K_X = K(g_n)$ with $h = g_n$) guarantees the existence of a continuous gaussian process $Z^2(D_n^T)$, with variance-covariance function $K(g_n)$ and such that for some constant κ and for $u = \min(u_5, u_2, \delta_3)$,

$$\forall D_n^T \in E, \mathbb{E}\left[\sup_{a_x\leqslant t\leqslant b_x} |Z^1(t) - Z^2(t)||D_n^T\right] \leqslant \kappa(1+x)^{\frac{7}{6}}\log^{\frac{2}{3}}(n) \times n^{-\frac{u}{12}}. \tag{5.50}$$

Since the conditional distribution of $Z^2(D_n^T)$ given D_n^T is entirely determined by the function g_n which does not depend on D_n^T, Z^2 is independent from D_n^T. In particular, Z^2 can be naturally extended to $D_n^T \notin E$. Moreover, since g_n increases, $W = Z^2 \circ g_n^{-1}$ is a continuous, centered gaussian process with covariance function

$$\mathrm{Cov}(Z_s, Z_t) = K(g_n)(g_n^{-1}(s), g_n^{-1}(t)) = \begin{cases} s \wedge t \text{ if } 0 \leqslant s,t \\ -(s \vee t) \text{ if } s,t \leqslant 0 \\ 0 \text{ else }, \end{cases} \tag{5.51}$$

it is therefore a two-sided Wiener process on $[g_n(a_x); g_n(b_x)]$ taking value 0 at 0. W can be extended to $\mathbb{R}$ by placing independent Wiener processes W_g, W_d on its left and on its right, by the equations $W(u) = W(g_n(a_x)) + W_g(u) - W_g(g_n(a_x))$ for $u < a_x$, $W(u) = W(g_n(b_x)) + W_d(u) - W_d(g_n(b_x))$ for $u > b_x$. Thus, by claim 5.4.3.2 and equation (5.50), with probability greater than $1 - \frac{2}{n^2}$,

$$\begin{aligned}
\mathbb{E}\left[\sup_{a_x \leqslant t \leqslant b_x} |Z(t) - W_{g_n(t)}||D_n^T\right] &= \mathbb{E}\left[\sup_{a_x \leqslant t \leqslant b_x} |Z(t) - Z^2(t)||D_n^T\right] \\
&\leqslant \mathbb{E}\left[\sup_{a_x \leqslant t \leqslant b_x} |Z^1(t) - Z(t)||D_n^T\right] + \mathbb{E}\left[\sup_{a_x \leqslant t \leqslant b_x} |Z^1(t) - W_{g_n(t)}||D_n^T\right] \\
&\leqslant \kappa(1+x)^{\frac{7}{6}} \log^{\frac{2}{3}}(n) n^{-\frac{u_2}{12}} + \kappa_5(c_1)(1 + 2\log n)(1+x)^{\frac{3}{2}} n^{-\frac{\delta_4}{3}} \\
&\leqslant \kappa(1+x)^{\frac{3}{2}} n^{-u},
\end{aligned}$$

for all $u < \min\left(\frac{u_5}{12}, \frac{u_2}{12}, \frac{\delta_3}{12}, \frac{\delta_4}{3}\right)$ and a constant $\kappa(u)$. Finally, by claim 5.4.3.1, with probability greater than $1 - \frac{3}{n^2}$,

$$\begin{aligned}
\sup_{\alpha \in [a_x; b_x]} \left|\hat{R}^{ho}(\alpha) - [f_n(\alpha) - W_{g_n(\alpha)}]\right| &\leqslant \sup_{\alpha \in [a_x; b_x]} |L(\alpha) - f_n(\alpha)| + \sup_{\alpha \in [a_x; b_x]} |Z(\alpha) - W_{g_n(\alpha)}| \\
&\leqslant \kappa_1(1+x)[\log^2(n) + \log^2(2+x)] n^{-\min(\frac{1}{12}, \frac{\delta_3}{2})} \\
&\quad + \kappa(1+x)^{\frac{3}{2}} n^{-u} \\
&\leqslant \kappa(1+x)^{\frac{3}{2}} n^{-u_1},
\end{aligned}$$

for all $u_1 < \min\left(\frac{u_5}{12}, \frac{u_2}{12}, \frac{\delta_3}{12}, \frac{\delta_4}{3}\right)$ and a constant κ. This proves Theorem 5.3.8.

5.5 Appendix

Lemma 5.5.1 *Let X be a random variable belonging to $[-1; 1]$, with pdf s. For all $j \in \mathbb{N}$, let $\theta_j = \langle s, \psi_j \rangle$. Then*

$$Var\left(\psi_j(X)\right) \underset{j \to +\infty}{\longrightarrow} 1$$

$$\forall k_0 \leqslant k, \sum_{j=k_0}^{k} |\, Var(\psi_j) - 1| \leqslant \|\theta\|_{\ell^1} = \sum_{j=0}^{+\infty} |\langle s, \psi_j \rangle|.$$

Proof $\mathbb{E}[\psi_j(X)] = \int_0^1 \psi_j(x) s(x) dx = \theta_j$. Moreover, $\psi_j(X)^2 = 2\cos^2(2\pi j X) = 1 + \cos(2\pi j X)$, therefore $\mathrm{Var}(\cos(\pi j X)) = 1 + \frac{\theta_j}{\sqrt{2}} - \theta_j^2$, therefore since $|\theta_j| \leqslant \sqrt{2}$, $|\mathrm{Var}(\cos(jX)) - 1| \leqslant \left|\sqrt{2} - \frac{1}{\sqrt{2}}\right| |\theta_j| \leqslant |\theta_j|$. ■

Lemma 5.5.2 *Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a function, $g, h : \mathbb{R}_+ \to \mathbb{R}$ be two non-increasing functions. Then*

$$\inf_{x\in\mathbb{R}_+} \{f(x)+g(x)+h(x)\} \leqslant \inf_{x\in\mathbb{R}_+} \{f(x)+g(x)\} + \inf_{x\in\mathbb{R}_+} \{f(x)+h(x)\}.$$

Proof Let $\delta > 0$. Let x_g be such that $f(x_g)+g(x_g) \leqslant \delta + \inf_{x\in\mathbb{R}_+}\{f(x)+g(x)\}$. Let x_h be such that $f(x_h)+h(x_h) \leqslant \inf_{x\in\mathbb{R}_+}\{f(x)+h(x)\}+\delta$. Let $x_* = \max(x_g, x_h)$. If $x_* = x_g$, then

$$\begin{aligned}
f(x_*)+g(x_*)+h(x_*) &\leqslant \inf_{x\in\mathbb{R}_+} \{f(x)+g(x)\} + \delta + h(x_*) \\
&\leqslant \inf_{x\in\mathbb{R}_+} \{f(x)+g(x)\} + \delta + h(x_h) \text{ by monotony of } h \\
&\leqslant \inf_{x\in\mathbb{R}_+} \{f(x)+g(x)\} + \delta + f(x_h) + h(x_h) \\
&\leqslant \inf_{x\in\mathbb{R}_+} \{f(x)+g(x)\} + \inf_{x\in\mathbb{R}_+} \{f(x)+h(x)\} + 2\delta
\end{aligned}$$

Symetrically, if $x_* = x_h$, then $f(x_*)+g(x_*)+h(x_*) \leqslant \inf_{x\in\mathbb{R}_+}\{f(x)+g(x)\} + \inf_{x\in\mathbb{R}_+}\{f(x)+h(x)\} + 2\delta$. As a result,

$$\begin{aligned}
\inf_{x\in\mathbb{R}_+} \{f(x)+g(x)+h(x)\} \leqslant f(x_*)+g(x_*)+h(x_*) \leqslant \inf_{x\in\mathbb{R}_+} \{f(x)+g(x)\} \\
+ \inf_{x\in\mathbb{R}_+} \{f(x)+h(x)\} + 2\delta.
\end{aligned}$$

Since no assumptions were made about $\delta > 0$, lemma 5.5.2 is proved. ■

Proposition 5.5.3 *With the hypotheses and notations above, for any integers $k_0 \leqslant k$, with probability greater than $1-e^{-y}$:*

$$\left| \sum_{j=k_0+1}^{k} (\hat{\theta}_j^T - \theta_j)^2 - \frac{|k-k_0|}{n_t} \right| \leqslant \frac{\|\theta\|_{\ell^1}}{n_t} + C\sqrt{y+\log n}\left[\frac{\sqrt{|k-k_0|}}{n_t} + \frac{|k-k_0|}{n_t^{\frac{5}{4}}}\right].$$

In particular, there exists a constant $\kappa_1 = \kappa_1(\|s\|_\infty, c_1, \|\theta\|_{\ell^1})$ such that for any α_1, α_2 such that $(\alpha_1\Delta, \alpha_2\Delta) \in \mathbb{N}^2$ and $\alpha_1 < \alpha_2$, with probability greater than $1-e^{-y}$,

$$\left| \sum_{j=k_*+\alpha_1\Delta}^{k_*+\alpha_2\Delta} (\hat{\theta}_j^T - \theta_j)^2 - \frac{|k-k_0|}{n_t} \right| \leqslant \kappa_1(\alpha_2-\alpha_1)[\log n + y]^2 \times n^{-\min(\frac{1}{12}, \frac{\delta_3}{2})}\mathfrak{e}(n). \quad (5.52)$$

Proof Let $(k_0, k) \in \mathbb{N}^2$ be such that $k_0 < k$. The proof rests on lemma 14 of Arlot and Lerasle [5] applied to $S_m = \langle \psi_{k_0+1}, \dots, \psi_k \rangle$. Let us compute $b_m = \sup_{u \in \mathbb{R}^{|k-k_0|} : \|u\| \leqslant 1} \sum_{j=k_0}^{k} u_j \psi_j(x) \leqslant \sup_x \sqrt{\sum_{j=k_0}^{k} \psi_j^2(x)} \leqslant \sqrt{|k-k_0|}$ and

$$\mathcal{D}_k = \sum_{j=k_0+1}^{k} \mathrm{Var}\left(\psi_j(X)\right) = |k-k_0| \pm \frac{\|\theta\|_{\ell^1}}{n_t}$$

(by lemma 5.5.1). Furthermore, $\mathcal{D}_k \leqslant \sqrt{2}|k-k_0|$ since $\psi_j = \sqrt{2}\cos(2\pi j \cdot) : [0;1] \to [-\sqrt{2};\sqrt{2}]$. By [5, lemma 14], with probability greater than $1-e^{-y}$, for any $\varepsilon > 0$,

$$\Big| \sum_{j=k_0+1}^{k} (\hat{\theta}_j^T - \theta_j)^2 - \frac{\mathcal{D}_k}{n_t} \Big| \leqslant \varepsilon \frac{\mathcal{D}_k}{n_t} + \kappa \left(\frac{\|s\|_\infty [\log n + y]}{(\varepsilon \wedge 1) n_t} + \frac{|k-k_0| [\log n + y]^2}{(\varepsilon \wedge 1)^3 n_t^2} \right).$$

Let $\varepsilon_1 = \sqrt{\frac{\|s\|_\infty (\log n + y)}{|k-k_0|}} \wedge 1$. If $\varepsilon_1 = 1$, then $|k-k_0| \leqslant \|s\|_\infty (y + \log n)$ therefore $\varepsilon_1 \frac{|k-k_0|}{n_t} + \kappa \frac{\|s\|_\infty [\log n + y]}{(\varepsilon_1 \wedge 1) n_t} \leqslant (1+\kappa) \frac{\|s\|_\infty (y + \log n)}{n_t}$. If $\varepsilon_1 < 1$, then $\varepsilon_1 \frac{|k-k_0|}{n_t} + \kappa \frac{\|s\|_\infty [\log n + y]}{(\varepsilon_1 \wedge 1) n_t} = (1+\kappa) \sqrt{\|s\|_\infty (y + \log n)} \frac{\sqrt{|k-k_0|}}{n_t}$. In all cases, if $k > k_0$,

$$\varepsilon_1 \frac{|k-k_0|}{n_t} + \kappa \frac{\|s\|_\infty [\log n + y]}{(\varepsilon_1 \wedge 1) n_t} \leqslant (1+\kappa) \|s\|_\infty (y + \log n) \frac{\sqrt{|k-k_0|}}{n_t}. \tag{5.53}$$

Let $\varepsilon_2 = \frac{\sqrt{\log n + y}}{n_t^{\frac{1}{4}}} \wedge 1$. If $\frac{\sqrt{y + \log n}}{n_t^{\frac{1}{4}}} \geqslant 1 = \varepsilon_2$, then $\varepsilon_2 \frac{|k-k_0|}{n_t} + \kappa \frac{|k-k_0| [\log n + y]^2}{(\varepsilon_2 \wedge 1)^3 n_t^2} \leqslant \sqrt{y + \log n} \frac{|k-k_0|}{n_t^{\frac{5}{4}}} + \kappa \frac{|k-k_0| (y + \log n)^2}{n_t^2} \leqslant (1+\kappa)(y + \log n)^2 \frac{|k-k_0|}{n_t^{\frac{5}{4}}}$. If $\varepsilon_2 = \frac{\sqrt{y + \log n}}{n_t^{\frac{1}{4}}} < 1$, then

$$\begin{aligned} \varepsilon_2 \frac{|k-k_0|}{n_t} + \kappa \frac{|k-k_0| [\log n + y]^2}{(\varepsilon_2 \wedge 1)^3 n_t^2} &= \sqrt{y + \log n} \frac{|k-k_0|}{n_t^{\frac{5}{4}}} + \kappa (y + \log n)^2 \frac{|k-k_0|}{n_t^2} \frac{n_t^{\frac{3}{4}}}{(y + \log n)^{\frac{3}{2}}} \\ &\leqslant (1+\kappa) \sqrt{y + \log n} \frac{|k-k_0|}{n_t^{\frac{5}{4}}}. \end{aligned}$$

In all cases,

$$\varepsilon_2 \frac{|k-k_0|}{n_t} + \kappa \frac{|k-k_0| [\log n + y]^2}{(\varepsilon_2 \wedge 1)^3 n_t^2} \leqslant (1+\kappa)(y + \log n)^2 \frac{|k-k_0|}{n_t^{\frac{5}{4}}}. \tag{5.54}$$

By lemma 5.5.2,

$$\begin{aligned}\Big|\sum_{j=k_0+1}^{k}(\hat{\theta}_j^T-\theta_j)^2-\frac{\mathcal{D}_k}{n_t}\Big| &\leqslant \inf_{\varepsilon\geqslant 0}\left\{\varepsilon\frac{\mathcal{D}_k}{n_t}+\kappa\frac{\|s\|_\infty[\log n+y]}{(\varepsilon\wedge 1)n_t}\right\}\\ &\quad+\inf_{\varepsilon\geqslant 0}\left\{\varepsilon\frac{\mathcal{D}_k}{n_t}+\kappa\frac{|k-k_0|[\log n+y]^2}{(\varepsilon\wedge 1)^3 n_t^2}\right\}\\ &\leqslant \varepsilon_1\frac{|k-k_0|}{n_t}+\kappa\frac{\|s\|_\infty[\log n+y]}{(\varepsilon_1\wedge 1)n_t}+\varepsilon_2\frac{|k-k_0|}{n_t}\\ &\quad+\kappa\frac{|k-k_0|[\log n+y]^2}{(\varepsilon_2\wedge 1)^3 n_t^2}+(\varepsilon_1+\varepsilon_2)\frac{\|\theta\|_{\ell^1}}{n_t}\\ &\leqslant (1+\kappa)\|s\|_\infty(y+\log n)\frac{\sqrt{|k-k_0|}}{n_t}\\ &\quad+(1+\kappa)(y+\log n)^2\frac{|k-k_0|}{n_t^{\frac{5}{4}}}+\frac{2\|\theta\|_{\ell^1}}{n_t},\end{aligned}$$

by equations (5.53), (5.54). In conclusion, on an event E_y of probability greater than $1-e^{-y}$,

$$\begin{aligned}\Big|\sum_{j=k_0+1}^{k}(\hat{\theta}_j^T-\theta_j)^2-\frac{|k-k_0|}{n_t}\Big| &\leqslant \Big|\sum_{j=k_0+1}^{k}(\hat{\theta}_j^T-\theta_j)^2-\frac{\mathcal{D}_k}{n_t}\Big|+\frac{\|\theta\|_{\ell^1}}{n_t}\\ &\leqslant \frac{3\|\theta\|_{\ell^1}}{n_t}+(1+\kappa)\|s\|_\infty(y+\log n)\\ &\quad\times\left[\frac{\sqrt{|k-k_0|}}{n_t}+(y+\log n)\frac{|k-k_0|}{n_t^{\frac{5}{4}}}\right].\end{aligned}\tag{5.55}$$

If $k_0=k_*+\alpha_1\Delta$ and $k=k_*+\alpha_2\Delta$, then by hypothesis H4 of Theorem 5.3.8,

$$\frac{\sqrt{|k-k_0|}}{n_t}=\sqrt{\alpha_2-\alpha_1}\sqrt{\frac{\Delta}{n_v n_t}}\sqrt{\frac{n_v}{n_t}}=\sqrt{\alpha_2-\alpha_1}\sqrt{\frac{n-n_t}{n_t}}\mathfrak{e}\leqslant\sqrt{\alpha_2-\alpha_1}n^{-\frac{\delta_3}{2}}\mathfrak{e}.\tag{5.56}$$

Furthermore,

$$\begin{aligned}\frac{|k-k_0|}{n_t^{\frac{5}{4}}}&=(\alpha_2-\alpha_1)\frac{\mathcal{E}}{n_t^{\frac{1}{4}}}\\ &=(\alpha_2-\alpha_1)\sqrt{\frac{\mathcal{E}}{n_v}}\frac{\sqrt{n_v\mathcal{E}}}{n_t^{\frac{1}{4}}}\\ &\leqslant(\alpha_2-\alpha_1)\mathfrak{e}\frac{\sqrt{2n_v\mathrm{or}(n_t)+1}}{n_t^{\frac{1}{4}}}.\end{aligned}$$

Let $k_1 = \lceil n_t^{\frac{1}{3+\delta_1}} \rceil$, so that $n_t^{\frac{1}{3+\delta_1}} \leqslant k_1 \leqslant 2n_t^{\frac{1}{3+\delta_1}}$. By hypothesis H1 of Theorem 5.3.8, $\sum_{j=k+1}^{+\infty} \theta_j^2 \leqslant \frac{c_1}{k^{2+\delta_1}}$ therefore

$$\mathrm{or}(n_t) \leqslant \inf_{k\in\mathbb{N}^*} \frac{c_1}{k^{2+\delta_1}} + \frac{k}{n_t} \leqslant \frac{c_1}{k_1^{2+\delta_1}} + \frac{k_1}{n_t} \leqslant \frac{c_1}{n_t^{\frac{2}{3+\delta_1}}} + \frac{2n_t^{\frac{1}{3+\delta_1}}}{n_t} \leqslant \frac{2+c_1}{n_t^{\frac{2}{3+\delta_1}}}.$$

Thus $1 + 2n_v\mathrm{or}(n_t) \leqslant 1 + 2n_t\mathrm{or}(n_t) \leqslant (5+2c_1)n_t^{\frac{1+\delta_1}{3+\delta_1}}$, hence

$$\begin{aligned}
\frac{|k-k_0|}{n_t^{\frac{5}{4}}} &\leqslant (\alpha_2-\alpha_1)\mathfrak{e}\sqrt{5+2c_1}\frac{n_t^{\frac{1+\delta_1}{6+2\delta_1}}}{n_t^{\frac{1}{4}}} \\
&\leqslant (\alpha_2-\alpha_1)\sqrt{5+2c_1}n_t^{-\frac{1}{12}}\mathfrak{e} \\
&\leqslant (\alpha_2-\alpha_1)\sqrt{5+2c_1}\frac{2^{\frac{1}{12}}}{n^{\frac{1}{12}}}\mathfrak{e}. \qquad (5.57)
\end{aligned}$$

Finally, $\frac{\|\theta\|_{\ell^1}}{n_t} = \frac{n_v}{n_t}\frac{\|\theta\|_{\ell^1}}{n_v} \leqslant \|\theta\|_{\ell^1}\frac{n-n_t}{n_t}\mathfrak{e} \leqslant \|\theta\|_{\ell^1} n^{-\delta_3}$. Equation (5.52) follows from equations (5.55), (5.56) and (5.57).

■

Lemma 5.5.4 *Let $(c_{i,j})_{(i,j)\in\mathbb{N}^2}$ be real coefficients. Let $I_1, I_2 \subset \mathbb{N}$ be two finite sets. Let $(\theta_j)_{j\in\mathbb{N}}$ be a sequence. Let $C = \max\left\{\sup_{i\in I_1}\sum_{j\in I_2}|c_{i,j}|, \sup_{i\in I_2}\sum_{j\in I_1}|c_{i,j}|\right\}$. Then*

$$\sum_{i\in I_1}\left(\sum_{j\in I_2} c_{i,j}\theta_j\right)^2 \leqslant C^2\sum_{j\in I_2}\theta_j^2$$

and

$$\left|\sum_{i\in I_1}\sum_{j\in I_2}\theta_i\theta_j c_{i,j}\right| \leqslant C\max\left\{\sum_{i\in I_1}\theta_i^2, \sum_{j\in I_2}\theta_j^2\right\}.$$

Proof Let $C_i = \sum_{j\in I_2} |c_{i,j}|$. Then

$$\begin{aligned}\sum_{i\in I_1}\left(\sum_{j\in I_2} c_{i,j}\theta_j\right)^2 &= \sum_{i\in I_1} C_i^2\left(\frac{1}{C_i}\sum_{j\in I_2}\operatorname{sgn}(c_{i,j})|c_{i,j}|\theta_j\right)^2\\ &\leqslant \sum_{i\in I_1}\frac{C_i^2}{C_i}\sum_{j\in I_2}|c_{i,j}|\theta_j^2 \text{ by Jensen's inequality}\\ &\leqslant \left(\max_{i\in I_1} C_i\right)\sum_{j\in I_2}\theta_j^2\sum_{i\in I_1}|c_{i,j}|\\ &\leqslant C^2\sum_{j\in I_2}\theta_j^2.\end{aligned}$$

This proves the first equation. Furthermore,

$$\begin{aligned}\left|\sum_{i\in I_1}\sum_{j\in I_2}\theta_i\theta_j c_{i,j}\right| &\leqslant \sum_{i\in I_1}\sum_{j\in I_2}\frac{\theta_i^2+\theta_j^2}{2}|c_{i,j}|\\ &= \frac{1}{2}\sum_{i\in I_1}\theta_i^2\sum_{j\in I_2}|c_{i,j}| + \frac{1}{2}\sum_{j\in I_2}\theta_j^2\sum_{i\in I_1}|c_{i,j}|\\ &\leqslant C\max\left\{\sum_{i\in I_1}\theta_i^2, \sum_{j\in I_2}\theta_j^2\right\},\end{aligned}$$

which proves the second equation. ■

Lemma 5.5.5 *Under the assumptions of Theorem 5.3.8, there exists a constant* $\kappa(c_1, c_2) > 0$ *such that for any* $x \geqslant 0$,

$$k_* + a_x\Delta \geqslant \frac{\kappa}{(1+x)^{\frac{1}{\rho_1}}} n_t^{\frac{2}{3\rho_1}}.$$

Proof By hypothesis H2 of Theorem 5.3.8,

$$\begin{aligned}c_2(k_* + a_x\Delta)^{-\rho_1} &\leqslant \sum_{j=k_*+a_x\Delta+1}^{+\infty}\theta_j^2\\ &\leqslant \sum_{j=k_*+a_x\Delta+1}^{k_*}\left[\theta_j^2 - \frac{1}{n_t}\right] + |a_x|\mathcal{E} + \sum_{j=k_*+1}^{+\infty}\theta_j^2\\ &\leqslant \mathfrak{e}f_n(a_x) + |a_x|\mathcal{E} + \operatorname{or}(n_t). \qquad (5.58)\end{aligned}$$

By definition, $f_n(a_x) \leqslant x$ and by lemma 5.4.1, $|a_x| \leqslant 2(1+x)$. Furthermore, by lemma 5.3.4, $\mathfrak{e} \leqslant \mathcal{E} \leqslant 2\mathrm{or}(n_t) + \frac{1}{n_v}$. Since by hypothesis H5 of Theorem 5.3.8, $n_v \geqslant n^{\frac{2}{3}+\delta_4}$, it follows that: $\mathcal{E} \leqslant 2\mathrm{or}(n_t) + \frac{1}{n^{\frac{2}{3}+\delta_4}}$. Equation (5.58) thus yields

$$c_2(k_* + a_x\Delta)^{-\rho_1} \leqslant 6(1+x)\left[\mathrm{or}(n_t) + \frac{1}{n^{\frac{2}{3}}}\right].$$

On the other hand, by hypothesis H1 of Theorem 5.3.8,

$$\mathrm{or}(n_t) \leqslant \min_{k\in\mathbb{N}} \frac{c_1}{k^2} + \frac{k}{2n_t} \leqslant \frac{3c_1^{\frac{1}{3}}}{n_t^{\frac{2}{3}}}.$$

It follows finally that, for some constant $\kappa(c_1, c_2)$,

$$k_* + a_x\Delta \geqslant \frac{\kappa}{(1+x)^{\frac{1}{\rho_1}}} n_t^{\frac{2}{3\rho_1}}.$$

■

Claim 5.5.5.1 *Let $u_2 = \min\left(\frac{2\delta_2}{3\rho_1}, \delta_3\right)$. Let x be a non-negative real number. Let a_x, b_x be such that $a_x \leqslant 0 \leqslant b_x$ and $\max(f_n(a_x), f_n(b_x)) \leqslant x$. Assume also that $a_x\Delta - 1 \geqslant \frac{-k_*}{\Delta}$. There exists a constant $\kappa_3 \geqslant 0$ such that for all $j \in [a_x\Delta; b_x\Delta + 1]$,*

$$\left|f_n\left(\tfrac{j}{\Delta}\right) - f_n\left(\tfrac{j-1}{\Delta}\right)\right| \leqslant \kappa_3(1+x)^2 n^{-u_2} \tag{5.59}$$

$$\theta^2_{k_*+j} \leqslant \kappa_3(1+x)^2 n^{-u_2}\mathfrak{e}. \tag{5.60}$$

Proof By hypothesis H3 of Theorem 5.3.8, for all $k \geqslant 1$, $\theta^2_{k+k^{\delta_2}} \geqslant c_3\theta^2_{k-k^{\delta_2}}$. Thus, for all $k \geqslant 1$ and any $j \in [k - k^{\delta_2}; k + k^{\delta_2}]$,

$$\begin{aligned}\max\left(\theta_k^2, \frac{1}{n_t}\right) &\leqslant \max\left(\frac{\theta_j^2}{c_3}, \frac{1}{n_t}\right) \\ &\leqslant \frac{1+c_3}{c_3 n_t} + \frac{1}{c_3}\left|\theta_j^2 - \frac{1}{n_t}\right|. \end{aligned} \tag{5.61}$$

Let $k \in [k_* + a_x\Delta; k_* + b_x\Delta + 1]$. Assume without loss of generality (up to a change in the constant κ_3) that $x \geqslant 1$. Thus by lemma 5.3.7, $\max(-a_x, b_x) \geqslant 1$.

- If $|b_x| \geqslant 1$, then two cases can be distinguished.

– If $k \leqslant k_* + \frac{\Delta}{2}$, then $k + k^{\delta_2} \wedge \frac{\Delta}{2} \leqslant k_* + \Delta \leqslant k_* + b_x\Delta$, therefore by definition of a_x, b_x,

$$2x\mathfrak{e} \geqslant \mathfrak{e}[f_n(a_x) + f_n(b_x)] = \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \left|\theta_j^2 - \frac{1}{n_t}\right| \geqslant \sum_{j=k+1}^{k+k^{\delta_2}\wedge\frac{\Delta}{2}} \left|\theta_j^2 - \frac{1}{n_t}\right|.$$

– If $k_* + \frac{\Delta}{2} < k \leqslant k_* + b_x\Delta + 1$, then $k - k^{\delta_2} \wedge \frac{\Delta}{2} \geqslant k_*$, therefore

$$2x\mathfrak{e} \geqslant \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \left|\theta_j^2 - \frac{1}{n_t}\right| \geqslant \sum_{j=k-k^{\delta_2}\wedge\frac{\Delta}{2}}^{k-1} \left|\theta_j^2 - \frac{1}{n_t}\right|.$$

- If $|a_x| \geqslant 1$, then we likewise consider two possibilities.

– If $k > k_* - \frac{\Delta}{2}$, then $k - k^{\delta_2} \wedge \frac{\Delta}{2} > k_* - \Delta \geqslant k_* + a_x\Delta$, therefore by definition of a_x, b_x,

$$2x\mathfrak{e} \geqslant \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \left|\theta_j^2 - \frac{1}{n_t}\right| \geqslant \sum_{j=k-k^{\delta_2}\wedge\frac{\Delta}{2}}^{k-1} \left|\theta_j^2 - \frac{1}{n_t}\right|.$$

– If $k \leqslant k_* - \frac{\Delta}{2}$, then $k + k^{\delta_2} \wedge \frac{\Delta}{2} \leqslant k_*$, therefore

$$2x\mathfrak{e} \geqslant \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \left|\theta_j^2 - \frac{1}{n_t}\right| \geqslant \sum_{j=k+1}^{k+k^{\delta_2}\wedge\frac{\Delta}{2}} \left|\theta_j^2 - \frac{1}{n_t}\right|.$$

In all cases, by equation (5.61),

$$\left(k^{\delta_2} \wedge \frac{\Delta}{2}\right) \max\left(\theta_k^2, \frac{1}{n_t}\right) \leqslant k^{\delta_2} \wedge \frac{\Delta}{2} \frac{1+c_3}{c_3 n_t} + \frac{2x}{c_3}\mathfrak{e},$$

in other words

$$\max\left(\theta_k^2, \frac{1}{n_t}\right) \leqslant \frac{1+c_3}{c_3 n_t} + \frac{2x}{c_3} \frac{\mathfrak{e}}{k^{\delta_2} \wedge \frac{\Delta}{2}}.$$

Furthermore, by hypothesis H4 of Theorem 5.3.8, $\Delta \geqslant \frac{n_t}{n-n_t} \geqslant n^{\delta_3}$, and by lemma 5.5.5,

$$k^{\delta_2} \geqslant (k_* + a_x\Delta)^{\delta_2} \geqslant \frac{\kappa}{(1+x)^{\frac{\delta_2}{\rho_1}}} n_t^{\frac{2\delta_2}{3\rho_1}}.$$

Let $u_2 = \min\left(\frac{2\delta_2}{3\rho_1}, \delta_3\right)$ Since $\delta_2 \leqslant \rho_1$, there exists therefore a constant κ such that

$$\max\left(\theta_k^2, \frac{1}{n_t}\right) \leqslant \kappa(1+x)^2 n^{-u_2}\mathfrak{e}.$$

In conclusion, for all $j \in \{a_x\Delta, \dots, b_x\Delta + 1\}$,

$$\begin{aligned}
\theta_{k_*+j}^2 &\leqslant \max\left(\theta_{k_*+j}^2, \frac{1}{n_t}\right) \leqslant \kappa(1+x)^2 n^{-u_2}\mathfrak{e} \\
\left|f_n\left(\tfrac{j}{\Delta}\right) - f_n\left(\tfrac{j-1}{\Delta}\right)\right| &= \frac{1}{\mathfrak{e}}|\theta_{j+k_*}^2 - \frac{1}{n_t}| \\
&\leqslant \frac{1}{\mathfrak{e}}\max\left(\theta_{k_*+j}^2, \frac{1}{n_t}\right) \\
&\leqslant \kappa(1+x)^2 n^{-u_2}.
\end{aligned}$$

This proves claim 5.5.5.1. ■

Proposition 5.5.6 *Let P be the probability measure with pdf s on $[0;1]$. Let $\theta_j = \langle s, \psi_j\rangle = P(\psi_j)$ and $\theta_j^2 = \theta_j^2$, and assume that they satisfy the hypotheses of Theorem 5.3.8. Let $\hat{\theta}_j^T = P^T\psi_j$. Let $I_k^1, I_k^2 \subset \{k_* + a_x\Delta + 1, \dots, k_* + b_x\Delta\}$ be two intervals. Then the statistics*

$$U_{I_k^1, I_k^2} = \sum_{i\in I_k^1}\sum_{j\in I_k^2} \hat{\theta}_i^T\hat{\theta}_j^T[P(\psi_i\psi_j) - P\psi_i P\psi_j]$$

can be approximated in the following way. There exists two constants κ_4 and $u_3 > 0$ such that, with probability greater than $1 - e^{-y}$,

$$\begin{aligned}
U_{I_k^1, I_k^2} = \frac{1}{2}\frac{|I_k^1 \cap I_k^2|}{n_t} + \left(1 - \frac{1}{\sqrt{2}}\right)\sum_{i\in I_k^1\cap I_k^2}\theta_i^2 + \frac{1}{\sqrt{2}}\sum_{i\in I_k^1}\sum_{j\in I_k^2}\theta_i\theta_j\theta_{|i-j|} + \frac{1}{2n_t}\sum_{i\in I_k^1}\sum_{j\in I_k^2}\theta_{|i-j|}^2 \\
\pm \kappa_4(y + \log n)^2(1+x)n^{-u_3}\mathcal{E}.
\end{aligned}$$

Proof First, by lemma 5.4.1,

$$\max\left(\sum_{i\in I_k^1}\theta_i^2, \sum_{j\in I_k^2}\theta_j^2\right) \leqslant \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta}\theta_j^2 \leqslant 4(1+x)\mathcal{E}. \tag{5.62}$$

Let $c_{i,j} = \frac{\theta_{i+j}}{\sqrt{2}} + \left(\frac{1-\delta_{i,j}}{\sqrt{2}} + \delta_{i,j}\right)\theta_{|i-j|} - \theta_i\theta_j$. $U_{I_k^1, I_k^2}$ can be expressed as the sum of 4

terms: $U_{I_k^1,I_k^2} = V_1 + V_2 + V_3 + V_4 + V_5 + V_6$, where

$$V_1 = \sum_{i\in I_k^1}\sum_{j\in I_k^2}\theta_i\theta_j\left[\frac{\theta_{i+j}}{\sqrt{2}} + \big(\frac{1-\delta_{i,j}}{\sqrt{2}} + \delta_{i,j}\big)\theta_{|i-j|} - \theta_i\theta_j\right]$$

$$V_2 = (P^T - P)\sum_{i\in I_k^1}\psi_i\sum_{j\in I_k^2}\theta_j c_{i,j}$$

$$V_3 = (P^T - P)\sum_{j\in I_k^2}\psi_j\sum_{i\in I_k^1}\theta_i c_{i,j}$$

$$V_4 = \frac{1}{\sqrt{2}}\sum_{i\in I_k^1}\sum_{j\in I_k^2}(P^T - P)\psi_i(P^T - P)\psi_j\theta_{|i-j|}$$

$$V_5 = \left(1 - \frac{1}{\sqrt{2}}\right)\sum_{j\in I_k^1\cap I_k^2}\left(\hat{\theta}_j^T - \theta_j\right)^2$$

$$V_6 = \sum_{i\in I_k^1}\sum_{j\in I_k^2}(P^T - P)\psi_i(P^T - P)\psi_j\left[\frac{\theta_{i+j}}{\sqrt{2}} - \theta_i\theta_j\right]$$

The first term is

$$V_1 = \left(1 - \frac{1}{\sqrt{2}}\right)\sum_{i\in I_k^1\cap I_k^2}\theta_i^2 + \frac{1}{\sqrt{2}}\sum_{i\in I_k^1}\sum_{j\in I_k^2}\theta_i\theta_j\theta_{|i-j|} + \sum_{i\in I_k^1}\sum_{j\in I_k^2}\theta_i\theta_j\left[\frac{\theta_{i+j}}{\sqrt{2}} - \theta_i\theta_j\right].$$

For all $i \in I_k^1$,

$$\sum_{j\in I_k^2}\frac{|\theta_{i+j}|}{\sqrt{2}} + |\theta_i||\theta_j| \leqslant 2\sum_{j\geqslant k_* + a_x\Delta + 1}|\theta_j|.$$

Furthermore, for all $k \geqslant 2$, by hypothesis H1 of Theorem 5.3.8,

$$\sum_{j\geqslant k}|\theta_j| \leqslant \sum_{j=k}^{+\infty}\sqrt{\sum_{i=j}^{+\infty}\theta_i^2} \leqslant \sum_{j=k}^{+\infty}\frac{c_1}{(j-1)^{1+\frac{\delta_1}{2}}} \leqslant \frac{2c_1}{\delta_1}(k-1)^{\frac{\delta_1}{2}}. \tag{5.63}$$

Since $k_* + a_x\Delta \geqslant \frac{\kappa}{(1+x)^{\frac{1}{\rho_1}}}n_t^{\frac{2}{3\rho_1}}$ by lemma 5.5.5, there is a constant $\kappa(c_1, c_2)$ such that

$$\sum_{j\in I_k^2}\frac{|\theta_{i+j}|}{\sqrt{2}} + |\theta_i||\theta_j| \leqslant \kappa\frac{(1+x)^{\frac{\delta_1}{2\rho_1}}}{n_t^{\frac{\delta_1}{3\rho_1}}}. \tag{5.64}$$

The same argument applies to $\sum_{i\in I_k^1}\frac{|\theta_{i+j}|}{\sqrt{2}} + |\theta_i||\theta_j|$. Thus, by lemma 5.5.4,

$$\sum_{i\in I_k^1}\sum_{j\in I_k^2}\theta_i\theta_j\left[\theta_{i+j} - \theta_i\theta_j\right] \leqslant 2\kappa\frac{(1+x)^{\frac{\delta_1}{2\rho_1}}}{n_t^{\frac{\delta_1}{3\rho_1}}}\left[\sum_{i\in I_k^1}\theta_i^2 + \sum_{j\in I_k^2}\theta_j^2\right].$$

By equation (5.62), it follows that for a certain constant $\kappa(c_1, c_2)$,

$$\sum_{i\in I_k^1}\sum_{j\in I_k^2}\theta_i\theta_j\left[\frac{\theta_{i+j}}{\sqrt{2}}-\theta_i\theta_j\right]\leqslant\kappa\frac{(1+x)^{1+\frac{\delta_1}{2\rho_1}}}{n_t^{\frac{\delta_1}{3\rho_1}}}\mathcal{E}.$$

Thus

$$V_1=\left(1-\frac{1}{\sqrt{2}}\right)\sum_{i\in I_k^1\cap I_k^2}\theta_i^2+\frac{1}{\sqrt{2}}\sum_{i\in I_k^1}\sum_{j\in I_k^2}\theta_i\theta_j\theta_{|i-j|}\pm\kappa\frac{(1+x)^{1+\frac{\delta_1}{2\rho_1}}}{n_t^{\frac{\delta_1}{3\rho_1}}}\mathcal{E}\qquad(5.65)$$

Bernstein's inequality applies to V_2 and V_3. By symmetry, let us only consider V_2. Its variance satisfies the following inequality.

$$\begin{aligned}\operatorname{Var}\left(\sum_{i\in I_k^1}\psi_i\sum_{j\in I_k^2}\theta_j c_{i,j}\right)&\leqslant\|s\|_\infty\left\|\sum_{i\in I_k^1}\psi_i\sum_{j\in I_k^2}\theta_j c_{i,j}\right\|^2\\&\leqslant\|s\|_\infty\sum_{i\in I_k^1}\left(\sum_{j\in I_k^2}\theta_j c_{i,j}\right)^2.\end{aligned}$$

Let us now apply lemma 5.5.4. For all $i\in I_k^1$,

$$\begin{aligned}\sum_{j\in I_k^2}|c_{i,j}|&\leqslant\frac{1}{\sqrt{2}}\sum_{j\in I_k^2}|\theta_{i+j}|+\frac{1}{\sqrt{2}}\sum_{j\in I_k^2}|\theta_{|i-j|}|+|\theta_i|\sum_{j\in I_k^2}|\theta_j|\\&\leqslant\left(\sqrt{2}+\sup_{i\in\mathbb{N}}|\theta_i|\right)\sum_{r\in\mathbb{N}}|\theta_r|\\&\leqslant 3\|\theta\|_{\ell^1}\end{aligned}$$

In the same way, for all $j\in I_k^2$, $\sum_{i\in I_k^1}|c_{i,j}|\leqslant 3\|\theta\|_{\ell^1}$, hence by lemma 5.5.4,

$$\begin{aligned}\operatorname{Var}\left(\sum_{i\in I_k^1}\psi_i\sum_{j\in I_k^2}\theta_j c_{i,j}\right)&\leqslant 3\|\theta\|_{\ell^1}\|s\|_\infty\sum_{j\in I_k^2}\theta_j^2\\&\leqslant 12\|\theta\|_{\ell^1}\|s\|_\infty(1+x)\mathcal{E}\text{ by equation (5.62).}\end{aligned}\qquad(5.66)$$

As for the upper bound on the uniform norm, it follows from lemma 5.5.4 and the

elementary upper bound $\|\psi_i\|_\infty \leqslant \sqrt{2}$ that

$$\begin{aligned}\sup_{x\in\mathbb{R}}\left|\sum_{i\in I_k^1}\psi_i(x)\sum_{j\in I_k^2}\theta_j c_{i,j}\right| &\leqslant \sqrt{\sum_{i\in I_k^1}\left(\sum_{j\in I_k^2}\theta_j c_{i,j}\right)^2}\sup_{x\in\mathbb{R}}\sqrt{\sum_{i\in I_k^1}\psi_i(x)^2}\\ &\leqslant 3\left\|\theta\right\|_{\ell^1}\sqrt{2|I_k^1|}\sqrt{\sum_{j\in I_k^2}\theta_j^2}\\ &\leqslant 3\left\|\theta\right\|_{\ell^1}\sqrt{2(b_x-a_x)\Delta}\sqrt{4(1+x)\mathcal{E}}\\ &\leqslant \kappa(1+x)\sqrt{\Delta\mathcal{E}} \text{ by lemma 5.4.1,}\end{aligned} \tag{5.67}$$

for some constant $\kappa = \kappa(\|\theta\|_{\ell^1})$. By Bernstein's inequality, there exists an event $E_2(y) \subset \mathbb{R}^{n_t}$ with probability $\mathbb{P}(D_n^T \in E_2(y)) \geqslant 1 - e^{-y}$ such that, for any $D_n^T \in E_2(y)$,

$$\begin{aligned}|V_2| &\leqslant \sqrt{\frac{2y}{n_t}}\sqrt{\operatorname{Var}\left(\sum_{i\in I_k^1}\psi_i\sum_{j\in I_k^2}\theta_j c_{i,j}\right)} + \frac{y}{3n_t}\sup_{x\in\mathbb{R}}\left|\sum_{i\in I_k^1}\psi_i(x)\sum_{j\in I_k^2}\theta_j c_{i,j}\right|\\ &\leqslant \sqrt{24\left\|\theta\right\|_{\ell^1}\left\|s\right\|_\infty}\sqrt{\frac{y(1+x)\mathcal{E}}{n_t}} + \frac{\kappa y}{3n_t}(1+x)\sqrt{\Delta\mathcal{E}} \text{ by (5.66), (5.67).}\end{aligned}$$

Setting $\kappa = \max\left(\sqrt{24\left\|\theta\right\|_{\ell^1}\left\|s\right\|_\infty}, \frac{\kappa}{3}\right)$, it follows that on $E_2(y)$,

$$|V_2| \leqslant \kappa\sqrt{y(1+x)}\sqrt{\frac{n_v}{n_t}}\mathfrak{e} + \kappa y(1+x)\sqrt{\frac{n_v}{n_t}}\mathfrak{e}\sqrt{\mathcal{E}}.$$

$\mathcal{E}$ is uniformly bounded: $\mathcal{E} \leqslant \sum_{j=1}^{n_t}\theta_j^2 + \frac{1}{n_t} \leqslant 1 + \left\|s\right\|^2 \leqslant 1 + \left\|s\right\|_\infty$. Furthermore, by hypothesis H4 of Theorem 5.3.8, $\sqrt{\frac{n_v}{n_t}} = \sqrt{\frac{n-n_t}{n_t}} \leqslant n^{-\frac{\delta_3}{2}}$. Thus, there exists a constant $\kappa(\left\|\theta\right\|_{\ell^1}, \left\|s\right\|_\infty)$ such that, on $E_2(y)$,

$$|V_2| \leqslant \kappa y(1+x)n^{-\frac{\delta_3}{2}}\mathfrak{e}. \tag{5.68}$$

Symmetrically, there exists an event $E_3(y)$ of probability greater than $1-e^{-y}$, such that for any $D_n^T \in E_3(y)$,

$$|V_3| \leqslant \kappa y(1+x)n^{-\frac{\delta_3}{2}}\mathfrak{e}. \tag{5.69}$$

Now consider V_4. This term can be expressed as a finite sum of sums of squares:

$$V_4 = \frac{1}{\sqrt{2}} \sum_{r\in\mathbb{Z}} \sum_{i\in I_k^1\cap(I_k^2-r)} (P^T-P)\psi_i(P^T-P)\psi_{i+r}\theta_{|r|}$$
$$= \frac{1}{4\sqrt{2}} \sum_{r\in\mathbb{Z}} \theta_{|r|} \sum_{i\in I_k^1\cap(I_k^2-r)} \left[(P^T-P)(\psi_i+\psi_{i+r})\right]^2 - \left[(P^T-P)(\psi_i-\psi_{i+r})\right]^2.$$

Let $J_0 = \{j \in \mathbb{N} : \lfloor \frac{j}{r} \rfloor \text{ is even}\}$ and $J_1 = \{j \in \mathbb{N} : \lfloor \frac{j}{r} \rfloor \text{ is odd}\}$. Thus

$$V_4 = \frac{1}{4\sqrt{2}} \sum_{r\in\mathbb{Z}} \theta_{|r|} \sum_{(z,\varepsilon)\in\{0;1\}\times\{-1;1\}} \sum_{j\in J_z} \varepsilon(P^T-P)(\psi_i+\varepsilon\psi_{i+r})^2 \mathbb{I}_{I_k^1}(i)\mathbb{I}_{I_k^2}(i+r).$$

For any fixed $r \neq 0$, $(z,\varepsilon) \in \{0;1\}\times\{-1;1\}$, $\frac{1}{\sqrt{2}}\big(\psi_i+\varepsilon\psi_{i+r}\big)_{i\in J_z}$ is an orthonormal collection of functions, since for any $(i,j) \in J_z^2$,

$$\begin{aligned} <\psi_i+\varepsilon\psi_{i+r}, \psi_j+\varepsilon\psi_{j+r}> &= <\psi_i,\psi_j> + \varepsilon <\psi_{i+r},\psi_j> + \varepsilon <\psi_i,\psi_{j+r}> + <\psi_{i+r},\psi_{j+r}> \\ &= 2\delta_{i,j} + \varepsilon <\psi_{i+r},\psi_j> + \varepsilon <\psi_i,\psi_{j+r}> \\ &= 2\delta_{i,j} \text{ puisque } i,j \in J_z \text{ and } i+r, j+r \in J_{1-z}. \end{aligned}$$

[5, Lemma 14] applied to $S_m = \langle(\psi_i+\varepsilon\psi_{i+r})_{i\in J_z\cap I_k^1\cap I_k^2}\rangle$ for all $(z,\varepsilon) \in \{0;1\}\times\{-1;1\}$, $r \in \{-n_t,\dots,n_t\}$ and a union bound yield an event $E_4(y)$ of probability $\mathbb{P}(D_n^T \in E_4(y)) \geqslant 1-e^{-y}$ such that, for some absolute constant κ and for all $D_n^T \in E_4(y)$, $(z,\varepsilon) \in \{0;1\}\times\{-1;1\}$ and $r \in \mathbb{Z}$,

$$\sum_{i\in J_z\cap I_k^1\cap(I_k^2-r)} \varepsilon(P^T-P)(\psi_i+\varepsilon\psi_{i+r})^2 = (1\pm\delta)\frac{\varepsilon}{n_t} \sum_{i\in J_z\cap I_k^1\cap(I_k^2-r)} \big[\mathrm{Var}(\psi_i)+\mathrm{Var}(\psi_{i+r}) + 2\varepsilon\,\mathrm{Cov}(\psi_i,\psi_{i+r})\big] + \kappa\frac{\|s\|_\infty[\log(1+r)+\log n_t+y]}{(\delta\wedge 1)n_t} + \kappa\frac{|I_k^1|[\log(1+r)+\log n_t+y]^2}{(\delta\wedge 1)^3 n_t^2}.$$

By summing on $(r,z,\varepsilon) \in \mathbb{Z}\times\{0;1\}\times\{-1;1\}$ and since $\|\psi_i\|_\infty \leqslant \sqrt{2}$, it follows

that for all $D_n^T \in E_4(y)$,

$$\begin{aligned}\left|V_4 - \frac{1}{n_t}\sum_{r\in\mathbb{Z}}\frac{\theta_{|r|}}{\sqrt{2}}\sum_{i\in I_k^1\cap(I_k^2-r)} c_{i,i+r}\right| &= \left|V_4 - \frac{1}{n_t}\sum_{r=-n_t}^{n_t}\frac{\theta_{|r|}}{\sqrt{2}}\sum_{i\in I_k^1\cap(I_k^2-r)} c_{i,i+r}\right| \\ &\leqslant \frac{\delta}{n_t\sqrt{2}}\sum_{r\in\mathbb{Z}}|\theta_{|r|}|\sum_{i\in I_k^1\cap(I_k^2-r)}[\mathrm{Var}(\psi_i)+\mathrm{Var}(\psi_{i+r})] \\ &\quad + \kappa\sum_{r\in\mathbb{Z}}\frac{|\theta_{|r|}|}{\sqrt{2}}\times\frac{\|s\|_\infty[\log n_t\log(1+r)+y]}{(\delta\wedge 1)n_t} \\ &\quad + \kappa\sum_{r\in\mathbb{Z}}\frac{|\theta_{|r|}|}{\sqrt{2}}\times\frac{|I_k^1|[\log n_t+\log(1+r)+y]^2}{(\delta\wedge 1)^3 n_t^2}\end{aligned}$$

By hypothesis H1, $|\theta_j| \leqslant \frac{\sqrt{c_1}}{j^{1+\frac{\delta_1}{2}}}$, hence the sum $\sum_{r\in\mathbb{Z}}|\theta_{|r|}|\log(1+r)^2$ converges to a finite value $\|\theta\|_{1,\log^2}$. Moreover, by lemma 5.4.1, $|I_k^1| \leqslant (b_x - a_x)\Delta \leqslant 2(1+x)\Delta$, hence

$$\begin{aligned}\left|V_4 - \frac{1}{n_t}\sum_{r\in\mathbb{Z}}\frac{\theta_{|r|}}{\sqrt{2}}\sum_{i\in I_k^1\cap(I_k^2-r)} c_{i,i+r}\right| &\leqslant 6\|\theta\|_{\ell^1}(1+x)\frac{\delta}{n_t}\Delta + 2\kappa\|\theta\|_{1,\log^2}\frac{\|s\|_\infty[1+y]}{(\delta\wedge 1)n_t} \\ &\quad + 8\kappa(1+x)\|\theta\|_{1,\log^2}\frac{\Delta[1+y]^2}{(\delta\wedge 1)^3 n_t^2}.\end{aligned}$$

There exists therefore a constant $\kappa(\|\theta\|_{1,\log^2})$ such that, for all $D_n^T \in E_4(y)$,

$$\left|V_4 - \frac{1}{n_t}\sum_{r\in\mathbb{Z}}\frac{\theta_{|r|}}{\sqrt{2}}\sum_{i\in I_k^1\cap(I_k^2-r)} c_{i,i+r}\right| \leqslant \kappa\delta(1+x)\mathcal{E} + \frac{[\log n_t+y]}{(\delta\wedge 1)n_t} + \kappa(1+x)\frac{[\log n_t+y]^2}{(\delta\wedge 1)^3 n_t}\mathcal{E}.$$

Let now $\delta = \max\left\{\frac{n-n_t}{n_t}, n^{-\frac{1}{3}}\right\}^{\frac{3}{4}}$. By hypothesis H4 of Theorem 5.3.8, $\frac{n-n_t}{n_t} \leqslant n^{-\delta_3}$, therefore $\delta\mathcal{E} \leqslant n^{-\min\left(\frac{1}{4},\delta_3\right)}\mathcal{E}$. Moreover, $\mathcal{E} \geqslant \frac{1}{n_v}$ therefore $\frac{1}{\delta n_t} \leqslant \left(\frac{n-n_t}{n_t}\right)^{\frac{1}{4}}\frac{1}{n_v} \leqslant n^{-\frac{\delta_3}{4}}\mathcal{E}$. Finally, since $\delta \geqslant n^{-\frac{1}{4}}$ and $n_t \geqslant \frac{n}{2}$, $\frac{\mathcal{E}}{\delta^3 n_t} \leqslant 2n^{-\frac{1}{4}}\mathcal{E}$. Since $\delta_3 \leqslant 1$, there exists therefore a constant κ such that for all $D_n^T \in E_4(y)$,

$$\left|V_4 - \frac{1}{n_t\sqrt{2}}\sum_{i\in I_k^1}\sum_{j\in I_k^2}\theta_{|i-j|}c_{i,j}\right| \leqslant \kappa(1+x)[\log n_t+y]^2 n^{-\frac{\delta_3}{4}}\mathcal{E}. \tag{5.70}$$

Moreover, since $c_{i,j} = \frac{\theta_{i+j}}{\sqrt{2}} + \left(\frac{1-\delta_{i,j}}{\sqrt{2}} + \delta_{i,j}\right)\theta_{|i-j|} - \theta_i\theta_j$ and $\theta_0 = 1$,

$$\frac{1}{n_t}\sum_{i\in I_k^1}\sum_{j\in I_k^2}\frac{\theta_{|i-j|}}{\sqrt{2}}c_{i,j} = \sum_{i\in I_k^1\cap I_k^2}\frac{1}{\sqrt{2}}\left(1-\frac{1}{\sqrt{2}}\right)\frac{1}{n_t} + \frac{1}{n_t}\sum_{i\in I_k^1}\sum_{j\in I_k^2}\frac{\theta_{|i-j|}^2}{2} + \frac{1}{n_t}\sum_{i\in I_k^1}\sum_{j\in I_k^2}\frac{\theta_{|i-j|}\theta_{i+j}}{2} - \frac{1}{n_t}\sum_{i\in I_k^1}\sum_{j\in I_k^2}\frac{\theta_{|i-j|}}{\sqrt{2}}\theta_i\theta_j.$$

Since for all $j \in \mathbb{N}$, $|\theta_j| \leqslant 1$,

$$\begin{aligned}\left|\frac{1}{n_t}\sum_{i\in I_k^1}\sum_{j\in I_k^2}\frac{\theta_{|i-j|}}{\sqrt{2}}c_{i,j} - \left(1-\frac{1}{\sqrt{2}}\right)\frac{|I_k^1\cap I_k^2|}{n_t\sqrt{2}} - \frac{1}{n_t}\sum_{i\in I_k^1}\sum_{j\in I_k^2}\frac{\theta_{|i-j|}^2}{2}\right| &\leqslant \frac{2}{n_t}\left(\sum_{r\in\mathbb{N}}|\theta_r|\right)^2 \\ &\leqslant 2\frac{n-n_t}{n_t}\frac{1}{n_v}\left\|\theta\right\|_{\ell^1}^2 \\ &\leqslant 2\left\|\theta\right\|_{\ell^1}^2 n^{-\delta_3}\mathcal{E}\end{aligned} \tag{5.71}$$

since $\mathcal{E} \geqslant \frac{1}{n_v}$ and $\frac{n-n_t}{n_t} \geqslant n^{-\delta_3}$, by hypothesis H4 of Theorem 5.3.8. From equations (5.70) and (5.71), it follows that, for some constant $\kappa(\|\theta\|_{1,\log^2})$,

$$\left|V_4 - \left(1-\tfrac{1}{\sqrt{2}}\right)\frac{|I_k^1\cap I_k^2|}{n_t\sqrt{2}} - \frac{1}{2n_t}\sum_{i\in I_k^1}\sum_{j\in I_k^2}\theta_{|i-j|}^2\right| \leqslant \kappa(1+x)[\log n_t + y]^2 n^{-\frac{\delta_3}{4}}\mathcal{E}. \tag{5.72}$$

V_5 can be expressed as

$$V_5 = \left(1-\frac{1}{\sqrt{2}}\right)\sum_{j\in I_k^1\cap I_k^2}\left(\hat{\theta}_j^T - \theta_j\right)^2,$$

therefore by proposition 5.5.3, there exists an event $E_5(y)$ of probability greater than $1-e^{-y}$ such that for all $D_n^T \in E_5(y)$,

$$V_5 = \left(1-\frac{1}{\sqrt{2}}\right)\sum_{i\in I_k^1\cap I_k^2}\frac{1}{n_t} \pm \left(1-\frac{1}{\sqrt{2}}\right)\kappa_1(b_x-a_x)[\log n + y]^2 n^{-\min(\frac{1}{12},\frac{\delta_3}{2})}\mathfrak{e}(n)$$

It follows by lemma 5.4.1 that on $E_5(y)$,

$$\left|V_5 - \left(1-\frac{1}{\sqrt{2}}\right)\frac{|I_k^1\cap I_k^2|}{n_t}\right| \leqslant 4\kappa_1(1+x)[\log n + y]^2 n^{-\min(\frac{1}{12},\frac{\delta_3}{2})}\mathfrak{e}(n). \tag{5.73}$$

Finally, V_6 can be bounde in the following manner.

$$V_6 \leqslant \frac{1}{2}\sum_{i\in I_k^1}\sum_{j\in I_k^2}\left[(P^T-P)^2\psi_i+(P^T-P)^2\psi_j\right]\left[\frac{|\theta_{i+j}|}{\sqrt{2}}+|\theta_i||\theta_j|\right]$$
$$=\frac{1}{2}\sum_{i\in I_k^1}(P^T-P)^2\psi_i\sum_{j\in I_k^2}\left[\frac{|\theta_{i+j}|}{\sqrt{2}}+|\theta_i||\theta_j|\right]+\frac{1}{2}\sum_{j\in I_k^2}(P^T-P)^2\psi_j\sum_{i\in I_k^1}\left[\frac{|\theta_{i+j}|}{\sqrt{2}}+|\theta_i||\theta_j|\right].$$

Thus, by equation (5.64),

$$V_6 \leqslant \kappa\frac{(1+x)^{\frac{\delta_1}{2\rho_1}}}{n_t^{\frac{\delta_1}{3\rho_1}}}\times\left[\frac{1}{2}\sum_{i\in I_k^1}(P^T-P)^2\psi_i+\frac{1}{2}\sum_{j\in I_k^2}(P^T-P)^2\psi_j\right]$$
$$\leqslant \kappa\frac{(1+x)^{\frac{\delta_1}{2\rho_1}}}{n_t^{\frac{\delta_1}{3\rho_1}}}\times\sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta}\left(\hat{\theta}_j^T-\theta_j\right)^2. \tag{5.74}$$

By proposition 5.5.3, there exists an event $E_6(y)$ of probability greater than $1-e^{-y}$, such that for any $D_n^T \in E_6(y)$,

$$\sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta}\left(\hat{\theta}_j^T-\theta_j\right)^2 \leqslant (b_x-a_x)\mathcal{E}+\kappa_1(b_x-a_x)(y+\log n)^2 n^{-\min(\frac{1}{12},\frac{\delta_3}{2})}\mathfrak{e}$$
$$\leqslant 2\max(\kappa_1,1)(b_x-a_x)(y+\log n)^2\mathcal{E}$$
$$\leqslant 8\max(\kappa_1,1)(1+x)(y+\log n)^2\mathcal{E} \text{ by lemma 5.4.1.}$$

It follows by equation (5.74) that on $E_6(y)$, for a certain constant $\kappa(\kappa_1,\delta_1,c_1,\kappa_6)$,

$$V_6 \leqslant \kappa[y+\log n]^2\frac{(1+x)^{\frac{\delta_1}{2\rho_1}}}{n_t^{\frac{\delta_1}{3\rho_1}}}\mathcal{E}. \tag{5.75}$$

Combining equations (5.65), (5.68), (5.69), (5.72), (5.73), (5.75) on the event $\cap_{i=2}^6 E_i(\log 6+y)$ yields the result. ■

Lemma 5.5.7 *Let $s \in L^\infty([0;1])$ be a probability density function. For all $j\in\mathbb{N}$, let $\theta_j=\langle s,\psi_i\rangle$, where $\psi_0(x)=1$ and $\psi_j(x)=\sqrt{2}\cos(2j\pi x)$ for all $j\in\mathbb{N}^*$. Then for any finite set $I\subset\mathbb{N}$ and for all functions $u\in\mathbb{R}^I$,*

$$0 \leqslant \sum_{i\in I}\sum_{j\in I}u(i)u(j)\left(\frac{1-\delta_{i,j}}{\sqrt{2}}+\delta_{i,j}\right)\theta_{|i-j|} \leqslant \|s\|_\infty\sum_{i\in I}u(i)^2.$$

Proof Let $X \sim s$ be a random variable with distribution $s(x)dx$ on $[0;1]$. For any $x \in \mathbb{R}$, and any $i \neq j$,

$$\psi_i(x)\psi_j(x) = 2\cos(2i\pi x)\cos(2j\pi x) = \cos(2(i+j)\pi x)+\cos(2(i-j)\pi x) = \frac{\psi_{i+j}+\psi_{|i-j|}}{\sqrt{2}}.$$

If $i \neq j$, then $\mathrm{Cov}(\psi_i(X),\psi_j(X)) = \frac{\theta_{i+j}+\theta_{|i-j|}}{\sqrt{2}} - \theta_i\theta_j$. If $i = j$, $\mathrm{Var}(\psi_i(X)) = 1 + \frac{\theta_{2i}}{\sqrt{2}} - \theta_i^2$. Let $u \in \mathbb{R}^I$, $k \in \mathbb{N}$ and $t_k = \sum_{i\in I} u(i)\psi_{i+k}$, then

$$\mathrm{Var}(t_k(X)) = \sum_{i\in I}\sum_{j\in I} u(i)u(j)\left[\left(\frac{1-\delta_{i,j}}{\sqrt{2}}+\delta_{i,j}\right)\theta_{|i-j|} + \frac{\theta_{i+j+2k}}{\sqrt{2}} - \theta_{i+k}\theta_{j+k}\right]$$

Furthermore, $\lim_{n\to+\infty}\theta_n = 0$, hence

$$\lim_{k\to+\infty}\mathrm{Var}(t_k(X)) = \sum_{i\in I}\sum_{j\in I} u(i)u(j)\left(\frac{1-\delta_{i,j}}{\sqrt{2}}+\delta_{i,j}\right)\theta_{|i-j|}.$$

It immediately follows that $\sum_{i\in I}\sum_{j\in I} u(i)u(j)\left(\frac{1-\delta_{i,j}}{\sqrt{2}}+\delta_{i,j}\right)\theta_{|i-j|} \geqslant 0$. Moreover, for all $k \in \mathbb{N}$,

$$\begin{aligned}\mathrm{Var}(t_k(X)) &\leqslant \mathbb{E}\left[t_k(X)^2\right]\\ &= \int_0^1 t_k(x)^2 s(x)dx\\ &\leqslant \|s\|_\infty \|t_k\|^2\\ &\leqslant \|s\|_\infty \sum_{i\in I} u(i)^2.\end{aligned}$$

Thus

$$\sum_{i\in I}\sum_{j\in I} u(i)u(j)\left(\frac{1-\delta_{i,j}}{\sqrt{2}}+\delta_{i,j}\right)\theta_{|i-j|} \leqslant \|s\|_\infty \sum_{i\in I} u(i)^2.$$

■

Lemma 5.5.8 *Let $\varepsilon : \mathbb{R}_+ \to \mathbb{R}_+$ be a non-decreasing function such that $\varepsilon(0) > 0$ and $h_+ : \mathbb{R}_+ \to \mathbb{R}_+$ be a continuous, non-decreasing function. Let $g_0 : \mathbb{R}_+ \to \mathbb{R}$ be a continuous function such that, for any $s < t$,*

$$-\varepsilon(\max(s,t)) \leqslant g_0(t) - g_0(s) \leqslant \max\{h_+(t) - h_+(s), \varepsilon(\max(s,t))\}.$$

Assume that $\varepsilon(0) > 0$. Then there exists a continuous, non-decreasing function $g : \mathbb{R}_+ \to \mathbb{R}_+$ such that $g_0(0) = g(0)$,

$$\left\| \frac{g_0 - g}{\varepsilon} \right\|_\infty \leqslant 6,$$

and moreover

$$\forall x, y, |g(y) - g(x)| \leqslant |h_+(y) - h_+(x)|.$$

Proof Assume to begin with that ε is right-continuous. Let $r > 0, \delta > 0$. We define by induction a sequence $(x_i)_{i\in\mathbb{N}}$ and a function g on $[x_i; x_{i+1}]$. Let $x_0 = 0$ and $g(x_0) = g_0(x_0)$. For any $i \in \mathbb{N}$, assuming x_i and $g(x_i)$ have been defined, let

$$\begin{aligned} x_{i+1} &= \inf \left\{ x \geqslant x_i : g_0(x) \geqslant g_0(x_i) + 2\varepsilon(x_i) \text{ or } \varepsilon(x) \geqslant \frac{3}{2}\varepsilon(x_i) \right\} \\ \forall x \in]x_i; x_{i+1}], g(x) &= \begin{cases} g(x_i) \text{ if } \varepsilon(x_{i+1}) \geqslant \frac{3}{2}\varepsilon(x_i) \\ g(x_i) + \frac{g_0(x_{i+1}) - g_0(x_i)}{h_+(x_{i+1}) - h_+(x_i)}[h_+(x) - h_+(x_i)] \text{ otherwise.} \end{cases} \end{aligned} \tag{5.76}$$

If $x_{i+1} = +\infty$, the above definitions still make sense and the induction stops. Notice first that for any $x \in [x_i; x_{i+1}[$, $g_0(x) - g_0(x_i) \leqslant [h_+(x) - h_+(x_i)] \vee \varepsilon(x) \leqslant [h_+(x) - h_+(x_i)] \vee \frac{3}{2}\varepsilon(x_i)$. Thus,by continuity of g_0,

$$g_0(x_{i+1}) - g_0(x_i) \leqslant [h_+(x_{i+1}) - h_+(x_i)] \vee \frac{3}{2}\varepsilon(x_i).$$

By assumption, ε is right-continuous, therefore if $\varepsilon(x_{i+1}) < \frac{3}{2}\varepsilon(x_i)$, it must be that $\inf\{x \geqslant x_i : \varepsilon(x) \geqslant \frac{3}{2}\varepsilon(x_i)\} > x_{i+1}$. Then by definition of x_{i+1} and continuity of g_0, $g_0(x_{i+1}) = g_0(x_i) + 2\varepsilon(x_i)$, therefore

$$2\varepsilon(x_i) = g_0(x_{i+1}) - g_0(x_i) \leqslant [h_+(x_{i+1}) - h_+(x_i)] \vee \frac{3}{2}\varepsilon(x_i),$$

which implies that

$$0 < 2\varepsilon(x_i) = g_0(x_{i+1}) - g_0(x_i) \leqslant [h_+(x_{i+1}) - h_+(x_i)]. \tag{5.77}$$

This proves that g is well defined. g is non-decreasing and continuous since h_+ has these properties. If $\varepsilon(x_{i+1}) < \frac{3}{2}\varepsilon(x_i)$, then the previous equation implies that

$$\forall i \in \mathbb{N}, \forall (x, y) \in]x_i; x_{i+1}], x \leqslant y \implies g(y) - g(x) \leqslant h_+(y) - h_+(x),$$

else g is constant on $]x_i; x_{i+1}]$ and the above equation is trivially true. Hence, since g, h_+ are non-decreasing and continuous,

$$\forall (x, y) \in \mathbb{R}, x \leqslant y \implies g(y) - g(x) \leqslant h_+(y) - h_+(x).$$

We will now prove by induction that for all $i \in \mathbb{N}^*$,

$$0 \leqslant g_0(x_i) - g(x_i) \leqslant 4\varepsilon(x_i). \tag{5.78}$$

Base case: This equation is true for $i = 1$ since $x_0 = 0$ and $g(0) = g_0(0) = 0$, therefore by definition of g, x_1, $0 \leqslant g(x_1) \leqslant g_0(x_1) \leqslant 2\varepsilon(x_0) \leqslant 2\varepsilon(x_1)$.

Inductive step: Assume that equation (5.78) is true for some $i \in \mathbb{N}$. Then by definition of x_{i+1} and g,

- If $\varepsilon(x_{i+1}) \geqslant \frac{3}{2}\varepsilon(x_i)$, then $g(x_{i+1}) = g(x_i)$ therefore $g_0(x_{i+1}) - g(x_{i+1}) = g_0(x_{i+1}) - g_0(x_i) + g_0(x_i) - g(x_i)$. By the induction hypothesis and the definition of x_{i+1},

  $$0 \leqslant g_0(x_{i+1}) - g(x_{i+1}) \leqslant 2\varepsilon(x_i) + 4\varepsilon(x_i) \leqslant 6 \times \frac{2}{3}\varepsilon(x_{i+1}) \leqslant 4\varepsilon(x_{i+1}),$$

  which proves equation (5.78) for $i + 1$.

- Otherwise, by definition of g, $g(x_{i+1}) = g(x_i) + [g_0(x_{i+1}) - g_0(x_i)]$ therefore by the induction hypothesis and since ε is non-decreasing,

  $$0 \leqslant g_0(x_{i+1}) - g(x_{i+1}) = g_0(x_i) - g(x_i) \leqslant 4\varepsilon(x_i) \leqslant 4\varepsilon(x_{i+1}).$$

  This proves equation (5.78) for $i + 1$.

By induction, equation (5.78) is therefore true for all $i \in \mathbb{N}$ (such that $x_i < +\infty$). Let now $i \in \mathbb{N}$ and $x \in]x_i; x_{i+1}]$. By definition of g,

$$g(x_i) \leqslant g(x) \leqslant g(x_i) + (g_0(x_{i+1}) - g_0(x_i))_+.$$

By equation (5.78) and definition of x_{i+1},

$$\begin{aligned} g(x) - g_0(x) &\leqslant g(x) - g_0(x_i) \\ &\leqslant g(x_i) - g_0(x_i) + (g_0(x_{i+1}) - g_0(x_i))_+ \\ &\leqslant 2\varepsilon(x_i) \\ &\leqslant 2\varepsilon(x). \end{aligned}$$

Moreover, by equation (5.78) and definition of the x_i,

$$\begin{aligned} g(x) - g_0(x) &\geqslant g(x_i) - g_0(x_{i+1}) \\ &\geqslant g(x_i) - g_0(x_i) - [g_0(x_{i+1}) - g_0(x_i)] \\ &\geqslant -4\varepsilon(x_i) - 2\varepsilon(x_i) \\ &\geqslant -6\varepsilon(x_i) \\ &\geqslant -6\varepsilon(x). \end{aligned}$$

It has been proved that for all $i \in \mathbb{N}$ such that x_i is finite,

$$\forall x \in]x_i; x_{i+1}], |g(x) - g_0(x)| \leqslant 6\varepsilon(x).$$

It must now be proved that $\lim_{n\to+\infty} x_n = +\infty$. Since ε is non-decreasing and right-continuous, by definition of x_n, $g_0(x_{n+1}) \geqslant g_0(x_n) + 2\varepsilon(x_n) \geqslant g_0(x_n) + 2\varepsilon(0)$ or $\varepsilon(x_{n+1}) \geqslant \frac{3}{2}\varepsilon(x_n)$. Since $\varepsilon(0) > 0$ by assumption, this implies that $\max(g_0, \varepsilon)(x_n) \to +\infty$. The function $\max(g_0, \varepsilon)$ is non-decreasing, thus it is bounded on every interval of the form $[0; x]$, which implies that $x_n \to +\infty$. This proves the proposition under the assumption that ε is right-continuous.

In the general case, let $\varepsilon_+ : x \mapsto \inf_{y>x} \varepsilon(y)$, which is non-decreasing and right-continuous. Since ε is non-decreasing, $\varepsilon_+ \geqslant \varepsilon$, therefore the assumptions of the proposition hold with ε_+ instead of ε. By the right-continuous case of the proposition, which we already proved, there exists a non-decreasing function g such that $\left\| \frac{g-g_0}{\varepsilon_+} \right\|_\infty \leqslant 6$ and

$$\forall x, y, x \leqslant y \implies g(y) - g(x) \leqslant h_+(y) - h_+(x).$$

Let $x > 0$. By continuity of g, g_0,

$$|g(x) - g_0(x)| \leqslant \sup_{y<x} |g(y) - g_0(y)| \leqslant \sup_{y<x} 6\varepsilon_+(y) = 6 \sup_{y<x} \inf_{y'>y} \varepsilon(y') \leqslant 6\varepsilon(x).$$

This proves the proposition in the general case.

■

Proposition 5.5.9 *Let $([x_i; x_{i+1}[)_{1\leqslant i\leqslant M-1}$ be a partition of the interval $[a; b]$. Let $Y : \{x_1, \ldots, x_M\} \to \mathbb{R}$ be such that $(Y(x_j))_{1\leqslant j\leqslant M}$ is a zero-mean gaussian vector. Abusing notation, we also denote by Y the extension of Y to $[a; b]$ by linear interpolation. Let $K_Y : [a; b]^2 \to \mathbb{R}$ be the variance-covariance function of Y. Let $h : [a; b] \to \mathbb{R}$ be a continuous, increasing function and let $K_X : [a; b]^2 \to \mathbb{R}$ be a positive semi-definite function such that:*

$$\forall (s, t) \in [a; b]^2, \ |K_X(s, s) + K_X(t, t) - 2K_X(s, t)| \leqslant |h(s) - h(t)|.$$

Assume that there exists constants $L > 0$ and $\varepsilon \in [0; 1]$ such that:

- $\sup_{t\in[a;b]} \sqrt{K_X(t, t)} \leqslant L$
- *For any $i \in \{1, \ldots, M-1\}, h(x_{i+1}) - h(x_i) \leqslant \varepsilon$*
- $\max_{(i,j)\in\{1,\ldots,M\}^2} |K_X(x_i, x_j) - K_Y(x_i, x_j)| \leqslant \varepsilon.$

There exists a universal constant κ and a measurable function $f : C([a;b],\mathbb{R}) \to C([a;b],\mathbb{R})$ such that for all random variables $\nu \sim \mathcal{U}([0;1])$ independent from Y, $X = f(Y,\nu)$ is a zero-mean gaussian process with variance-covariance function K_X and moreover,

$$\mathbb{E}\left[\sup_{a\leqslant t\leqslant b}|X_t - Y_t|\right] \leqslant \kappa\sqrt{(1+L)\log M\,[(h(b)-h(a))\vee 1]}\varepsilon^{\frac{1}{12}}.$$

Proof We assume without loss of generality that $h(b) - h(a) \geqslant 1$. We shall moreover use the following notation. For A, B two symmetric matrices, $A \prec B$ means that $B - A$ is positive definite. $\|A\|_{op}$ denotes the matrix operator norm corresponding to the euclidean norm, i.e $\|A\|_{op} = \sup_{x:\|x\|_2\leqslant 1}|Ax|$. We will need the following lemmas:

Lemma 5.5.10 *For all $A \in \mathbb{R}^{m\times m}$, $\|A\|_{op} \leqslant m\max_{1\leqslant i,j\leqslant m}|A_{i,j}|$.*

Proof Let $v \in \mathbb{R}^m$ be such that $\sum_{i=1}^m v_i^2 = 1$. By the Cauchy-Schwartz inequality,

$$\|Av\|^2 = \sum_{i=1}^m\left(\sum_{j=1}^m A_{i,j}v_j\right)^2 \leqslant \sum_{i=1}^m\sum_{j=1}^m A_{i,j}^2 \leqslant m^2\max_{i=1,\dots,m}A_{i,j}^2.$$

This is true for any v, which proves lemma 5.5.10. ■

Lemma 5.5.11 below is a special case of Mc-Carthy's trace inequality ([78], Lemma 2.6).

Lemma 5.5.11 *Let A, B be two symmetric, positive semi-definite matrices, then*

$$Tr(\sqrt{A+B}) \leqslant Tr(\sqrt{A}) + Tr(\sqrt{B}).$$

The hypotheses imply that h is bijective from $[a;b]$ to $[h(a);h(b)]$. Let $m \in \mathbb{N}$. For all $j \in \{1,\dots,m\}$, let

$$t_j = \max\left\{x_i | i \in \{1,\dots,M\}, h(x_i) \leqslant h(a) + \frac{j-1}{m-1}[h(b)-h(a)]\right\}. \tag{5.79}$$

Let $K_{X,m} = \left(K_X(t_i,t_j)\right)_{1\leqslant i,j\leqslant m}$ and $K_{Y,m} = \left(K_Y(t_i,t_j)\right)_{1\leqslant i,j\leqslant m}$. The Wasserstein distance between two gaussian vectors is known [85]: there exists a coupling $\tilde{X}^m, \tilde{Y}^m$ of the distributions $\mathcal{N}(0,K_{X,m})$ and $\mathcal{N}(0,K_{Y,m})$ such that:

$$\mathbb{E}\left[\sum_{i=1}^m(\tilde{X}_i^m - \tilde{Y}_i^m)^2\right] = \mathrm{Tr}\left(K_{X,m} + K_{Y,m} - 2\big(K_{X,m}^{\frac{1}{2}}K_{Y,m}K_{X,m}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right).$$

Thus

$$\begin{aligned} K_{X,m}^2 &= K_{X,m}^{\frac{1}{2}} K_{Y,m} K_{X,m}^{\frac{1}{2}} + K_{X,m}^{\frac{1}{2}} \big(K_{Y,m} - K_{X,m}\big) K_{X,m}^{\frac{1}{2}} \\ &\prec K_{X,m}^{\frac{1}{2}} K_{Y,m} K_{X,m}^{\frac{1}{2}} + \|K_{Y,m} - K_{X,m}\|_{op} K_{X,m}. \end{aligned}$$

By lemma 5.5.11,

$$\mathrm{Tr}(K_{X,m}) \leqslant \mathrm{Tr}\left(\big(K_{X,m}^{\frac{1}{2}} K_{Y,m} K_{X,m}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right) + \|K_{Y,m} - K_{X,m}\|_{op}^{\frac{1}{2}} \mathrm{Tr}\left(K_{X,m}^{\frac{1}{2}}\right).$$

By the same argument (exchangeing X and Y),

$$\mathrm{Tr}(K_{Y,m}) \leqslant \mathrm{Tr}\left(\big(K_{Y,m}^{\frac{1}{2}} K_{X,m} K_{Y,m}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right) + \|K_{Y,m} - K_{X,m}\|_{op}^{\frac{1}{2}} \mathrm{Tr}\left(K_{Y,m}^{\frac{1}{2}}\right).$$

It follows that

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^{m} (\tilde{X}_i^m - \tilde{Y}_i^m)^2\right] &\leqslant \|K_{Y,m} - K_{X,m}\|_{op}^{\frac{1}{2}} \mathrm{Tr}\left(K_{X,m}^{\frac{1}{2}} + K_{Y,m}^{\frac{1}{2}}\right) \\ &\leqslant m^{\frac{1}{2}} \|K_{Y,m} - K_{X,m}\|_{\infty}^{\frac{1}{2}} \left(\sqrt{\mathrm{Tr}(K_{X,m})} + \sqrt{\mathrm{Tr}(K_{Y,m})}\right) \\ &\leqslant m^{\frac{1}{2}} \|K_{Y,m} - K_{X,m}\|_{\infty}^{\frac{1}{2}} m^{\frac{1}{2}} \max_{1 \leqslant i,j \leqslant m} \left\{\sqrt{K_X(t_i,t_i)} + \sqrt{K_Y(t_j,t_j)}\right\} \\ &\leqslant 2m\sqrt{\varepsilon}\sqrt{L^2+\varepsilon}. \end{aligned} \tag{5.80}$$

By the transfer principle (Kallenberg, Theorem 5.10), there exists f_1 such that for all uniform random variables variables ν_1 independent from Y, $(\tilde{X}^m, \tilde{Y}^m) \sim (f_1(Y^m, \nu_1), Y^m)$.

Let X_0 be a gaussian process with variance-covariance function K_X. Let $W_0 = X_0 \circ h^{-1}$. For any $(s,t) \in [h(a); h(b)]^2$, $W_0(t) - W_0(s)$ is a centred gaussian random variable, hence for all $r > 0$, there exists a universal constant $C(r)$ such that

$$\begin{aligned} \mathbb{E}\left[(W_0(t) - W_0(s))^r\right] &\leqslant C(r) \mathbb{E}\left[(W_0(t) - W_0(s))^2\right]^{\frac{r}{2}} \\ &\leqslant C(r)\Big(K_X(h^{-1}(t), h^{-1}(t)) + K_X(h^{-1}(s), h^{-1}(s)) \\ &\qquad - 2K_X(h^{-1}(s), h^{-1}(t))\Big)^{\frac{r}{2}} \\ &\leqslant C(r)|t-s|^{\frac{r}{2}}. \end{aligned}$$

By the Kolmogorov continuity theorem [89, Chapter 1, Theorem 2.1], applied to $x \mapsto \frac{W_0(h(a)+(h(b)-h(a))x)}{\sqrt{h(b)-h(a)}}$, there exists a continuous version W_1 of W_0 such that for any $\theta \in [0;1)$, and all $(s,t) \in [h(a); h(b)]^2$,

$$\mathbb{E}\left[\left(\sup_{s \neq t} \frac{|W_1(t) - W_1(s)|}{|t-s|^{\theta\left(\frac{1}{2} - \frac{1}{r}\right)}}\right)^r\right] \leqslant \frac{(h(b)-h(a))^{\frac{r}{2}}}{(h(b)-h(a))^{\theta r\left(\frac{1}{2}-\frac{1}{r}\right)}} B(\theta, r) < +\infty,$$

where $B(\theta, r)$ is a universal constant. Let $X_1 = W_1 \circ h$, which is still a gaussian process, with variance-covariance function K_X. Then, for any $(s,t) \in [a;b]^2$,

$$\mathbb{E}\left[\left(\sup_{s\neq t} \frac{|X_1(t) - X_1(s)|}{|h(t) - h(s)|^{\theta\left(\frac{1}{2}-\frac{1}{r}\right)}}\right)^r\right] \leqslant [h(b) - h(a)]^{\frac{r}{2}} B(\theta, r) < +\infty. \tag{5.81}$$

The $C([a;b], \mathbb{R})$-valued process X_1 induces a probability distribution Q on the Borel space $C([a;b], \mathbb{R})$. Furthermore, $(X_1(t_j))_{1\leqslant j\leqslant m} \sim \tilde{X}^m \sim f_1(Y^m, \nu_1)$. By (Kallenberg, Theorem 5.10), there exists a measurable function f_2 such that for all uniform random variables ν_3 independent from Y^m, ν_1, $(X_1, (X_1(t_j))_{1\leqslant j\leqslant m}) \sim (f_2(f_1(Y^m, \nu_1), \nu_3), f_1(Y^m, \nu_1))$. Let $X = f_2(f_1(Y^m, \nu_1), \nu_3)$ and $X^m = (X(t_j))_{1\leqslant j\leqslant m}$. Almost surely,

Claim 5.5.11.1 *1.* $X^m = (X(t_j))_{1\leqslant j\leqslant m} = f_1(Y^m, \nu_1)$ *p.s, so*

2. $(X^m, Y^m) \sim (\widetilde{X}^m, \widetilde{Y}^m)$*, in particular by equation* (5.80)*,*

$$\mathbb{E}\left[\|X^m - Y^m\|^2\right] \leqslant 2m\sqrt{\varepsilon}\sqrt{L^2 + \varepsilon}.$$

3. $X \sim X^1$ *as a random continuous function, in particular by equation* (5.81) *with* $\theta = \frac{3}{4}$ *and* $r = 6$*,*

$$\forall \delta > 0, \mathbb{E}\left[\sup_{(s,t)\in[a;b]:|h(t)-h(s)|\leqslant\delta} |X(t) - X(s)|^6\right]^{\frac{1}{6}} \leqslant \sqrt{h(b) - h(a)} B(\tfrac{3}{4}, 6)^{\frac{1}{6}} \delta^{\frac{1}{4}}. \tag{5.82}$$

By abuse of notation, denote X^m, Y^m the random processes obtained by linear interpolation between the points (t_j, X_j^m) and (t_j, Y_j^m), respectively. For all $t \in [a;b]$, there exists $j \in \{1, \ldots, m\}$ such that $t_j \leqslant t \leqslant t_{j+1}$, therefore $|h(t) - h(t_j)| \leqslant h(t_{j+1}) - h(t_j)$, since h is non-decreasing. By definition of t_{j+1} (equation (5.79)), $h(t_{j+1}) \leqslant h(a) + \frac{j}{m-1}(h(b) - h(a))$ and furthermore, there exists $i \in \{1, \ldots, M\}$ such that $x_i = t_j$. By equation (5.79) which defines t_j, $h(x_{i+1}) > h(a) + \frac{j-1}{m-1}(h(b) - h(a))$, which yields

$$\begin{aligned} |h(t) - h(t_j)| &\leqslant h(a) + \frac{j}{m-1}(h(b) - h(a)) - h(x_{i+1}) + h(x_{i+1}) - h(x_i) \\ &\leqslant \frac{h(b) - h(a)}{m-1} + h(x_{i+1}) - h(x_i). \end{aligned}$$

By assumption, $h(x_{i+1}) - h(x_i) \leqslant \varepsilon$, which yields

$$\forall t \in [a;b], \exists j(t) \in \{1, \ldots, m\}, t_j \leqslant t \leqslant t_{j+1} \text{ and } |h(t) - h(t_{j(t)})| \leqslant \frac{h(b) - h(a)}{m} + \varepsilon. \tag{5.83}$$

Since Y is piecewise linear on the partition $([x_i; x_{i+1}[)_{1\leqslant i\leqslant M-1}$,

$$\begin{aligned}\sup_{a\leqslant t\leqslant b}|X(t)-Y(t)| &\leqslant \sup_{t\in[a;b]}|X(t)-X(t_{j(t)})| + \max_{j\in\{1,\dots,m\}}|X(t_j)-Y(t_j)| \\ &\quad + \sup_{t\in[a;b]}|Y(t)-Y(t_{j(t)})| \\ &\leqslant \sup_{(s,t):|h(s)-h(t)|\leqslant\varepsilon+\frac{h(b)-h(a)}{m}}|X(s)-X(t)| + \sqrt{\sum_{j=1}^{m}|X(t_j)-Y(t_j)|^2} \\ &\quad + \max_{j\in\{1,\dots,m\}}\max\left\{|Y(x_i)-Y(t_j)| : i\in\{1,\dots,M\}\cap[t_j;t_{j+1}]\right\}.\end{aligned}$$

Thus, by claim 5.5.11.1,

$$\begin{aligned}&\mathbb{E}\left[\sup_{a\leqslant t\leqslant b}|X(t)-Y(t)|\right] \\ &\quad\leqslant \mathbb{E}\Big[\sup_{(s,t):|h(s)-h(t)|\leqslant\varepsilon+\frac{h(b)-h(a)}{m}}|X(s)-X(t)|\Big] + \mathbb{E}\Big[\sum_{i=1}^{m}(X_i^m-Y_i^m)^2\Big]^{\frac{1}{2}} \\ &\qquad + \mathbb{E}\left[\max_{j\in\{1,\dots,m\}}\max\left\{|Y(x_i)-Y(t_j)| : i\in\{1,\dots,M\}\cap[t_j;t_{j+1}]\right\}\right] \\ &\quad\leqslant \sqrt{h(b)-h(a)}B(\tfrac{3}{4},6)^{\frac{1}{6}}\left(\varepsilon+\frac{h(b)-h(a)}{m}\right)^{\frac{1}{4}} + \left(2m\sqrt{\varepsilon}\sqrt{L^2+\varepsilon}\right)^{\frac{1}{2}} \\ &\qquad + \sqrt{2\log M}\max_{j\in\{1,\dots,m\}}\max\left\{\sqrt{\mathbb{E}[|Y(x_i)-Y(t_j)|^2]} : i\in\{1,\dots,M\}\cap[t_j;t_{j+1}]\right\}.\end{aligned} \tag{5.84}$$

Furthermore, for any $(i,j)\in[1;M]^2$,

$$\begin{aligned}\mathbb{E}[(Y(x_i)-Y(x_j))^2] &= K_Y(x_i,x_i)+K_Y(x_j,x_j)-2K_Y(x_i,x_j) \\ &\leqslant K_X(x_i,x_i)+K_X(x_j,x_j)-2K_X(x_i,x_j) \\ &\quad + 4\max_{(r,s)\in[1;M]^2}|K_X(x_r,x_s)-K_Y(x_r,x_s)| \\ &\leqslant |h(x_i)-h(x_j)|+4\varepsilon.\end{aligned}$$

Setting $\kappa = B(\frac{3}{4},6)^{\frac{1}{6}}$, it follows by equation (5.84) and the non-decreasing nature

of h that

$$\begin{aligned}\mathbb{E}\left[\sup_{a\leqslant t\leqslant b}|X_t-Y_t|\right] &\leqslant \sqrt{h(b)-h(a)}B(\tfrac{3}{4},6)^{\frac{1}{6}}\left(\varepsilon+\frac{h(b)-h(a)}{m}\right)^{\frac{1}{4}}+\sqrt{2m(L+1)}\varepsilon^{\frac{1}{4}}\\ &\quad+\sqrt{2\log M}\sqrt{h(t_{j+1})-h(t_j)+4\varepsilon}\\ &\leqslant \kappa\sqrt{h(b)-h(a)}\varepsilon^{\frac{1}{4}}+\kappa\frac{[h(b)-h(a)]^{\frac{3}{4}}}{m^{\frac{1}{4}}}+\sqrt{2m(L+1)}\varepsilon^{\frac{1}{4}}\\ &\quad+\sqrt{2\log M}\sqrt{\frac{h(b)-h(a)}{m}+5\varepsilon}\text{ by equation (5.83).}\end{aligned}$$

Let now $m=\left\lceil\frac{h(b)-h(a)}{\varepsilon^{\frac{1}{3}}}\right\rceil$. Since by assumption $\varepsilon\leqslant 1$, $h(b)-h(a)\geqslant 1$ it follows finally, by keeping only the dominant powers of $[h(b)-h(a)],\varepsilon,L$ and $\log M$, that

$$\mathbb{E}\left[\sup_{a\leqslant t\leqslant b}|X_t-Y_t|\right]\leqslant\kappa\sqrt{h(b)-h(a)}\sqrt{2(L+1)\log M}\varepsilon^{\frac{1}{12}}.$$

for an absolute constant κ. ■

Chapter 6

A detailed analysis of Agghoo: accurate oracle inequalities

6.1 Introduction

In this chapter, we investigate the performance of the hold-out and Agghoo in the same setting as in the previous chapter, in the form of oracle inequalities.

General oracle inequalities for cross-validation in L^2 density estimation were proved by Arlot and Lerasle [5]. In the specific case of orthogonal series estimators on the cosine basis, Hall [50] proved that cross-validation is asymptotically equivalent to the oracle in terms of risk. His result also applies to some other orthogonal bases, as well as some families of weighted orthogonal series estimators. More recently, Magalhaes in his thesis [71, Corollary 3.2] proved with an oracle inequality that "cross-validation penalties" could be used to select the best estimator within the family of weighted orthogonal series estimators with non-increasing weights. In the setting of this chapter, the methods of Chapter 3 of this thesis can also be used to derive an oracle inequality for Agghoo and the hold-out (Theorem 6.6.5). Under stronger assumptions (Section 6.2.2), this upper bound achieves greater accuracy than those mentioned above (indeed, it is optimal up to log terms), however this comes at the cost of introducing the non-explicit quantity $\mathfrak{e}$ (Definition 6.4.1) and losing control of the expectation (the bound holds with probability greater than $1 - \frac{2}{n^2}$).

The main focus in the literature has generally been on obtaining oracle inequalities with leading constant 1, to show that the proposed methods perform as well as the oracle. These results typically do not attempt to identify the true magnitude of the model selection error, their purpose is rather to show that it is negligible relative to the oracle. For example, Hall's results in [50] are asymptotic, while Magalhaes shows an oracle inequality with leading constant $1 + \frac{\kappa}{\sqrt{\log n}}$,

which is presumably not optimal. This is insufficient for fine comparisons between methods such as the hold-out, cross-validation or Agghoo which all satisfy similar inequalities. In the case of Agghoo, the risk-reducing effects of aggregation should introduce negative terms in the oracle inequalities, the magnitude of which should be compared to the excess risk of the hold-out in order to assess the overall performance of the method. If the effect of aggregation outweighs the excess risk of the hold-out, Agghoo could very well have a risk lower than that of the model selection oracle, as shown in Figure 2.1 of the Introduction.

The purpose of this chapter is to carry out an analysis of Agghoo precise enough to shed some light on this phenomenon. To this end, the asymptotic approximation of the hold-out risk estimator, developed in the previous chapter, plays a crucial role. Under the same assumptions, we derive an asymptotic expression for the excess risk of the hold-out relative to the oracle (Theorem 6.4.3), and an upper bound for the risk of Agghoo (Theorem 6.4.4). These results are sufficiently accurate to measure the effect of aggregation: compared to the asymptotic expression for the hold-out's risk, the oracle inequality for Agghoo contains two additional negative terms, at the same order of approximation.

In the special case where the Fourier coefficients of the density s decrease polynomially (in absolute value), it is possible to compare the various terms appearing in these oracle inequalities (notably $\mathcal{E}$ and $\mathfrak{e}$, which are defined in Chapter 5. As a consequence, Corollary 6.4.6 shows that for some values of its parameters, Agghoo's risk can be smaller than the oracle risk by a constant factor. This behaviour provides evidence of the advantages of Agghoo relative to model-selection procedures such as cross-validation, which by definition cannot have a risk below that of the oracle.

6.2 Setting and hypotheses

The setting and hypotheses are the same as in the previous chapter. This section contains a brief summary. A more detailed discussion can be found in Chapter 5.

6.2.1 Setting

The setting is the same as in Chapter 5. Let us recall the main notations and assumptions. The statistical problem is least-squares density estimation: given an i.i.d sample $D_n = (X_1, \dots, X_n)$ with common probability density function $s \in L^2([0;1])$, assumed to be even, estimate s.

The collection of estimators considered for this purpose are the orthogonal series estimators on the trigonometric basis of cosines:

Definition 6.2.1 *For all $k \in \mathbb{N}$ and all $T \subset \{1 \dots n\}$,*

$$\hat{s}_k^T = \sum_{j=0}^{k} P_n^T(\psi_j)\psi_j,$$

where $\psi_0 = 1$, for all $j \in \mathbb{N}^*$, $\psi_j : x \mapsto \sqrt{2}\cos(2\pi jx)$ and for any measurable function t,

$$P_n^T(t) = \frac{1}{|T|}\sum_{i \in T} t(X_i).$$

We assume that the Fourier coefficients of s, $\theta_j = \langle s, \psi_j \rangle$, are non-increasing in absolute value. Let $(n_t(n))_{n\in\mathbb{N}}$ be a sequence of integers such that for all n, $1 \leqslant n_t(n) \leqslant n-1$. $n_t(n)$ represents the amount of data used for computing the density estimators, as opposed to estimator selection. The optimal choice of k and the optimal achievable risk for $\hat{s}_k^T$, respectively, can be approximated by the following quantities.

Definition 6.2.2 *(Definition 5.3.2 of Chapter 5)*

$$k_* = k_*(n_t) = \max\left\{ j \in \mathbb{N} : \theta_j^2 \geqslant \frac{1}{n_t} \right\}$$
$$\mathrm{or}(n_t) = \inf_{k\in\mathbb{N}} \left\{ \sum_{j=k+1}^{+\infty} \theta_j^2 + \frac{k}{n_t} \right\}.$$

6.2.2 Hypotheses

The following assumptions are made on $(\theta_j^2)_{j\in\mathbb{N}}$ and $(\tau_n)_{n\in\mathbb{N}}$. They are identical to the assumptions of Theorem 5.3.8 of the previous Chapter.

H0. The sequence $(\theta_j^2)_{j\in\mathbb{N}}$ is non-increasing.

H1. For all $n \in \mathbb{N}$, $n_t \in \{1, \dots, n-1\}$.

H2. There exists constants $c_1 \geqslant 0$ and $\delta_1 \geqslant 0$ such that for all $k \in \mathbb{N}$, $\sum_{j=k+1}^{+\infty} \theta_j^2 \leqslant \frac{c_1}{k^{2+\delta_1}}$.

H3. There exists constants $c_2 \geqslant 0$, $\rho_1 \geqslant 0$ such that for all $k \in \mathbb{N}$, $\sum_{j=k+1}^{+\infty} \theta_j^2 \geqslant \frac{c_2}{k^{\rho_1}}$.

H4. There exists constants $c_3 > 0$, $\delta_2 > 0$ such that for all $k \geqslant 1$,

$$\theta_{k+k^{\delta_2}}^2 \geqslant c_3 \theta_{k-k^{\delta_2}}^2.$$

H5. There exists a constant $\delta_3 > 0$ such that for all $n \in \mathbb{N}$, $n - n_t \leqslant n^{1-\delta_3}$.

H6. There exists a constant $\delta_4 > 0$ such that for all $n \in \mathbb{N}$, $n - n_t \geqslant n^{\frac{2}{3}+\delta_4}$.

6.3 Hold-out and Agghoo

This chapter focuses on the performance of hold-out and aggregated hold-out when applied to the hyperparameter k of Definition 6.2.1. Let us now define these methods. Let $n \in \mathbb{N}$, $n_t \in \{1, \dots, n-1\}$ and $n_v = n - n_t$. The hold-out is defined below.

Definition 6.3.1 *Let $T \subset \{1 \dots n\}$ be a subset of cardinality $|T| = n_t$. Let*

$$\hat{k}_T = \min \operatorname*{argmin}_{k \in \{1,\dots,n-n_t\}} \left\{ \left\| \hat{s}_k^T \right\|^2 - 2P_n^{T^c}(\hat{s}_k^T) \right\},$$

which yields the non-parametric density estimator

$$\widehat{s}_T^{\mathrm{ho}} = \hat{s}_{\hat{k}_T}^T.$$

In practice, it is natural to impose some restriction on k in order to compute the estimator. It would be more natural to select $\hat{k}_T$ from the interval $\{1, \dots, n_t\}$ instead of $\{1, \dots, n - n_t\}$, yet this restriction is useful in the proof of the theorems of the next section (note that $n - n_t \leqslant n_t$ by hypothesis H5). It will have no negative impact on performance as long as $\hat{k}_T^* \in \{1, \dots, n - n_t\}$, where

$$\hat{k}_T^* = \min \operatorname*{argmin}_{k \in \mathbb{N}} \left\{ \left\| \hat{s}_k^T - s \right\|^2 \right\}.$$

By hypothesis (H2),

$$\mathrm{or}(n_t) \leqslant \inf_{k \in \mathbb{N}} \left\{ \frac{c_1}{k^2} + \frac{k}{n_t} \right\} = \mathcal{O}(n_t^{-\frac{2}{3}}).$$

Since $\mathrm{or}(n_t) \geqslant \frac{k_*(n_t)}{n_t}$, this implies that $k_*(n_t) = \mathcal{O}(n_t^{\frac{1}{3}})$. By a concentration argument (Proposition 6.9.2), the same is true of $\hat{k}_T^*$, with high probability. Therefore, by assumptions (H2) and (H6) of Section 6.2.2, $n - n_t \geqslant \hat{k}_T^*$ holds with high probability, for all n large enough.

For a given size n_t, the hold-out procedure has a free hyperparameter: the subset T which does not affect the distribution of $\widehat{s}_T^{\mathrm{ho}}$. This suggests that aggregating several hold-out estimator $\widehat{s}_{T_j}^{\mathrm{ho}}$ with different subsets T_j might be a way to reduce the variance of the hold-out estimator and improve its performance. This general

strategy, known as *Aggregated hold-out*, is defined in Chapter 3 of this thesis (Section 3.3.2). There are several ways to generate subsets T_j for aggregation of the $\widehat{s}^{\text{ho}}_T$. In this chapter, I will focus on two particular algorithms, analogous to V-fold CV and Monte-Carlo CV, respectively. First, consider the $V-$fold procedure defined below.

Definition 6.3.2 ***V-fold Agghoo*** *Let $n_t \in \{1, \dots, n-1\}$. Let $(T_j)_{j=1,\dots,V}$ be a potentially random collection of subsets of $\{1 \dots n\}$, having the same cardinality $n_t = n_t$, independent from the data D_n, and such that the sets T_j^c are pairwise disjoint. Let then*

$$\widehat{s}^{vf}_{n_t,V} = \frac{1}{V} \sum_{j=1}^{V} \widehat{s}^{\text{ho}}_{T_j}$$

be the V-fold aggregated hold-out estimator.

The pointwise value of the estimator $\widehat{s}^{vf}_{n_t,V}$ depends on the exact choice of the T_j, but its distribution only depends on n_t and V, hence the notation.

Here, the blocks T_j^c are not required to form a partition of $\{1 \dots n\}$, but only assumed to be pairwise disjoint: there are therefore two parameters, n_t (size of the training sample) and V (size of the ensemble), which in the case of a complete V-fold procedure would be linked by the equation $(n - n_t)V = n$. By comparison, the definition of $\widehat{s}^{vf}_{n_t,V}$ only requires that $V \leqslant \frac{n}{n-n_t}$.

By uncoupling V and n_t, Definition 6.3.2 makes it easier to understand their respective influence on the procedure and its performance.

Instead of being pairwise disjoint, the test sets T_j^c can be randomly generated in a Monte-Carlo procedure. This leads to the following variant of Agghoo.

Definition 6.3.3 ***Monte-Carlo Agghoo*** *Let $(T_j)_{j=1,\dots,V}$ be i.i.d subsets of $\{1, \dots, n\}$, independent from the data D_n and drawn uniformly from the set*

$$\{T \subset \{1 \dots n\}, |T| = n_t\}.$$

Let then

$$\widehat{s}^{mc}_{n_t,V} = \frac{1}{V} \sum_{j=1}^{V} \widehat{s}^{\text{ho}}_{T_j}$$

be the Monte-Carlo aggregated hold-out estimator.

Thus, $\widehat{s}^{mc}_{n_t,V}$ is random even given a fixed dataset D_n; however, its distribution only depends on n_t and V, as for $\widehat{s}^{vf}_{n_t,V}$.

6.4 Oracle inequalities

This section contains the main results of this chapter. We state oracle inequalities which are sufficiently precise to allow a comparison between Agghoo, the hold-out and the oracle. Subsection 6.4.2 contains oracle inequalities valid in the general setting of section 6.2.1, under the hypotheses of section 6.2.2. Subsection 6.4.3 gives an example in which Agghoo can perform better than the oracle by a constant factor in the asymptotic $n \to +\infty$.

6.4.1 Key quantities

Let us briefly recall the following definitions from Chapter 5 (Definition 5.3.3 of that Chapter).

Definition 6.4.1 *For any $n \in \mathbb{N}$, let*

$$\begin{aligned}
\Delta_d(s, n_t, n) &= \max\left\{ l \in \mathbb{N} : \theta^2_{k_*(n_t)+l} \geqslant \left[1 - \sqrt{\frac{n_t}{n-n_t}}\frac{1}{\sqrt{l}}\right]\frac{1}{n_t} \right\} \\
\Delta_g(s, n_t, n) &= \min\left\{ l \in \{0, \dots, k_*(n_t)\} : \theta^2_{k_*(n_t)-l} \geqslant \left[1 + \sqrt{\frac{n_t}{n-n_t}}\frac{1}{\sqrt{l}}\right]\frac{1}{n_t} \right\} \\
\Delta(s, n_t, n) &= \max\left(\Delta_d(s, n_t, n), \Delta_g(s, n_t, n)\right) \\
\mathcal{E}(s, n_t, n) &= \frac{\Delta(s, n_t, n)}{n_t}. \\
\mathfrak{e}(s, n_t, n) &= \sqrt{\frac{\mathcal{E}(s, n_t, n)}{n-n_t}}.
\end{aligned}$$

The quantities $\Delta, \mathfrak{e}$ were used in Chapter 5 to rescale the hold-out risk estimator in order to derive an asymptotic approximation for the rescaled process. It was argued there that Δ is the order of magnitude of $\hat{k}_T - k_*(n_t)$, where $\hat{k}_T$ is the parameter selected by the hold-out, as in Definition 6.3.1. Recall that by Lemma 5.3.4 of Chapter 5, $\mathcal{E}$ and $\mathfrak{e}$ satisfy the following inequalities.

$$\mathfrak{e} \geqslant \frac{1}{n-n_t} \tag{6.1}$$

$$\mathfrak{e} \leqslant \mathcal{E} \tag{6.2}$$

$$\mathcal{E} \leqslant 2\mathrm{or}(n_t) + \frac{1}{n-n_t}. \tag{6.3}$$

Recall also the definition of the function f_n that plays a key role in Chapter 5 (Definition 5.3.6 of Chapter 5).

Definition 6.4.2 *For all $k \in \mathbb{N}$, let $R(k) = \sum_{j=k+1}^{+\infty} \theta_j^2$. Extend R to $\mathbb{R}_+$ by linear interpolation: $\forall x \in \mathbb{R}_+, R(x) = (1 + \lfloor x \rfloor - x)R(\lfloor x \rfloor) + (x - \lfloor x \rfloor)R(\lfloor x \rfloor + 1)$. Let then $f_n : \left[\frac{-k_*(n_t)}{\Delta}; +\infty\right[\to \mathbb{R}_+$ be defined by*

$$\forall \alpha \in \mathbb{R}, \qquad f_n(\alpha) = \frac{1}{\mathfrak{e}} \left(R(k_* + \alpha\Delta) - R(k_*) + \frac{\alpha\Delta}{n_t} \right). \tag{6.4}$$

Thus, for all $k \in \mathbb{N}$, $k \neq k_*$,

$$\mathfrak{e} f_n \left(\frac{k - k_*}{\Delta} \right) = \sum_{j=k\wedge k_*+1}^{k\vee k_*} \left| \theta_j^2 - \frac{1}{n_t} \right|. \tag{6.5}$$

f_n approximates a rescaled version of the *excess risk* $\left\|\hat{s}_k^T - s\right\|^2 - \left\|\hat{s}_{k_*}^T - s\right\|^2$. It is a convex, piecewise linear function that satisfies certain bounds on its increments (Lemma 5.3.7). More discussion of f_n and of its properties can be found in Chapter 5 (Definition 5.3.6).

6.4.2 Main results

Theorem 5.3.8 of the previous chapter suggests that the excess risk of the hold-out relative to the oracle, $\left\|\widehat{s}_T^{\text{ho}} - s\right\|^2 - \text{or}(n_t)$, is of order $\mathfrak{e}$. Theorem 6.4.3 below confirms this supposition, and gives an asymptotic expansion of the risk of the hold-out at first order in $\mathfrak{e}$, as $n \to +\infty$.

Theorem 6.4.3 *Let*

$$\hat{\alpha} = \underset{\alpha \in \left[\frac{-k_*}{\Delta}; +\infty\right[}{\text{argmin}} \{f_n(\alpha) - W_{g_n(\alpha)}\},$$

where g_n is the same as in Theorem 5.3.8 of the previous chapter and W is symmetrized Brownian motion ($(W_t)_{t\geqslant 0}$ and $(W_{-t})_{t\geqslant 0}$ are independent standard BMs). Then, under the assumptions of Section 6.2.2, as $n \to +\infty$,

$$\mathbb{E}\left[\left\|\widehat{s}_T^{\text{ho}} - s\right\|^2\right] = \text{or}(n_t) + \mathbb{E}[f_n(\hat{\alpha})]\mathfrak{e} + o(\mathfrak{e}) \tag{6.6}$$

$$\leqslant \text{or}(n_t) + \kappa_{ho}\mathfrak{e} + o(\mathfrak{e}), \tag{6.7}$$

where κ_{ho} depends on $\|s\|_\infty$ only.

Theorem 6.4.3 is proved in Section 6.7.3, using the results of Sections 6.7.1 and 6.7.2. Theorem 6.4.3 guarantees that the excess risk of the hold-out, relative to the oracle, is at most of order $\mathfrak{e}$.

Together with equation (6.3), Theorem 6.4.3 implies that

$$\mathbb{E}\left[\left\|\widehat{s}_T^{\mathrm{ho}} - s\right\|^2\right] \leqslant \mathrm{or}(n_t) + (1+o(1))\left[\kappa_{ho}\sqrt{\frac{2\mathrm{or}(n_t)}{n-n_t}} + \frac{\kappa_{ho}}{n-n_t}\right].$$

In particular the expected risk of the hold-out is asymptotically equivalent to $\mathrm{or}(n_t)$ whenever $\frac{1}{n-n_t} = o(\mathrm{or}(n_t))$. This is the same, up to log terms, as what could be obtained using the general oracle inequalities of chapter 3 (Theorem 3.7.3, optimized in θ). However, the bound $\mathcal{E} \leqslant 2\mathrm{or}(n_t) + \frac{1}{n-n_t}$ (equation (6.3)) may be far from optimal, since it amounts to the upper bounds $\Delta_g \leqslant k_*(n_t)$ and $\Delta_d \leqslant +\infty$.

In the other direction, there is no lower bound on $\mathbb{E}[f_n(\hat{\alpha})]$; however, it can be expected not to tend to 0 in general, given that the functions f_n, g_n are lower and upper-bounded by Lemma 5.3.7 and Theorem 5.3.8 of Chapter 5. Assuming that $\mathbb{E}[f_n(\hat{\alpha})]$ is lower bounded, equation (6.1) and Theorem 6.4.3 show that a remainder term of order $\frac{1}{n-n_t}$ is unavoidable in general.

The rate of convergence implicit in the o is proportional to a power of $\frac{1}{n}$; the constant and the exponent can be upper bounded entirely in term of the constants featuring in Section 6.2.2, as will be apparent in the proof.

When several hold-out estimators are aggregated, as in Definitions 6.3.2 and 6.3.3, performance improves significantly, and we have the following oracle inequality.

Theorem 6.4.4 *Let*

$$\hat{\alpha} = \underset{\alpha \in \left[\frac{-k_*}{\Delta}; +\infty\right[}{\mathrm{argmin}} \{f_n(\alpha) - W_{g_n(\alpha)}\}$$

as in Theorem 6.4.3. Let $F : \mathbb{R} \to [0;1]$ be the distribution function of $\hat{\alpha}$. In the limit $n \to +\infty$, under the assumptions of section 6.2.2,

$$\mathbb{E}\left[\left\|\widehat{s}_{n_t,V}^{vf} - s\right\|^2\right] \leqslant \mathbb{E}\left[\left\|\widehat{s}_T^{\mathrm{ho}} - s\right\|^2\right] - \frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*(n_t)}{n_t} - \frac{V-1}{V}\mathcal{E} \times \left(2\int_{-\infty}^{0}[F(1-F)](x)dx + \int_0^{+\infty}[F(1-F)](x)dx\right) + o(\mathcal{E}). \tag{6.8}$$

A similar inequality holds for the Monte-Carlo aggregate $\widehat{s}_{n_t,V}^{mc}$:

$$\mathbb{E}\left[\left\|\widehat{s}_{n_t,V}^{mc} - s\right\|^2\right] \leqslant \mathbb{E}\left[\left\|\widehat{s}_T^{\mathrm{ho}} - s\right\|^2\right] - \frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*(n_t)}{n_t} + [1+o(1)]\left(\frac{n-n_t}{n}\right)^2\frac{k_*(n_t)}{n_t} - \frac{V-1}{V}\mathcal{E} \times \left(2\int_{-\infty}^{0}[F(1-F)](x)dx + \int_0^{+\infty}[F(1-F)](x)dx\right) + o(\mathcal{E}). \tag{6.9}$$

Furthermore, there exists a constant $\kappa_{ag}(\|s\|_\infty, \|s\|^2) > 0$ *such that, for all* $n \in \mathbb{N}$, $\int_{-\infty}^{+\infty}[F(1-F)](x)dx \geqslant \kappa_{ag}$, *and therefore:*

$$\mathbb{E}\left[\left\|\widehat{s}^{vf}_{n_t,V} - s\right\|^2\right] \leqslant \mathbb{E}\left[\left\|\widehat{s}^{\text{ho}}_{T} - s\right\|^2\right] - \frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*(n_t)}{n_t} - \frac{V-1}{V}\kappa_{ag}\mathcal{E} + o(\mathcal{E}). \tag{6.10}$$

Similarly, for $\widehat{s}^{mc}_{n_t,V}$,

$$\mathbb{E}\left[\left\|\widehat{s}^{vf}_{n_t,V} - s\right\|^2\right] \leqslant \mathbb{E}\left[\left\|\widehat{s}^{\text{ho}}_{T} - s\right\|^2\right] - \frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*(n_t)}{n_t} + [1+o(1)]\left(\frac{n-n_t}{n}\right)^2\frac{k_*(n_t)}{n_t} - \frac{V-1}{V}\kappa_{ag}\mathcal{E} + o(\mathcal{E}).$$

In particular, if $\mathfrak{e} = o(\mathcal{E})$, *then by Theorem 6.4.3,*

$$\mathbb{E}\left[\left\|\widehat{s}^{vf}_{n_t,V} - s\right\|^2\right] \leqslant or(n_t) - \frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*(n_t)}{n_t} - \frac{V-1}{V}\kappa_{ag}\mathcal{E} + o(\mathcal{E}) \tag{6.11}$$

and the same holds for $\widehat{s}^{mc}_{n_t,V}$.

Theorem 6.4.4 is proved in Section 6.7.4, using the results of Sections 6.7.1 and 6.7.2. Assume to simplify the discussion that $\frac{V-1}{V} = \frac{n_t}{n}$, as in the classical V-fold procedure. In particular, $\frac{V-1}{V} \to 1$. In that case, Theorem 6.4.4 shows that compared to the hold-out, the risk of Agghoo is decreased by two terms: $\kappa_{ag}\mathcal{E}$ and the term $\frac{n-n_t}{n}\frac{k_*(n_t)}{n_t}$. These terms have distinct origins. The term $\frac{n-n_t}{n}\frac{k_*(n_t)}{n_t}$ corresponds to a bagging effect: for small values of j, the basis functions ψ_j are selected in all the models $E_{\hat{k}_j^{ho}} = \langle(\psi_i)_{1\leqslant i\leqslant \hat{k}_j^{ho}}\rangle$, and their coefficients are averaged, leading to a decreased risk, comparable to replacing $P_n^T(\psi_j)$ with $P_n(\psi_j)$ (i.e retraining the estimator on the whole dataset).

The term proportional to $\mathcal{E}$, on the other hand, arises from the aggregation of the $\hat{s}_k^T$ for different values of the parameter k (rather than different values of T). The inequality $\mathcal{E} \geqslant \mathfrak{e}$ from Lemma 5.3.4 ensures that the reduction in risk due to aggregation is always significant relative to the excess risk of the hold-out, which is of order $\mathfrak{e}$ by Theorem 6.4.3. The interest of Theorem 6.4.4 lies in the fact that $\mathcal{E}$ can be much bigger than $\mathfrak{e}$, leading the oracle inequality (6.11), in which Agghoo performs asymptotically better than $or(n_t)$ by a factor proportional to $\mathcal{E}$. Thus, Agghoo's performance depends on $\mathcal{E}$ and $\frac{\mathcal{E}}{\mathfrak{e}}$. If $\frac{\mathcal{E}}{\mathfrak{e}}$ is sufficiently large, then Agghoo performs better than the oracle (trained on a sample of size n_t). When this occurs, the amount by which the risk decreases is greater than $\kappa\mathcal{E}$, for a constant $\kappa > 0$.

By definition, both $\mathcal{E}$ and $\frac{\mathcal{E}}{\mathfrak{e}}$ are increasing functions of Δ. By definition, for fixed n_t, n, $\Delta(s, n_t, n)$ is the larger the slower the sequence θ_j^2 decreases in the neighbourhood of $k_*(n_t)$. Therefore, one should expect Agghoo to perform

better for sequences θ_j^2 that decrease slowly as they reach the value $\frac{1}{n_t}$ (since $\theta_{k_*}^2 \geqslant \frac{1}{n_t} \geqslant \theta_{k_*+1}^2$ by definition).

The oracle inequalities of Theorem 6.4.4 also depend explicitly on V, the number of subsets used in the aggregation, which gives an indication of how this parameter influences Agghoo's performance. Inequality (6.11) implies that even for $V = 2$, Agghoo performs better than the oracle when $\mathfrak{e} = o(\mathcal{E})$. The amount by which the risk is shown to decrease depends on V through the factor $\frac{V-1}{V}$. Assuming that this upper-bound on Agghoo's risk reflects its actual performance, at least qualitatively, this provides theoretical evidence for the claim (made in chapters 3 and 4 based on simulations) that small values of V ($V = 5$ or 10) are sufficient to reap most of the benefits of aggregation.

6.4.3 Example

The theoretical results of section 6.4.2 are expressed in terms of the quantities $\mathfrak{e}, \mathcal{E}$ and $or(n_t)$. In particular, as discussed previously, Agghoo' performance improves relative to the oracle when $\mathcal{E}$ grows, as long as $\frac{\mathcal{E}}{\mathfrak{e}}$ is sufficiently large. The general bounds of lemma 5.3.4 only show that $\frac{\mathcal{E}}{\mathfrak{e}} \leqslant 1$, which implies that $\mathbb{E}\left[\left\|\widehat{s}_{n_t,V}^{vf} - s\right\|^2\right] \leqslant or(n_t) + \rho\mathfrak{e}$, just like the hold-out, although with a better constant $\rho < \kappa_{ho}$. To better compare the performance of Agghoo with the hold-out and with the oracle, it is necessary to obtain bounds on $\frac{\Delta}{\mathcal{E}}$ and $\mathcal{E}$ that are more accurate than those of lemma 5.3.4. However, as discussed in Chapter 5, lemma 5.3.4 can be optimal for some sequences θ_j^2. In this section, we consider some sequences θ_j^2 for which it is possible to obtain better bounds on $\mathcal{E}$ and $\frac{\mathfrak{e}}{\mathcal{E}}$. We consider densities s such that the squared Fourier coefficients θ_j^2 decrease like an inverse power of j, that is, $(\theta_j^2)_{j\geqslant 1}$ is nonincreasing and there exist two constants $L > 0, \beta > 1$ such that

$$\theta_j^2 \sim Lj^{-2\beta-1}. \tag{6.12}$$

The assumption $\beta > 1$ ensures that the hypotheses of section 6.2.2 hold. More precisely, they hold for all $\delta_1 \leqslant 2(\beta - 1), \rho_1 \geqslant 2\beta$ and $\delta_2 < 1$. Assumption (6.12) is sufficient to give an equivalent for or(n) and for $k_*(n)$. It is not quite enough to obtain asymptotic expressions for $\Delta, \mathcal{E}$ and $\mathfrak{e}$, however; the reason is that the local fluctuations $\theta_{j+k}^2 - \theta_j^2$ may be very much larger (or smaller) than they are for the asymptotically equivalent sequence $Lj^{-2\beta-1}$. Nonetheless, assumption (6.12) implies several asymptotic relationships between the sequences $\Delta, \mathcal{E}$ and $\mathfrak{e}$. These are stated in lemma 6.4.5 below.

Lemma 6.4.5 *Assuming equation* (6.12) *holds true, if* $\frac{n_t(n)}{n} \to 1$ *and* $n - n_t(n) \to$

$+\infty$ *as* $n \to +\infty$, *then*

$$k_*(n_t) \sim (Ln_t)^{\frac{1}{2\beta+1}} \tag{6.13}$$

$$\mathrm{or}(n) \sim \left[\frac{1}{2\beta}+1\right] \frac{L^{\frac{1}{2\beta+1}}}{n^{\frac{2\beta}{2\beta+1}}} = \left[\frac{1}{2\beta}+1\right] \frac{k_*(n)}{n} \tag{6.14}$$

$$\mathrm{or}(n_t) - \mathrm{or}(n) \sim \frac{n-n_t}{n}\frac{k_*(n)}{n}. \tag{6.15}$$

For all $\eta_0 > 0$ *and all sufficiently large integers* n,

$$\frac{n_t}{n-n_t} \leqslant \eta_0 k_*(n) \implies \frac{\mathcal{E}(s,n_t,n)}{\mathfrak{e}(s,n_t,n)} \geqslant \frac{(2\eta_0)^{-\frac{1}{3}}}{4\beta+2}. \tag{6.16}$$

For all constants $\varepsilon_0 > 0$ *and all sufficiently large* n,

$$\frac{n_t}{n-n_t} \geqslant \sqrt{\varepsilon_0 k_*(n)} \implies \mathcal{E}(s,n_t,n) \geqslant [1-o(1)]\varepsilon_0 \frac{n-n_t}{n_t}\frac{k_*(n_t)}{n_t}. \tag{6.17}$$

Finally, for all constants $\eta \in [0;\eta_0]$ *and all sufficiently large* n,

$$\eta k_*(n) \leqslant \frac{n_t}{n-n_t} \leqslant \eta_0 k_*(n) \implies \mathcal{E}(s,n_t,n) \geqslant [1-o(1)]\eta \frac{2\beta}{2\beta+1}\mathrm{or}(n_t). \tag{6.18}$$

Lemma 6.4.5 is proved in Section 6.8.2. It states three bounds and three asymptotic equivalents, which are all relevant to the performance of Agghoo. Equation (6.15) shows that as $V \to +\infty$ the term $\frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*(n)}{n}$ from Theorem 6.4.4 cancels $\mathrm{or}(n_t) - \mathrm{or}(n)$, at first order. Equation (6.16) gives a sufficient condition for $\frac{\mathcal{E}}{\mathfrak{e}}$ to be arbitrarily large. For small enough η_0, this guarantees that $\kappa_{ho}\mathfrak{e} - \kappa_{ag}\frac{V-1}{V}\mathcal{E} \leqslant \kappa_{ag}\frac{V-1}{2V}\mathcal{E}$, which implies by Theorems 6.4.3 and 6.4.4 that $\mathbb{E}[\|\widehat{s}^{vf}_{n_t,V} - s\|^2] < \mathrm{or}(n_t)$. Equation (6.17) gives a sufficient condition for $\mathcal{E}$ to be larger than $\mathrm{or}(n_t) - \mathrm{or}(n) \sim \frac{n-n_t}{n_t}\frac{k_*(n)}{n}$, which helps to prove that $\mathbb{E}[\|\widehat{s}^{vf}_{n_t,V} - s\|^2] < \mathrm{or}(n)$. Finally, equation (6.18) gives a sufficient condition for $\mathcal{E}$ to be of order $\mathrm{or}(n_t)$, which together with equation (6.16) allows to derive an oracle inequality with leading constant smaller than 1. More precisely, Theorems 6.4.3 and 6.4.4, together with Lemma 6.4.5, yield the following corollary.

Corollary 6.4.6 *Assume that the squared Fourier coefficients* θ_j^2 *of* s *are non-increasing and satisfy equation* (6.12) *for some* $\beta > 1$ *(they decrease at a fixed polynomial rate). Then all assumptions of section 6.2.2 are satisfied and moreover, there exists a constant* $\eta_0(\beta, \|s\|_\infty, \|s\|^2)$ *such that the following equations hold.*

- *Let* $\delta_3 > 0$. *For all* $V \geqslant 2$ *and all sufficiently large* n,

$$n^{\delta_3} \leqslant \frac{n_t}{n-n_t} \leqslant \eta_0 k_*(n) \sim \eta_0 (Ln_t)^{\frac{1}{2\beta+1}} \implies \mathbb{E}\left[\left\|\widehat{s}^{vf}_{n_t,V} - s\right\|^2\right] < \mathrm{or}(n_t). \tag{6.19}$$

- *For all $\varepsilon_0 > 0$, there exists an integer n_0 such that for all $n \geqslant n_0$ and all $V \geqslant \max\left(\lceil \frac{12}{\kappa_{ag}\varepsilon_0} \rceil, 5\right)$,*

$$\sqrt{\varepsilon_0 k_*(n)} \leqslant \frac{n_t}{n-n_t} \leqslant \eta_0 k_*(n) \implies \mathbb{E}\left[\left\|\widehat{s}^{vf}_{n_t,V} - s\right\|^2\right] < \mathrm{or}(n). \tag{6.20}$$

- *For all $\eta > 0$ and all $V \geqslant 5$, there exists an integer n_0 such that*

$$\begin{aligned} &\forall n \geqslant n_0, \eta k_*(n) \leqslant \frac{n_t}{n-n_t} \leqslant \eta_0 k_*(n) \implies \\ &\mathbb{E}\left[\left\|\widehat{s}^{vf}_{n_t,V} - s\right\|^2\right] \leqslant \left(1 - \frac{\eta}{6}\frac{2\beta}{2\beta+1}\kappa_{ag}\right)\mathrm{or}(n), \end{aligned} \tag{6.21}$$

where $\kappa_{ag}(\|s\|_\infty, \|s\|^2) > 0$ is the same as in Theorem 6.4.4.

The same results hold for $\widehat{s}^{mc}_{n_t,V}$ instead of $\widehat{s}^{vf}_{n_t,V}$. Moreover, one can take

$$\eta_0(\beta, \|s\|_\infty, \|s\|^2) = \frac{1}{250(4\beta+2)^3}\left(\frac{\kappa_{ag}}{\kappa_{ho}}\right)^3,$$

where κ_{ag}, κ_{ho} are the constants which appear in Theorems 6.4.3 and 6.4.4.

Corollary 6.4.6 is proved in section 6.8.2. It shows that, depending on the size of $\frac{n_t}{n-n_t}$ relative to $k_*(n)$, the risk of Agghoo $\mathbb{E}\left[\left\|\widehat{s}^{vf}_{n_t,V} - s\right\|^2\right]$ may be smaller than $\mathrm{or}(n_t)$, smaller than $\mathrm{or}(n)$ and even smaller than $\theta\mathrm{or}(n)$, for some constant $\theta < 1$. Interestingly, few constraints are imposed on V: at most, for equation (6.20), V is required to be larger than some constant depending on ε_0. For equations (6.19) and (6.21), it is only necessary that $V \geqslant 2$ (an obvious requirement) and $V \geqslant 5$, respectively. In particular, V is never required to tend to $+\infty$. This confirms the claim, made in previous chapters on the basis of experimental evidence, that the benefits of aggregation can be had for small values of V ($V \leqslant 10$).

Of the assumptions made on $\frac{n_t}{n-n_t}$, the most important is the upper bound $\frac{n_t}{n-n_t} \leqslant \eta_0 k_*(n_t)$, which appears in all three equations of Corollary 6.4.6. This assumption is not surprising, since

$$\frac{n_t}{n-n_t} \geqslant \eta_0 k_*(n_t) \implies \frac{1}{n-n_t} \geqslant \eta_0 \frac{k_*(n_t)}{n_t} \sim \eta_0 \frac{2\beta}{2\beta+1}\mathrm{or}(n_t),$$

which means that the hold-out "model selection error", which is typically larger than $\frac{1}{n-n_t}$, is becoming significant relative to the total risk. This is obviously bad for the hold-out. For Agghoo, it implies, since by Lemma 6.6.2 $\mathfrak{e} \geqslant \frac{1}{n-n_t}$ and

$\mathcal{E} \leqslant 2\mathrm{or}(n_t) + \frac{1}{n-n_t}$, that $\mathcal{E} = \mathcal{O}\left(\sqrt{\frac{\mathcal{E}}{n-n_t}}\right) = \mathcal{O}(\mathfrak{e})$. Thus, depending on the values of κ_{ho} and κ_{ag}, the upper bound on Agghoo's risk in equation 6.10 may be of order $or(n_t) + \kappa\mathfrak{e}$, like the hold-out, rather than $or(n_t) - \kappa'\mathcal{E}$ as in equation (6.11). This suggests that Agghoo's performance, like that of the hold-out, may deteriorate when $\frac{n_t}{n-n_t}$ is greater than some constant times $k_*(n_t)$.

Inequality (6.21) suggests, on the contrary, that performance improves as $\frac{n_t}{n-n_t}$ grows, when $\frac{n_t}{n-n_t} \leqslant \eta_0 k_*(n_t)$; thus, it is reasonable to conjecture that there is an optimal value $\eta_0 > 0$, at which the risk of Agghoo is lowest, and such that the performance worsens for $\frac{n_t}{n-n_t} > \eta_0 k_*(n_t)$. For $\frac{n_t}{n-n_t} \sim \eta k_*(n_t)$, $0 < \eta \leqslant \eta_0$, Corollary 6.4.6 shows that Agghoo improves on the model selection oracle by a constant factor. This result has the disadvantage that it requires $\frac{n_t}{n-n_t}$ to be chosen in a distribution-dependent way, since $k_*(n_t) \sim (Ln_t)^{\frac{1}{2\beta+1}}$ However, $k_*(n_t)$ can in theory be estimated, for example by running a first round of the hold-out with a training set of size $|T| = \frac{n}{2}$ and using the hold-out parameter, $\hat{k}_T^{ho}$, as a plug-in estimator for $k_*(\frac{n}{2})$. By lemma 6.4.5, $k_*(\frac{n}{2})$ differs asymptotically from $k_*(n)$ by a constant factor: hence, choosing $\frac{n_t}{n-n_t} = \eta\hat{k}_T^{ho}$ for a small enough constant $\eta > 0$ can be expected to yield an estimator which satisfies equation (6.21), at least for all pdfs s such that $\eta_0(\|s\|^2, \|s\|_\infty) > \eta$.

Less ambitiously, choosing $\frac{n_t}{n-n_t}$ so as to satisfy the hypothesis of equation (6.20) allows for more robustness with respect to the parameters $\beta, L, \|s\|_\infty, \|s\|^2$. As $k_*(n) \sim (Ln)^{\frac{1}{2\beta+1}}$ by Lemma 6.4.5, it is enough to choose n_t such that $\frac{n_t}{n-n_t} \sim n^{\frac{1}{2\alpha+1}}$, where $\frac{1}{4\beta+2} \leqslant \frac{1}{2\alpha+1} < \frac{1}{2\beta+1}$ for equation (6.20) to be satisfied for all n large enough. Conversely, if n_t is chosen such that $\frac{n_t}{n-n_t} \sim n^{\frac{1}{2\alpha+1}}$, equation (6.20) will hold eventually for all β in the semi-open interval $[\frac{\alpha}{2} - \frac{1}{4}; \alpha[$. Thus, a deterministic choice of n_t is possible if some information about β is available.

Finally, the weakest result, equation (6.19) requires no lower bound on $\frac{n_t}{n-n_t}$ other than the arbitrarily small polynomial lower bound n^{δ_3} required by the hypotheses of Section 6.2.2.

6.5 Conclusion

We have given an asymptotic expression for the risk of the hold-out estimator, exhibiting the general order of magnitude of the "model selection error" $\mathbb{E}\left[\left\|\widehat{s}_T^{\mathrm{ho}} - s\right\|^2\right] - \mathrm{or}(n_t)$. Moreover, we have shown, through an oracle inequality (Theorem 6.4.4), that Agghoo improves significantly on the performance of the hold-out. In Corollary 6.4.6, we showed that if the Fourier coefficients of s decrease polynomially, Agghoo can even satisfy an oracle inequality with leading constant less than 1, if its parameter n_t is well chosen.

A natural question is whether these results can be generalized. A first remark, as in Chapter 5, is that the proof strategy only involves the Fourier coefficients θ_j through sums on intervals of length of order Δ, hence assumptions bearing on θ_j^2 could be replaced by versions bearing on local averages of the form $\frac{1}{2m_n}\sum_{j=k-m_n}^{k+m_n}\theta_j^2$, for $m_n = o(\Delta)$. This applies in particular to the assumption that the θ_j^2 form a non-increasing sequence.

It could also be interesting to consider relaxations of hypothesis (6.12) to obtain explicit results under weaker assumptions. A possible approach would be to average the risk over several values of the sample size n, considering for example the *regret*

$$\frac{1}{N}\sum_{n=1}^{N}\left\|\hat{s}^{mc}_{n_t(n),V} - s\right\|^2,$$

instead of the risk for a single sample size N. Bounding the regret instead of the risk may allow to relax assumption (6.12) because this "averages" values of n where θ_j^2 decreases rapidly around $k_*(n)$ and values of n for which θ_j^2 is flat around $k_*(n)$. As discussed in Section 6.4.2, the upper bounds on Agghoo's risk in Theorem 6.4.4 depend on the speed at which θ_j^2 decreases around $k_*(n_t)$. Thus, Agghoo's performance may be good on average, even if there are exceptions for some values of $k_*(n)$.

Aside from relaxing assumptions on the pdf s, a different and more ambitious generalization would be to other estimators. For the general approach to work, these would need to be orthogonal series estimators associated with "nested models", i.e $\hat{t}_k = \sum_{j=1}^{r_k} P_n(\phi_j)\phi_j$, where r_k is an increasing sequence and $(\phi_j)_{1\leqslant j\leqslant k}$ is an orthonormal family. Moreover, since the computation of the covariance between two estimators $\hat{t}_{k_1}, \hat{t}_{k_2}$ involves products $\phi_i\phi j$, a very convenient property is that the $(\phi_j)_{1\leqslant j\leqslant r_k}$ form an algebra. This suggests that multidimensional Fourier series, spherical harmonics, or orthogonal polynomials might also be suitable, though the unboundedness of polynomial bases with respect to their argument x or to the index k might prove to be a problem.

6.6 Preliminary resuts

Before the beginning of the proofs, let us recall here the definition of a_x, b_x, identical to that given in Theorem 5.3.8 of the previous chapter.

Definition 6.6.1 *For any $x \geqslant 0$, let*

$$a_x = \min\left\{\frac{j}{\Delta} : j \in \mathbb{Z} \cap [-k_*(n_t); 0], f_n\left(\frac{j}{\Delta}\right) \leqslant x\right\}$$
$$b_x = \max\left\{\frac{j}{\Delta} : j \in \mathbb{N}, f_n\left(\frac{j}{\Delta}\right) \leqslant x\right\}.$$

Note that by convexity of f_n, $[a_x; b_x] \subset f_n^{-1}([0; x]) \subset \left[a_x - \frac{1}{\Delta}; b_x + \frac{1}{\Delta}\right]$.

6.6.1 Results proved in Chapter 5

The following results are stated and proved in Chapter 5, but as they are of much help in this chapter too, they are reproduced here.

Lemma 6.6.2 *(Lemma 5.3.4) For any density s such that the sequence $\theta_j^2 = \langle s, \psi_j\rangle^2$ is non-increasing,*

$$\Delta \geqslant \frac{n_t}{n - n_t} \tag{6.22}$$
$$\mathcal{E} \geqslant \frac{1}{n - n_t} \tag{6.23}$$
$$\mathfrak{e} \geqslant \frac{1}{n - n_t} \tag{6.24}$$
$$\mathfrak{e} \leqslant \mathcal{E} \tag{6.25}$$
$$\mathcal{E} \leqslant 2\mathrm{or}(n_t) + \frac{1}{n - n_t}. \tag{6.26}$$

Lemma 6.6.3 *(Lemma 5.3.7) For all $\alpha_1, \alpha_2 \in \mathbb{R}$ such that $\alpha_1\alpha_2 \geqslant 0$ and $|\alpha_2| \geqslant |\alpha_1| \geqslant 1$,*

$$f_n(\alpha_2) - f_n(\alpha_1) \geqslant |\alpha_2| - |\alpha_1|.$$

Furthermore,

- *If $\Delta = \Delta_d$, then for all $\alpha_1, \alpha_2 \in [0; 1]$ such that $\alpha_1 \leqslant \alpha_2$, $f_n(\alpha_2) - f_n(\alpha_1) \leqslant \alpha_2 - \alpha_1$.*
- *If $\Delta = \Delta_g$,then for all $\alpha_1, \alpha_2 \in [-1; 0]$ such that $\alpha_1 \leqslant \alpha_2$, $f_n(\alpha_1) - f_n(\alpha_2) \leqslant \alpha_2 - \alpha_1$.*

Lemma 6.6.4 *(Lemma 5.4.1) Let a_x, b_x as in theorem 5.3.8. For all $x > 0$:*

$$[b_x - a_x] \leqslant 2(1 + x)$$
$$\sum_{k_* + a_x\Delta}^{k_* + b_x\Delta} \theta_j^2 \leqslant 4(1 + x)\mathcal{E}.$$

6.6.2 A first oracle inequality for the hold-out

We begin by showing that the $\hat{k}_T$ parameter selected by hold-out lies with high probability in an interval $[a_x; b_x]$, for x of the order of $\log^2 n$. This result is required to apply Theorem 5.3.8 of the previous chapter, which focuses on what happens in these intervals $[a_x; b_x]$.

Theorem 6.6.5 *Let $T \subset \{1; n\}$ be a set of cardinality $|T| = n_t = n_t$. Let*

$$\hat{k}_T \in \underset{k\in\{1,\dots,n-n_t\}}{\operatorname{argmin}} \left\{ \left\|\hat{s}_k^T\right\|^2 - 2P_n^{T^c}\hat{s}_k^T \right\}.$$

There exists a constant $\kappa_2 = \kappa_2(\|s\|_\infty)$ and an integer $n_0 = n_0(\kappa_1, (n_t(n))_{n\in\mathbb{N}})$ such that for all $n \geqslant n_0$, with probability greater than $1 - \frac{2}{n^2}$,

$$\left\|\widehat{s}_T^{\mathrm{ho}} - s\right\|^2 - \left\|\hat{s}^T_{k_*(n_t)} - s\right\|^2 \leqslant \kappa_2 \log^2 n\mathfrak{e},$$

and on the same event,

$$f_n\left(\frac{\hat{k}_T - k_*}{\Delta}\right) \leqslant \kappa_2 \log^2 n.$$

Proof Let $r_n = \kappa_1(4\log n)^2 n^{-\frac{\delta_3}{2}}$, where κ_1 is the same as in Proposition 6.9.2. In this proof, we will also use the notation $n_v = n - n_t$ for the cardinality of the validation sample. By assumption H6, $n_v \geqslant n^{\frac{2}{3}}$ while by assumption (H2),

$$k_*(n_t) \leqslant n_t \mathrm{or}(n_t) \leqslant n_t \inf_{k\in\mathbb{N}} \left\{ \frac{c_1}{k^{2+\delta_1}} + \frac{k}{n_t} \right\} = \mathcal{O}\left(n_t^{\frac{1}{3}}\right).$$

It follows that there exists a constant $n_0 \in \mathbb{N}$ such that:

$$\forall n \geqslant n_0, r_n \leqslant \frac{1}{2} \text{ and } k_*(n_t) \leqslant n_v. \tag{6.27}$$

Let $n \geqslant n_0$. Let E_n be the D_n^T- mesurable event on which for all $(\alpha_1, \alpha_2) \in \left\{\frac{k-k_*}{\Delta} : k \in \{0, \dots, n_v\}\right\}^2$ such that $\alpha_1 < \alpha_2$,

$$\left| \sum_{j=k_*+\alpha_1\Delta}^{k_*+\alpha_2\Delta} \left(\hat{\theta}_j^T - \theta_j\right)^2 - \frac{(\alpha_2 - \alpha_1)\Delta}{n_t} \right| \leqslant [\alpha_2 - \alpha_1] r_n \mathfrak{e}.$$

By proposition 6.9.2, $\mathbb{P}(D_n^T \in E_n) \geqslant 1 - \frac{1}{n^2}$. Until further notice, let us fix a training set $D_n^T \in E_n$, and consider the collection $(\hat{s}_k^T)_{1\leqslant k\leqslant n_v}$ of estimators, as well

as the loss function $\gamma : (t, x) \mapsto \|t\|^2 - 2t(x)$ with domain $\left\{\hat{s}_k^T | 1 \leqslant k \leqslant n_v\right\}$. The corresponding "bayes estimator" is the oracle $\hat{s}_{\hat{k}_*}^T$, where

$$\hat{k}_* \in \underset{k\in\{0,\dots,n_v\}}{\operatorname{argmin}} \left\|\hat{s}_k^T - s\right\|^2 = \underset{k\in\{0,\dots,n_v\}}{\operatorname{argmin}} \sum_{j=k+1}^{+\infty} \theta_j^2 + \sum_{j=1}^{k} \left(\hat{\theta}_j^T - \theta_j^2\right),$$

and the excess risk is $\ell\left(\hat{s}_{\hat{k}_*}^T, \hat{s}_k^T\right) = \left\|\hat{s}_k^T - s\right\|^2 - \left\|\hat{s}_{\hat{k}_*}^T - s\right\|^2$. Let $k \in \{0, \dots, n_v\}$. If $\frac{k-k_*(n_t)}{\Delta} \geqslant 1$, equation 6.27 which defines n_0,

$$\begin{aligned}
\left\|\hat{s}_k^T - s\right\|^2 - \left\|\hat{s}_{\hat{k}_*}^T - s\right\|^2 &\geqslant \left\|\hat{s}_k^T - s\right\|^2 - \left\|\hat{s}_{k_*}^T - s\right\|^2 \\
&= -\sum_{j=k_*(n_t)+1}^{k} \theta_j^2 + \sum_{j=k_*(n)+1}^{k} \left(\hat{\theta}_j^T - \theta_j\right)^2 \\
&\geqslant -\sum_{j=k_*(n_t)+1}^{k} \theta_j^2 + \frac{k-k_*}{n_t} - \frac{k-k_*}{\Delta} r_n \mathfrak{e} \\
&= \sum_{j=k_*(n_t)+1}^{k} \left[\frac{1}{n_t} - \theta_j^2\right] - \frac{k-k_*}{\Delta} r_n \mathfrak{e}. \qquad (6.28)
\end{aligned}$$

Hence, by definition of f_n and lemma 6.6.3,

$$\begin{aligned}
\left\|\hat{s}_k^T - s\right\|^2 - \left\|\hat{s}_{\hat{k}_*}^T - s\right\|^2 &\geqslant \mathfrak{e} f_n\left(\frac{k-k_*}{\Delta}\right) - \frac{k-k_*}{\Delta} \mathfrak{e} r_n \\
&\geqslant \mathfrak{e}\left[\frac{|k-k_*|}{\Delta} - 1\right] - \frac{|k-k_*|}{\Delta} \mathfrak{e} r_n \\
&\geqslant \left[(1-r_n)\frac{|k-k_*|}{\Delta} - 1\right] \mathfrak{e} \\
&\geqslant \left[\frac{1}{2}\frac{|k-k_*|}{\Delta} - 1\right] \mathfrak{e}. \qquad (6.29)
\end{aligned}$$

By the same argument, if $k \leqslant k_*(n_t) - \Delta$,

$$\begin{aligned}
\left\|\hat{s}_k^T - s\right\|^2 - \left\|\hat{s}_{\hat{k}_*}^T - s\right\|^2 &\geqslant \left\|\hat{s}_k^T - s\right\|^2 - \left\|\hat{s}_{k_*}^T - s\right\|^2 \\
&= \sum_{j=k+1}^{k_*} \theta_j^2 - \sum_{j=k+1}^{k_*} \left(\hat{\theta}_j^T - \theta_j\right)^2 \\
&\geqslant \sum_{j=k+1}^{k_*} \theta_j^2 - \frac{k_* - k}{n_t} - \frac{|k - k_*|}{\Delta} r_n \mathfrak{e} \\
&= \sum_{j=k+1}^{k_*} \left[\theta_j^2 - \frac{1}{n_t}\right] - \frac{|k - k_*|}{\Delta} r_n \mathfrak{e}. \qquad (6.30)
\end{aligned}$$

Hence, by definition of f_n and lemma 6.6.3,

$$\begin{aligned}
&= \mathfrak{e} f_n\left(\frac{k - k_*}{\Delta}\right) - \frac{k - k_*}{\Delta}\mathfrak{e} r_n \\
&\geqslant \mathfrak{e}\left[\frac{|k - k_*|}{\Delta} - 1\right] - \frac{|k - k_*|}{\Delta}\mathfrak{e} r_n \\
&\geqslant \left[(1 - r_n)\frac{|k - k_*|}{\Delta} - 1\right]\mathfrak{e} \\
&\geqslant \left[\frac{1}{2}\frac{|k - k_*|}{\Delta} - 1\right]\mathfrak{e}. \qquad (6.31)
\end{aligned}$$

By changing variables to $\alpha = \frac{k - k_*}{\Delta}$, this yields

$$|\alpha| > 2 \implies \ell(\widehat{s}_{\hat{k}_*}^T, \widehat{s}_{k_* + \alpha\Delta}^T) \geqslant \frac{|\alpha| - 2}{2}\mathfrak{e}(n). \qquad (6.32)$$

Furthermore, for all $x \in \mathbb{R}$, $\gamma(\hat{s}_k^T, x) = \left\|\hat{s}_k^T\right\|^2 - 2\hat{s}_k^T(x)$. For all $(k_1, k_2) \in \mathbb{N}^2$, let $c_{k_1,k_2} = \left\|\hat{s}_{k_1}^T\right\|^2 - \left\|\hat{s}_{k_2}^T\right\|^2$. Then for any variable X independent from D_n, with distribution P and pdf s,

$$\begin{aligned}
\mathbb{E}\left[\left(\gamma\left(\hat{s}_{k_1}^T, X\right) - \gamma\left(\hat{s}_{k_2}^T, X\right) - c_{k_1,k_2}\right)^2\right] &= \int_0^1 (\hat{s}_{k_1}^T - \hat{s}_{k_2}^T)^2(x) s(x) dx \\
&\leqslant \|s\|_\infty \left\|\hat{s}_{k_1}^T - \hat{s}_{k_2}^T\right\|^2. \qquad (6.33)
\end{aligned}$$

Furthermore, let $(k_1, k_2) \in \{0, \dots, n_v\}^2$ and $\alpha_1 = \frac{k_1 - k_*}{\Delta}, \alpha_2 = \frac{k_2 - k_*}{\Delta}$. Then

$$\begin{aligned}
\left\|\hat{s}^T_{k_1} - \hat{s}^T_{k_2}\right\|^2 &= \sum_{j=k_1\wedge k_2+1}^{k_1\vee k_2} \left(\hat{\theta}^T_j\right)^2 \\
&\leqslant 2 \sum_{j=k_1\wedge k_2+1}^{k_1\vee k_2} \theta_j^2 + 2 \sum_{j=k_1\wedge k_2+1}^{k_1\vee k_2} \left(\hat{\theta}^T_j - \theta_j\right)^2 \\
&\leqslant 2 \sum_{j=k_1\wedge k_2+1}^{k_1\vee k_2} \theta_j^2 + 2\frac{|k_1 - k_2|}{n_t} + 2|\alpha_2 - \alpha_1| r_n \mathfrak{e} \\
&\leqslant 2 \sum_{j=k_*\wedge k_1+1}^{k_*\vee k_1} \theta_j^2 + 2 \sum_{j=k_*\wedge k_2+1}^{k_*\vee k_2} \theta_j^2 + 2\frac{|k_1 - k_*|}{n_t} + 2\frac{|k_2 - k_*|}{n_t} + |\alpha_2 - \alpha_1|\mathfrak{e} \ (r_n \leqslant \frac{1}{2}) \\
&\leqslant 2 \left| \sum_{j=k_*\wedge k_1+1}^{k_1\vee k_*} \theta_j^2 - \frac{1}{n_t} \right| + 2 \left| \sum_{j=k_*\wedge k_2+1}^{k_2\vee k_*} \theta_j^2 - \frac{1}{n_t} \right| + 4\frac{|k_1 - k_*|}{n_t} + 4\frac{|k_2 - k_*|}{n_t} \\
&\quad + [|\alpha_2| + |\alpha_1|]\mathcal{E}.
\end{aligned}$$

By definition, $\frac{k_1 - k_*}{n_t} = \frac{k_1 - k_*}{\Delta}\frac{\Delta}{n_t} = \alpha_1\mathcal{E}$, the same holds for k_2, therefore by equations (6.28) and (6.30) and since $r_n \leqslant \frac{1}{2}$ for all $n \geqslant n_0$,

$$\begin{aligned}
\left\|\hat{s}^T_{k_1} - \hat{s}^T_{k_2}\right\|^2 &\leqslant \ell(\widehat{s}^T_{\hat{k}_*}, \widehat{s}^T_{k_1}) + \ell(\widehat{s}^T_{\hat{k}_*}, \widehat{s}^T_{k_2}) + 6|\alpha_2|\mathcal{E}(n) + 6|\alpha_1|\mathcal{E}(n) \\
&\leqslant \ell(\widehat{s}^T_{\hat{k}_*}, \widehat{s}^T_{k_1}) + \ell(\widehat{s}^T_{\hat{k}_*}, \widehat{s}^T_{k_2}) + 24\mathcal{E}(n) + 6(|\alpha_1| - 2)_+\mathcal{E}(n) + 6(|\alpha_2| - 2)_+\mathcal{E}(n).
\end{aligned}$$

Finally, by equation (6.32), it follows

$$\left\|\hat{s}^T_{k_1} - \hat{s}^T_{k_2}\right\|^2 \leqslant \ell(\widehat{s}^T_{\hat{k}_*}, \widehat{s}^T_{k_1}) + \ell(\widehat{s}^T_{\hat{k}_*}, \widehat{s}^T_{k_2}) + 24\mathcal{E}(n) + 12\frac{\mathcal{E}(n)}{\mathfrak{e}(n)}\left[\ell(\widehat{s}^T_{\hat{k}_*}, \widehat{s}^T_{k_1}) + \ell(\widehat{s}^T_{\hat{k}_*}, \widehat{s}^T_{k_2})\right].$$

Let $w_1(u) = \sqrt{\|s\|_\infty}\sqrt{1 + 12\frac{\mathcal{E}}{\mathfrak{e}}}u + \sqrt{12\|s\|_\infty}\sqrt{\mathcal{E}}$. By equation (6.33),

$$P\left|\gamma\big(\hat{s}^T_{k_1}, \cdot\big) - \gamma\big(\hat{s}^T_{k_2}, \cdot\big) - c_{k_1,k_2}\right|^2 \leqslant \left(w_1(\sqrt{\ell(\widehat{s}^T_{\hat{k}_*}, \widehat{s}^T_{k_1})}) + w_1(\sqrt{\ell(\widehat{s}^T_{\hat{k}_*}, \widehat{s}^T_{k_2})})\right)^2. \tag{6.34}$$

Furthermore, for all $k_1, k_2 \in \{0, \dots, n_v\}$,

$$\begin{aligned}
\sup_{x\in[0;1]} \left|\gamma(\hat{s}^T_{k_1}, x) - \gamma(\hat{s}^T_{k_2}, x) - c_{k_1,k_2}\right| &\leqslant \sup_{x\in[0;1]} \left| \sum_{j=k_1\wedge k_2+1}^{k_1\vee k_2} \hat{\theta}^T_j \psi_j(x) \right| \\
&\leqslant \sqrt{|k_1 - k_2|}\sqrt{\sum_{j=k_1\wedge k_2+1}^{k_1\vee k_2} \left(\hat{\theta}^T_j\right)^2} \\
&\leqslant \sqrt{n_v}\left(w_1(\sqrt{\ell(\widehat{s}^T_{\hat{k}_*}, \widehat{s}^T_{k_1})}) + w_1(\sqrt{\ell(\widehat{s}^T_{\hat{k}_*}, \widehat{s}^T_{k_2})})\right).
\end{aligned}$$

$u \mapsto \frac{w_1(u)}{u}$ is non-increasing, therefore by Theorem 3.7.2 of Chapter 3, with probability greater than $1 - e^{-y}$ (still conditioning on $D_n^T \in E_n$), for all $\theta \in [0; 1]$,

$$(1-\theta)\ell(\widehat{s}^T_{\hat{k}_*}, \widehat{s}^T_{\hat{k}_T}) \leqslant (1+\theta) \min_{k \in \{0,\dots,n_v\}} \ell(\widehat{s}^T_{\hat{k}_*}, \widehat{s}^T_k) + \delta(w_1, \sqrt{n_v})^2 \left(\theta + \frac{2(y + \log n_v)}{\theta}\right) + \delta(\sqrt{n_v} w_1, n_v)^2 \left(\theta + \frac{(y + \log n_v)^2}{\theta}\right). \tag{6.35}$$

Furthermore, by definition of δ given in chapter 3, $\delta(\sqrt{n_v} w_1, n_v)$ is non-negative and solves the equation:

$$\sqrt{n_v} w_1(\delta) = n_v \delta^2 \iff w_1(\delta) = \sqrt{n_v}\delta^2.$$

Therefore, $\delta(\sqrt{n_v} w_1, n_v) = \delta(w_1, \sqrt{n_v})$, hence by equation (6.35) applied with $y = 2 \log n$ and $\theta = \frac{1}{2}$,

$$\ell(\widehat{s}^T_{\hat{k}_*}, \widehat{s}^T_{\hat{k}_T}) \leqslant 4(1 + 3\log n)^2 \delta(w_1, \sqrt{n_v})^2, \tag{6.36}$$

with probability greater than $1 - \frac{1}{n^2}$. It remains to bound $\delta(w_1, \sqrt{n_v})$. Since w_1 is non-increasing, $\delta(w_1, \sqrt{n_v})$ solves the equation

$$w_1(x) = \sqrt{n_v} x^2$$
$$\sqrt{1 + 12\frac{\mathcal{E}}{\mathfrak{e}}} x + \sqrt{12\mathcal{E}} = \sqrt{\frac{n_v}{\|s\|_\infty}} x^2.$$

Solving this polynomial equation yields a unique positive solution:

$$\delta(w_1, \sqrt{n_v}) = \sqrt{\frac{\|s\|_\infty}{n_v}} \times \left(\sqrt{1 + 12\frac{\mathcal{E}}{\mathfrak{e}}} + \sqrt{1 + 12\frac{\mathcal{E}}{\mathfrak{e}} + 4\sqrt{\frac{12 n_v \mathcal{E}}{\|s\|_\infty}}}\right)$$
$$\leqslant 2\sqrt{\frac{\|s\|_\infty}{n_v}}\sqrt{1 + 12\frac{\mathcal{E}}{\mathfrak{e}}} + 2\left(\frac{12 n_v \mathcal{E}}{\|s\|_\infty}\right)^{\frac{1}{4}} \sqrt{\frac{\|s\|_\infty}{n_v}}.$$

By definition of $\mathcal{E}, \mathfrak{e}$ and lemma 6.6.2, it follows therefore that

$$\begin{aligned}
\delta(w_1, \sqrt{n_v})^2 &\leqslant 8\frac{\|s\|_\infty}{n_v}\left(1 + 12\frac{\mathcal{E}}{\mathfrak{e}}\right) + 8\sqrt{12\mathcal{E}}\sqrt{\frac{\|s\|_\infty}{n_v}} \\
&\leqslant 8\frac{\|s\|_\infty}{n_v} + 96\|s\|_\infty \frac{\mathcal{E}}{n_v \mathfrak{e}} + 8\sqrt{12\|s\|_\infty}\mathfrak{e}. \\
&\leqslant 8\|s\|_\infty \mathfrak{e} + 96\|s\|_\infty \mathfrak{e} + 8\sqrt{12\|s\|_\infty}\mathfrak{e} \\
&\leqslant \left(104\|s\|_\infty + 8\sqrt{12\|s\|_\infty}\right)\mathfrak{e}.
\end{aligned}$$

Given $D_n^T \in E_n$, by equation (6.36), there exists therefore a constant κ such that, with probability greater than $1 - \frac{1}{n^2}$,

$$\ell(\widehat{s}_{\hat{k}_*}^T, \widehat{s}_{\hat{k}_T}^T) \leqslant \kappa \log^2 n \mathfrak{e},$$

for any $n \geqslant n_0$. Since $\mathbb{P}(D_n^T \in E_n) \geqslant 1 - \frac{1}{n^2}$, on the whole, with probability greater than $1 - \frac{2}{n^2}$,

$$\begin{aligned}\left\|\hat{s}_{\hat{k}_T}^T - s\right\|^2 - \left\|\hat{s}_{k_*}^T - s\right\|^2 &\leqslant \left\|\hat{s}_{\hat{k}_T}^T - s\right\|^2 - \left\|\hat{s}_{\hat{k}_*}^T - s\right\|^2 \\ &\leqslant \kappa \log^2 n \mathfrak{e}.\end{aligned}$$

On the same event, by equations (6.29) and (6.31),

$$\kappa \log^2 n \times \mathfrak{e} \geqslant \left\|\hat{s}_{\hat{k}_T}^T - s\right\|^2 - \left\|\hat{s}_{k_*}^T - s\right\|^2 \geqslant \left[\frac{1}{2}\frac{|\hat{k}_T - k_*|}{\Delta} - 1\right]\mathfrak{e},$$

therefore $\frac{|\hat{k}_T - k_*|}{\Delta} \leqslant 2 + 2\kappa \log^2 n$. Furthermore, by equations (6.28) and (6.30),

$$\left\|\hat{s}_{\hat{k}_T}^T - s\right\|^2 - \left\|\hat{s}_{k_*}^T - s\right\|^2 \geqslant \mathfrak{e} f_n\left(\frac{\hat{k}_T - k_*}{\Delta}\right) - \frac{|\hat{k}_T - k_*|}{\Delta} r_n \mathfrak{e} \geqslant \mathfrak{e} f_n\left(\frac{\hat{k}_T - k_*}{\Delta}\right) - \frac{1}{2}\frac{|\hat{k}_T - k_*|}{\Delta}\mathfrak{e}.$$

Hence, on the same event,

$$f_n\left(\frac{\hat{k}_T - k_*}{\Delta}\right) \leqslant \kappa \log^2 n + \frac{1}{2}\left(2 + 2\kappa \log^2 n\right).$$

This proves theorem 6.6.5 with $\kappa_2 = 3\kappa$. ■

6.7 Proof of theorems 6.4.3 and 6.4.4

As in chapter 5, we will call constant any quantity that only depends on $\|s\|_\infty, \|s\|^2$ and the assumptions of section 6.2.2. A constant exponent will be denoted by the letter u, with the understanding that u only depends on the constants $\delta_1, \delta_2, \delta_3, \delta_4, \rho_1$ of section 6.2.2. The letter κ will denote any other constant. Moreover, for integers $a \leqslant b$, we will denote by $[|a; b|]$ the "integer interval" $\mathbb{Z} \cap [a; b]$.

6.7.1 Approximation of $\hat{k}_T^{ho}$ by the argmin of the limit process

.

In the previous chapter, the hold-out risk-estimator was approximated on the interval $[a_x; b_x]$. On the other hand, to study $\widehat{s}_T^{\text{ho}}$ and $\widehat{s}_{n_t,V}^{mc}, \widehat{s}_{n_t,V}^{vf}$, it is the argmin $\hat{k}_T^{ho}$ of this risk estimator that we are interested in. The following result shows that $\hat{k}_T^{ho}$ can be approximated by the argmin of the process $f_n - W_{g_n}$ provided by Theorem 5.3.8 of the previous chapter.

Claim 6.7.0.1 *Let $x > 0$. Let $T \subset \{1 \dots n\}$ be a set of cardinality $n_t = n_t$. Let W and g_n be as given by Theorem 5.3.8 of Chapter 5. Let*

$$\hat{\alpha}_W^{\infty,x} \in \underset{\alpha \in [a_x; b_x]}{\operatorname{argmin}} f_n(\alpha) - W_{g_n(\alpha)}.$$

As above, let

$$\hat{k}_{T,x}^{ho} \in \underset{k \in [\lfloor k_* + a_x \Delta; k_* + b_x \Delta \rfloor]}{\operatorname{argmin}} \left\| \hat{s}_k^T \right\|^2 - 2P_n^{T^c}\left(\hat{s}_k^T\right).$$

There exists a constant $\kappa_4 \geqslant 0$ such that, with probability greater than $1 - \kappa_4(1+x)^{\frac{1}{4}} n^{-\frac{u_1}{6}}$,

$$\left| g_n\left(\frac{\hat{k}_{T,x}^{ho} - k_*}{\Delta}\right) - g_n(\hat{\alpha}_W^{\infty,x}) \right| \leqslant \kappa_4 (1+x)^{\frac{9}{8}} n^{-\frac{u_1}{4}}.$$

In this section, we will prove claim 6.7.0.1. We wish to use the theorem of the previous chapter giving an approximation of the hold-out risk-estimator, in order to get an approximation of the argmin of this process. To control the difference between the argmins of two functions Y and Z, it is sufficient to control $|Y - Z|$ uniformly and to control the size of the intervals on which $Y - \min Y \leqslant \varepsilon$. The following proposition formalizes this reasoning in the case of stochastic processes.

Proposition 6.7.1 *Let Y be an almost surely continuous stochastic process which reaches almost surely a unique minimum on the interval $[a; b]$. Let $\hat{\alpha}_Y = \operatorname{argmin}_{\alpha \in [a;b]} Y_\alpha$ be the point at which this minimum is reached. Assume that there exists constants $\rho > 0, c \geqslant 0$ such that for all $\varepsilon > 0$:*

$$\mathbb{E}\left[\sup\left\{ |\alpha - \hat{\alpha}_Y| \middle| \alpha : Y_\alpha - \min_{\alpha' \in [a;b]} Y_{\alpha'} < \varepsilon \right\}\right] \leqslant c\varepsilon^\rho. \tag{6.37}$$

Then for any other process Z such that $\mathbb{E}\left[\sup_{t \in [a;b]} |Y_t - Z_t|\right] \leqslant \varepsilon$,

$$\mathbb{P}\left(|\hat{\alpha}_Y - \hat{\alpha}_Z| \geqslant c 2^\rho \varepsilon^{\frac{\rho}{2}}\right) \leqslant 2\varepsilon^{\frac{\rho}{2(1+\rho)}},$$

where $\hat{\alpha}_Z \in \operatorname{argmin}_{\alpha \in [a;b]} Y_\alpha$.

Proof For all $\varepsilon > 0$, let

$$\hat{\delta}(\varepsilon) = \sup\left\{ |\alpha - \hat{\alpha}_Y| \middle| \alpha : Y_\alpha - \min_{\alpha' \in [a;b]} Y_{\alpha'} < \varepsilon \right\}.$$

Remark that:

$$\begin{aligned} Y_{\hat{\alpha}_Z} - \min_{\alpha' \in [a;b]} Y_{\alpha'} &= Z_{\hat{\alpha}_Z} + (Y_{\hat{\alpha}_Z} - Z_{\hat{\alpha}_Z}) - \min_{\alpha' \in [a;b]} Z_{\alpha'} + (Y_{\alpha'} - Z_{\alpha'}) \\ &\leqslant Z_{\hat{\alpha}_Z} - \min_{\alpha' \in [a;b]} Z_{\alpha'} + 2 \sup_{\alpha' \in [a;b]} |Y_{\alpha'} - Z_{\alpha'}| \\ &\leqslant 2 \sup_{\alpha' \in [a;b]} |Y_{\alpha'} - Z_{\alpha'}|. \end{aligned}$$

By definition of $\hat{\delta}(\varepsilon)$, it follows that

$$|\hat{\alpha}_Y - \hat{\alpha}_Z| \leqslant \hat{\delta}\left(2 \sup_{\alpha' \in [a;b]} |Y_{\alpha'} - Z_{\alpha'}| \right).$$

As a result, since $\hat{\delta}$ is non-decreasing by definition and by Markov's inequality, for all $y \geqslant 0$,

$$\mathbb{P}\left(|\hat{\alpha}_Y - \hat{\alpha}_Z| \geqslant c 2^\rho y^{1+\rho} \varepsilon^\rho \right) \leqslant \mathbb{P}(\hat{\delta}(2y\varepsilon) \geqslant y \times c 2^\rho y^\rho \varepsilon^\rho) + \mathbb{P}\left(\sup_{\alpha' \in [a;b]} |Y_{\alpha'} - Z_{\alpha'}| \geqslant y\varepsilon \right) \leqslant \frac{2}{y}.$$

We get the stated result by setting $y = \varepsilon^{\frac{-\rho}{2(1+\rho)}}$ in the above equation. ■

Since $g_n(\hat{\alpha}_W^{\infty,x}) = \operatorname{argmin}_{u \in [g_n(a_x); g_n(b_x)]} f_n \circ g_n^{-1}(u) - W_u$, we will now show that $Y = f_n \circ g_n^{-1} - W$ satisfies equation 6.37. This is easier to do than for the hold-out process. In order to control the diameter $\hat{\delta}_Y(\varepsilon)$ of the set $\{t : Y(t) - \min Y \leqslant \varepsilon\}$, we use the following lemma, adapted from the proof of Theorem 3.7 of [88] which treats the case $\varepsilon = 0$ (uniqueness of minimum). This lemma shows that one can bound the expectation of $\hat{\delta}_Y(\varepsilon)$ using the expectation of the minimum of small perturbations of the process Y.

Lemma 6.7.2 *Let G be a random continuous function on the interval $[s;t]$. For all $a \in \mathbb{R}$, let:*

$$M^a = \max_{x \in [s;t]} G(x) + ax.$$

Assume that the function: $m : a \to \mathbb{E}[M^a]$ exists and is differentiable on the interval $]-a_0; a_0[$, where $a_0 > 0$. For all $\varepsilon > 0$, let

$$Z_{1,\varepsilon} = \inf\{x \in [s;t], G(x) \geqslant M - \varepsilon\}, Z_{2,\varepsilon} = \sup\{x \in [s;t], G(x) \geqslant M - \varepsilon\}.$$

Then for all $a \in (0; a_0)$,

$$\mathbb{E}\left[Z_{2,\varepsilon} - Z_{1,\varepsilon}\right] \leqslant \frac{2\varepsilon}{a} + 2 \sup_{x\in[-a;a]} |m'(x) - m'(0)|.$$

Proof For both $i \in \{1; 2\}$,

$$M - \varepsilon + aZ_{i,\varepsilon} \leqslant G(Z_{i,\varepsilon}) + aZ_{i,\varepsilon} \leqslant M^a.$$

It follows that:

$$-\varepsilon \leqslant M^a - M - aZ_{i,\varepsilon}. \tag{6.38}$$

Thus, for all $a > 0$,

$$\begin{aligned} \frac{\varepsilon}{a} &\geqslant \frac{M^{-a} - M}{-a} - Z_{i,\varepsilon} \\ -\frac{\varepsilon}{a} &\leqslant \frac{M^a - M}{a} - Z_{i,\varepsilon} \end{aligned}$$

therefore for all indices $i \in \{1; 2\}$,

$$-\frac{\varepsilon}{a} + \frac{M^{-a} - M}{-a} \leqslant Z_{i,\varepsilon} \leqslant \frac{M^a - M}{a} + \frac{\varepsilon}{a}.$$

In particular,

$$Z_{2,\varepsilon} - Z_{1,\varepsilon} \leqslant \frac{M^a - M}{a} - \frac{M^{-a} - M}{-a} + 2\frac{\varepsilon}{a}.$$

By taking expectations, it follows that

$$\mathbb{E}\left[Z_{2,\varepsilon} - Z_{1,\varepsilon}\right] \leqslant \frac{2\varepsilon}{a} + \frac{m(a) - m(0)}{a} - \frac{m(-a) - m(0)}{-a}.$$

Finally, by the mean value theorem, for all $a \in]-a_0; a_0[$,

$$\mathbb{E}\left[Z_{2,\varepsilon} - Z_{1,\varepsilon}\right] \leqslant \frac{2\varepsilon}{a} + \sup_{(x,y)\in[-a;a]^2} |m'(x) - m'(y)| \leqslant \frac{2\varepsilon}{a} + 2 \sup_{x\in[-a;a]} |m'(x) - m'(0)|.$$

■

For processes $G = f - W$ on an interval $[s; t]$, where W is a Wiener process and f a continuous function, Pimentel [88] shows that $m'(a) = \mathrm{Cov}(W_t - W_s, M^a)$ for all $a > 0$ (this is a consequence of Theorems 1 and 2 of [88]). With this result and lemma 6.7.2 below, it is possible to prove the following claim.

Claim 6.7.2.1 *Let $x > 0$. Let g_n satisfy the properties stated in Theorem 5.3.8 of Chapter 5. For all $u \in [g_n(a_x); g_n(b_x)]$, let $Y : u \mapsto (f_n \circ g_n^{-1})(u) - W_u$, where W is two-sided Brownian motion such that $W_0 = 0$. Let $\hat{u}_Y = \operatorname{argmin}_{u \in [g_n(a_x); g_n(b_x)]} Y_u$ and*

$$\hat{\delta}_Y(\varepsilon) = \sup \left\{ |u - \hat{u}_Y| : Y_u - \min_{\alpha' \in [g_n(a_x); g_n(b_x)]} Y_{u'} \leqslant \varepsilon \right\}.$$

Then there exists an absolute constant κ such that for all $\varepsilon > 0$,

$$\mathbb{E}[\hat{\delta}_Y(\varepsilon)] \leqslant \kappa(1+x)^{\frac{3}{4}}\sqrt{\varepsilon}.$$

Proof Let $\tilde{Y}_u = -Y_u - W_{g_n(a_x)}$, so that $\hat{u}_Y = \operatorname{argmax}_{u \in [g_n(a_x); g_n(b_x)]} \tilde{Y}_u$ and $\tilde{Y}_u = \tilde{f}_n(u) + W_u - W_{g_n(a_x)}$ where $\tilde{f}_n = -f_n \circ g_n^{-1}$. Since f_n is continuous, piecewise linear by definition and g_n^{-1} is Lipschitz continuous by Theorem 5.3.8 of Chapter 5, $\tilde{f}_n$ is locally Lipschitz and so is the function $x \mapsto \tilde{f}_n(x) + ax$, for any $a \in \mathbb{R}$.

For all $a \in \mathbb{R}$, let

$$\begin{aligned} M^a &= \max_{u \in [g_n(a_x); g_n(b_x)]} \tilde{Y}_u + au \\ Z_a &\in \operatorname*{argmax}_{u \in [g_n(a_x); g_n(b_x)]} \tilde{Y}_u + au \\ m(a) &= \mathbb{E}[M^a]. \end{aligned} \tag{6.39}$$

By [88, Theorem 2], Z_a is uniquely defined and

$$\mathbb{E}[Z_a] = g_n(a_x) + \operatorname{Cov}\left(W_{g_n(b_x)} - W_{g_n(a_x)}, M^a\right). \tag{6.40}$$

By [88, Theorem 3.7],

$$\begin{aligned} m'(a) &= \mathbb{E}[Z_a] \\ &= g_n(a_x) + \operatorname{Cov}\left(W_{g_n(b_x)} - W_{g_n(a_x)}, M^a\right) \text{ by equation (6.40).} \end{aligned}$$

Hence:

$$\begin{aligned} |m'(a) - m'(0)| &= \left|\operatorname{Cov}\left(W_{g_n(b_x)} - W_{g_n(a_x)}, M^a - M\right)\right| \\ &\leqslant \sqrt{\operatorname{Var}(W_{g_n(b_x)} - W_{g_n(a_x)})}\sqrt{\operatorname{Var}(M^a - M)} \\ &\leqslant \sqrt{|g_n(b_x) - g_n(a_x)|}\sqrt{\operatorname{Var}(M^a - M)}. \end{aligned} \tag{6.41}$$

Furthermore, by definition (6.39),

$$\begin{aligned} \tilde{Y}_{Z_0} + aZ_0 &= M + aZ_0 \\ &\leqslant M^a \\ &= \max_{u \in [g_n(a_x); g_n(b_x)]} \tilde{Y}_u + au \\ &\leqslant M + \max_{u \in [g_n(a_x); g_n(b_x)]} au. \end{aligned}$$

It follows that $|M^a - M| \leqslant |a| \max(|g_n(a_x)|; |g_n(b_x)|)$, and therefore, by equation (6.41),

$$|m'(a) - m'(0)| \leqslant \max(|g_n(a_x)|; |g_n(b_x)|)\sqrt{|g_n(b_x) - g_n(a_x)|}|a|. \tag{6.42}$$

Since for all $u' \in [g_n(a_x); g_n(b_x)]$ $\tilde{Y}_{u'} = -Y_{u'} - W_{g_n(a_x)}$,

$$Y_u - \min_{u' \in [g_n(a_x); g_n(b_x)]} Y_{u'} \leqslant \varepsilon \iff \tilde{Y}_u - \max_{u' \in [g_n(a_x); g_n(b_x)]} \tilde{Y}_{u'} \geqslant -\varepsilon.$$

By lemma 6.7.2 and equation (6.42), it follows that for all $a \in \mathbb{R}$:

$$\begin{aligned}
\mathbb{E}[\hat{\delta}_Y(\varepsilon)] &= \mathbb{E}\left[\sup\left\{|u - \hat{u}_Y| : Y_u - \min_{u' \in [g_n(a_x); g_n(b_x)]} Y_{u'} \leqslant \varepsilon\right\}\right] \\
&= \mathbb{E}\left[\sup\left\{|u - \hat{u}_Y| : \tilde{Y}_u - \max_{u' \in [g_n(a_x); g_n(b_x)]} \tilde{Y}_{u'} \geqslant -\varepsilon\right\}\right] \\
&\leqslant \frac{2\varepsilon}{a} + 2a \max(|g_n(a_x)|; |g_n(b_x)|)\sqrt{|g_n(a_x) - g_n(a_x)|}.
\end{aligned}$$

Setting $a = \sqrt{\frac{\varepsilon}{\max(|g_n(a_x)|;|g_n(b_x)|)\sqrt{|g_n(b_x) - g_n(a_x)|}}}$, it follows that

$$\mathbb{E}[\hat{\delta}_Y(\varepsilon)] \leqslant 4\sqrt{\varepsilon \max(|g_n(a_x)|; |g_n(b_x)|)}\,(g_n(b_x) - g_n(a_x))^{\frac{1}{4}}.$$

Finally, by point 3. of Theorem 5.3.8 of Chapter 5,

$$\max(|g_n(a_x)|; |g_n(b_x)|) \leqslant 20\,\|s\|_\infty\,(1 + x),$$

which yields the expected result with $\kappa = 4 \times 2^{\frac{1}{4}}(20\,\|s\|_\infty)^{\frac{3}{4}}$. ■

We can now proceed to prove claim 6.7.0.1. In the following, the letter κ will denote constants – which depend only on the assumptions of section 6.2.2 – the value of which can change from line to line. Let

$$Y_u = (f_n \circ g_n^{-1})(u) - W_u.$$

For all $x > 0$, $\hat{\alpha}_W^{\infty,x} = \operatorname{argmin}_{\alpha \in [a_x; b_x]} f_n(\alpha) - W_{g_n(\alpha)}$, in other words $g_n(\hat{\alpha}_W^{\infty,x}) \in \operatorname{argmin}_{u \in [g_n(a_x); g_n(b_x)]} Y_u$. By claim 6.7.2.1, for all $\varepsilon > 0$,

$$\mathbb{E}\left[\sup\left\{|u - g_n(\hat{\alpha}_W^{\infty,x})| : Y_u - \min_{u' \in [g_n(a_x); g_n(b_x)]} Y_{u'} \leqslant \varepsilon\right\}\right] \leqslant \kappa(1 + x)^{\frac{3}{4}}\sqrt{\varepsilon}. \tag{6.43}$$

Furthermore, by Theorem 5.3.8 of Chapter 5, there exists an event E of probability $\mathbb{P}(E) \geqslant 1 - \frac{1}{n^2}$ such that for all $D_n^T \in E$,

$$\mathbb{E}\left[\sup_{u\in[g_n(a_x);g_n(b_x)]}\left|\hat{R}^{ho}(g_n^{-1}(u)) - Y_u\right| \middle| D_n^T\right] = \mathbb{E}\left[\sup_{\alpha\in[a_x;b_x]}\left|\hat{R}^{ho}(\alpha) - Y_{g_n(\alpha)}\right| \middle| D_n^T\right] \leqslant \kappa(1+x)^{\frac{3}{2}} n^{-u_1}, \tag{6.44}$$

where $\hat{R}^{ho}$ denotes the process:

$$\hat{R}^{ho} : \frac{j}{\Delta} \mapsto \frac{1}{\mathfrak{e}}\left(\left\|\hat{s}^T_{k_*+j} - s\right\|^2 - \left\|\hat{s}^T_{k_*} - s\right\|^2\right) - \frac{2}{\mathfrak{e}}\left(P_n^{T^c} - P\right)\left(\hat{s}^T_{k_*+j} - \hat{s}^T_{k_*}\right),$$

extended by linear interpolation. By the definition of $\hat{k}^{ho}_{T,x}$ stated in claim 6.7.0.1, $\frac{\hat{k}^{ho}_{T,x}-k_*}{\Delta} \in \operatorname{argmin}_{\alpha\in[a_x;b_x]} \hat{R}^{ho}(\alpha)$, therefore

$$g_n\left(\frac{\hat{k}^{ho}_{T,x}-k_*}{\Delta}\right) \in \operatorname*{argmin}_{u\in[g_n(a_x);g_n(b_x)]} \hat{R}^{ho}(g_n^{-1}(u)).$$

By equations 6.43 and (6.44), proposition 6.7.1 can be applied conditionally, given $D_n^T \in E$, with $\varepsilon = \kappa(1+x)^{\frac{3}{2}} n^{-u_1}$, $c = \kappa(1+x)^{\frac{3}{4}}$ and $\rho = \frac{1}{2}$. Therefore, for all $D_n^T \in E$,

$$\mathbb{P}\left(|g_n(\tfrac{\hat{k}^{ho}_{T,x}-k_*}{\Delta}) - g_n(\hat{\alpha}_W^{\infty,x})| \geqslant \kappa(1+x)^{\frac{9}{8}} n^{-\frac{u_1}{4}} \middle| D_n^T\right) \leqslant \kappa(1+x)^{\frac{1}{4}} n^{-\frac{u_1}{6}}. \tag{6.45}$$

By the law of total probability,

$$\mathbb{P}\left(|g_n(\tfrac{\hat{k}^{ho}_{T,x}-k_*}{\Delta}) - g_n(\hat{\alpha}_W^{\infty,x})| \geqslant \kappa(1+x)^{\frac{9}{8}} n^{-\frac{u_1}{4}}\right) \leqslant \kappa(1+x)^{\frac{1}{4}} n^{-\frac{u_1}{6}} + \mathbb{P}(E^c) \leqslant \kappa(1+x)^{\frac{1}{4}} n^{-\frac{u_1}{6}} + \frac{1}{n^2}.$$

This proves claim 6.7.0.1 with $u_1 = \min(u_1, 12)$.

6.7.2 A bound on the distributional distance between the argmins

As will be evident in the next section, the risk of the hold-out estimator involves the distribution function of the selected parameter $\hat{k}_T^{ho}$, while the risk of the $\widehat{s}^{vf}_{n_t,V}$ estimator involves the distribution functions of $\min(\hat{k}^{ho}_{T_1}, \hat{k}^{ho}_{T_2})$ and $\max(\hat{k}^{ho}_{T_1}, \hat{k}^{ho}_{T_2})$, where T_1, T_2 are as in Definition 6.3.2. Here, we will try to approximate these distribution functions. This is the purpose of result 6.7.2.2 below.

Claim 6.7.2.2 *Let $T_1, T_2 \subset \{1 \dots n\}$ be two subsets such that $|T_1| = |T_2| = n_t$ and $T_1^c \cap T_2^c = \emptyset$. Let $x > 0$. For $i \in \{1;2\}$, let $\hat{k}^{ho}_{T_i,x}$ be the parameter chosen by minimization of the hold-out risk estimator on the set $[|k_* + a_x\Delta; k_* + b_x\Delta|]$, more precisely,*

$$\hat{k}^{ho}_{T_i,x} = \min \operatorname*{argmin}_{k \in [|k_*+a_x\Delta; k_*+b_x\Delta|]} \left\| \hat{s}_k^{T_i} \right\|^2 - 2\left(P_n^{T_i^c} - P\right)\left(\hat{s}_k^{T_i}\right).$$

Let W_1, W_2 be two independent Wiener processes satisfying equation (5.11) of Theorem 5.3.8 with $T = T_1$ and $T = T_2$, respectively. For both $i \in \{1;2\}$, let $\hat{\alpha}^{\infty,x}_{W_i} = \operatorname{argmin}_{\alpha \in [a_x; b_x]} f_n(\alpha) - W_{g_n(\alpha)}$. There exists a constant $\kappa_5 \geqslant 0$ such that:

$$\mathbb{E}\left[\sup_{y\in\mathbb{R}} \left| \mathbb{P}\left(\hat{k}^{ho}_{T_1,x} \leqslant y | D_n^{T_1}\right) - \mathbb{P}\left(k_* + \hat{\alpha}^{\infty,x}_{W_1}\Delta \leqslant y\right)\right|\right] \leqslant \kappa_5 (1+x)^{\frac{9}{32}} n^{-\frac{u_1}{16}} \tag{6.46}$$

$$\mathbb{E}\left[\sup_{y\in\mathbb{R}} \left| \mathbb{P}\left(\hat{k}^{ho}_{T_1,x} \wedge \hat{k}^{ho}_{T_2,x} \leqslant y | D_n^{T_1\cap T_2}\right) - \mathbb{P}\left(k_* + (\hat{\alpha}^{\infty,x}_{W_1} \wedge \hat{\alpha}^{\infty,x}_{W_2})\Delta \leqslant y\right)\right|\right] \leqslant \kappa_5 (1+x)^{\frac{9}{32}} n^{-\frac{u_1}{16}} \tag{6.47}$$

$$\mathbb{E}\left[\sup_{y\in\mathbb{R}} \left| \mathbb{P}\left(\hat{k}^{ho}_{T_1,x} \vee \hat{k}^{ho}_{T_2,x} \leqslant y | D_n^{T_1\cap T_2}\right) - \mathbb{P}\left(k_* + (\hat{\alpha}^{\infty,x}_{W_1} \vee \hat{\alpha}^{\infty,x}_{W_2})\Delta \leqslant y\right)\right|\right] \leqslant \kappa_5 (1+x)^{\frac{9}{32}} n^{-\frac{u_1}{16}}. \tag{6.48}$$

In the previous section, an upper bound on the gap between $\frac{\hat{k}^{ho}_{T_1,x} - k_*(n_t)}{\Delta}$ and $\hat{\alpha}^{\infty,x}_{W_1}$ has been proven. Lemma 6.7.3 below shows that any such relationship between two random variables implies a uniform upper bound on the distance between their distribution functions, provided that one of the distribution functions has some Hölder regularity.

Lemma 6.7.3 *Assume that for all $(t_1, t_2) \in \mathbb{R}^2$,*

$$|\mathbb{P}(\hat{\alpha} \leqslant t_2) - \mathbb{P}(\hat{\alpha} \leqslant t_1)| \leqslant L|t_2 - t_1|^\gamma.$$

Then if $\hat{\alpha}'$ is another random variable such that $\mathbb{P}\left(|\hat{\alpha} - \hat{\alpha}'| \geqslant \varepsilon\right) \leqslant \delta$,

$$\sup_{\alpha\in\mathbb{R}} |\mathbb{P}(\hat{\alpha} \leqslant \alpha) - \mathbb{P}(\hat{\alpha}' \leqslant \alpha)| \leqslant L\varepsilon^\gamma + 2\delta.$$

Proof For all $x > 1$, let E be the event:

$$E = \left\{|\hat{\alpha} - \hat{\alpha}'| \leqslant \varepsilon\right\}.$$

By definition of E, for all $t \in \mathbb{R}$,

$$\mathbb{I}_E \mathbb{I}_{\hat{\alpha}+\varepsilon \leqslant t} \leqslant \mathbb{I}_E \mathbb{I}_{\hat{\alpha}' \leqslant t} \leqslant \mathbb{I}_E \mathbb{I}_{\hat{\alpha}-\varepsilon \leqslant t}.$$

Taking expectations, it follows that for all $t \in \mathbb{R}$:

$$\mathbb{P}(\hat{\alpha} \leqslant t - \varepsilon) - 2\mathbb{P}(E^c) \leqslant \mathbb{P}(\hat{\alpha}' \leqslant t) \leqslant \mathbb{P}(\hat{\alpha} \leqslant t + \varepsilon) + 2\mathbb{P}(E^c),$$

therefore by Hölder continuity of the function $t \to \mathbb{P}(\hat{\alpha} \leqslant t)$,

$$\mathbb{P}(\hat{\alpha} \leqslant t) - L\varepsilon^\gamma - 2\mathbb{P}(E^c) \leqslant \mathbb{P}(\hat{\alpha}' \leqslant t) \leqslant \mathbb{P}(\hat{\alpha} \leqslant t) + L\varepsilon^\gamma + 2\mathbb{P}(E^c).$$

■

It must now be shown that the argmin distribution function has the Hölder regularity required by the above lemma. The following proposition shows that this is the case for processes of the form $f - W$ (where W is a Wiener process), under some assumptions on f relevant to the present setting.

Proposition 6.7.4 *Let $a < 0 < 1 < b$. Let $f : [a; b] \to \mathbb{R}$ be a continuous function which is non-increasing on $[a; 0]$, non-decreasing on $[0; b]$ and Lipschitz-continuous with constant L on $[0; 1]$. Let W be a two-sided Wiener process. Let $\hat{t} = \operatorname{argmin}_{t\in[a;b]} f(t) + W_t$ (the minimum is a.s. unique by [88, Theorem 2]). Then for all $(t_1, t_2) \in \mathbb{R}^2$,*

$$\left|\mathbb{P}\left(\hat{t} \leqslant t_2\right) - \mathbb{P}\left(\hat{t} \leqslant t_1\right)\right| \leqslant \left[\frac{8}{\pi} + \frac{8L}{\sqrt{\pi}}\right] |t_2 - t_1|^{\frac{1}{4}}. \tag{6.49}$$

Proof Let $\delta \in \left(0; \frac{1}{8}\right)$. Let $x \in]a; b[$. Let $I_\delta = [x - \delta; x + \delta]$. Let $m \in [2; 1/(2\delta)]$.

- If $0 \in [x - \frac{m}{2}\delta; x + \frac{m}{2}\delta]$, let $J_\delta = [x + \frac{m}{2}\delta; x + m\delta]$. Then

$$\begin{aligned}\mathbb{P}(\hat{t} \in I_\delta) &\leqslant \mathbb{P}\left(\min_{t\in I_\delta} f(t) + W_t \leqslant \min_{u\in J_\delta} f(u) + W_u\right)\\ &\leqslant \mathbb{P}\left(f(0) + \min_{t\in I_\delta} W_t \leqslant f(m\delta) + \min_{u\in J_\delta} W_u\right)\end{aligned}$$

  Since $x - \frac{m}{2}\delta \leqslant 0$, $x + m\delta \leqslant \frac{3}{2}m\delta \leqslant \frac{3}{4}$, f is $L-$Lip on $[0; x + m\delta]$. Therefore

$$\mathbb{P}(\hat{t} \in I_\delta) \leqslant \mathbb{P}\left(\min_{t\in I_\delta} W_t - W_{x+\delta} \leqslant Lm\delta + \min_{u\in J_\delta} W_u - W_{x+\delta}\right).$$

- If $0 < x - \frac{m}{2}\delta$, let $J_\delta = [x - \frac{m}{2}\delta; x - \delta]$. Since f is non-decreasing on $[x - \frac{m}{2}\delta; x + \delta]$,

$$\begin{aligned}\mathbb{P}(\hat{t} \in I_\delta) &\leqslant \mathbb{P}\left(\min_{t\in I_\delta} f(t) + W_t \leqslant \min_{u\in J_\delta} f(u) + W_u\right)\\ &\leqslant \mathbb{P}\left(f(x-\delta) - \min_{t\in I_\delta} W_t \leqslant f(x-\delta) + \min_{u\in J_\delta} W_u\right)\\ &\leqslant \mathbb{P}\left(\min_{t\in I_\delta} W_t - W_{x-\delta} \leqslant \min_{u\in J_\delta} W_u - W_{x-\delta}\right).\end{aligned}$$

- If $x+\frac{m}{2}\delta < 0$, symmetrically, let $J_\delta = [x+\delta; x+\frac{m}{2}\delta]$. Since f is non-increasing on $[x-\delta; x+\frac{m}{2}\delta]$,

$$\begin{aligned}\mathbb{P}(\hat{t} \in I_\delta) &\leqslant \mathbb{P}\left(\min_{t\in I_\delta} f(t) + W_t \leqslant \min_{u\in J_\delta} f(u) + W_u\right)\\ &\leqslant \mathbb{P}\left(f(x+\delta) - \min_{t\in I_\delta} W_t \leqslant f(x+\delta) - \min_{u\in J_\delta} W_u\right)\\ &\leqslant \mathbb{P}\left(\min_{t\in I_\delta} W_t - W_{x+\delta} \leqslant \min_{u\in J_\delta} W_u - W_{x+\delta}\right).\end{aligned}$$

Let Y_1, Y_2 be two independent random variables with standard normal distribution $\mathcal{N}(0,1)$. Since I_δ and J_δ are disjoint, of lengths 2δ and $\left(\frac{m}{2}-1\right)\delta$ respectively, in all cases, by the reflexion principle,

$$\begin{aligned}\mathbb{P}(\hat{t} \in I_\delta) &\leqslant \mathbb{P}\left(-\sqrt{2\delta}|Y_1| \leqslant -\sqrt{\left(\tfrac{m}{2}-1\right)\delta}|Y_2| + Lm\delta\right)\\ &\leqslant \mathbb{P}\left(-|Y_1| \leqslant -\frac{1}{2}\sqrt{m-2}|Y_2| + L\frac{m\sqrt{\delta}}{\sqrt{2}}\right)\\ &= \mathbb{P}\left(|Y_1| \geqslant \frac{1}{2}\sqrt{m-2}|Y_2| - L\frac{m\sqrt{\delta}}{\sqrt{2}}\right)\\ &\leqslant \mathbb{P}\left(|Y_1| \geqslant \frac{\sqrt{m-2}}{4}|Y_2|\right) + \mathbb{P}\left(L\frac{m\sqrt{\delta}}{\sqrt{2}} > \frac{\sqrt{m-2}}{4}|Y_2|\right)\\ &\leqslant \mathbb{P}\left(\left|\frac{Y_1}{Y_2}\right| \geqslant \frac{\sqrt{m-2}}{4}\right) + \mathbb{P}\left(|Y_2| \leqslant 2L\sqrt{2\delta}\frac{m}{\sqrt{m-2}}\right).\end{aligned}$$

$\frac{Y_1}{Y_2}$ follows the standard Cauchy distribution, and $Y_2 \sim \mathcal{N}(0,1)$. Moreover, for all $x > 0$, $\int_x^{+\infty} \frac{dt}{\pi(1+t^2)} \leqslant \frac{1}{\pi x}$, therefore:

$$\mathbb{P}\left(\hat{t} \in I_\delta\right) \leqslant \frac{8}{\pi\sqrt{m-2}} + \frac{4\sqrt{\delta}}{\sqrt{\pi}}\frac{Lm}{\sqrt{m-2}}.$$

Let m be such that $m-2 = \frac{1}{\sqrt{\delta}} \geqslant \sqrt{8} \geqslant 2$, then $m \leqslant 2(m-2)$, which yields

$$\mathbb{P}\left(\hat{t} \in I_\delta\right) \leqslant \left[\frac{8}{\pi} + \frac{8L}{\sqrt{\pi}}\right]\delta^{\frac{1}{4}}.$$

Furthermore, $m = 2 + \frac{1}{\sqrt{\delta}}$ satisfies the assumption $m \leqslant \frac{1}{2\delta}$: indeed, since $\delta \leqslant \frac{1}{8}$ by assumption,

$$\frac{1}{2\delta} \geqslant \frac{\sqrt{8}}{2\sqrt{\delta}} \geqslant \frac{1}{2}\times\frac{8}{2} + \frac{1}{2}\times\frac{\sqrt{8}}{2\sqrt{\delta}} \geqslant 2 + \sqrt{\frac{2}{\delta}} \geqslant m.$$

On the other hand, $\mathbb{P}\left(\hat{t} \in I_\delta\right) = \mathbb{P}(\hat{t} \leqslant x+\delta) - \mathbb{P}(\hat{t} \leqslant x-\delta)$. Since no further conditions were imposed on $x \in \mathbb{R}$, $\delta \in \left(0; \frac{1}{8}\right)$, we have thus proved that

$$\forall (t_1,t_2) \in \mathbb{R}^2, |t_2-t_1| \leqslant \frac{1}{8} \implies \left|\mathbb{P}(\hat{t} \leqslant t_2) - \mathbb{P}(\hat{t} \leqslant t_1)\right| \leqslant \left[\frac{8}{\pi} + \frac{8L}{\sqrt{\pi}}\right] |t_2-t_1|^{\frac{1}{4}}.$$

To conclude, it is enough to remark that if $|t_2-t_1| \geqslant \frac{1}{8}$, then $\frac{8}{\pi}|t_2-t_1|^{\frac{1}{4}} \geqslant \frac{8^{\frac{3}{4}}}{\pi} \geqslant 1.5 > 1$, hence equation (6.49) is still true. ■

The above proposition can now be used to prove that the distribution function of

$$g_n(\hat{\alpha}_W^{\infty,x}) = \underset{u \in [g_n(a_x); g_n(b_x)]}{\operatorname{argmin}} f_n \circ g_n^{-1}(u) - W_u$$

has some Hölder regularity. This is the purpose of the following claim.

Claim 6.7.4.1 *For all $x > 0$, let $\hat{\alpha}_W^{\infty,x} = \operatorname{argmin}_{\alpha \in [a_x; b_x]} f_n(\alpha) - W_{g_n(\alpha)}$, where W is a Wiener process such that $W_0 = 0$. Let $x \geqslant 1$. For all $(y_1,y_2) \in \mathbb{R}^2$,*

$$\left|\mathbb{P}\left(g_n(\hat{\alpha}_W^{\infty,x}) \leqslant y_2\right) - \mathbb{P}\left(g_n(\hat{\alpha}_W^{\infty,x}) \leqslant y_1\right)\right| \leqslant 5|y_1-y_2|^{\frac{1}{4}}. \tag{6.50}$$

Furthermore, there exists a constant κ such that for all $j \in [|a_x\Delta; b_x\Delta - 1|]$,

$$\left|\mathbb{P}\left(\hat{\alpha}_W^{\infty,x} \leqslant \frac{j+1}{\Delta}\right) - \mathbb{P}\left(\hat{\alpha}_W^{\infty,x} \leqslant \frac{j}{\Delta}\right)\right| \leqslant \kappa\sqrt{1+x}n^{-\frac{u_2 \wedge \delta_3}{4}}. \tag{6.51}$$

Proof Let $x \geqslant 1$. Proposition 6.7.4 will be applied to the function $f_n \circ g_n^{-1}$ (or to $x \mapsto f_n \circ g_n^{-1}(-x)$). Let us check that all hypotheses are satisfied. f_n is non-increasing on $[a_x; 0]$ and non-decreasing on $[0; b_x]$ and g_n is non-decreasing on $[a_x; b_x]$ therefore $f_n \circ g_n^{-1}$ is non-increasing on $[g_n(a_x); 0]$ and non-decreasing on $[0; g_n(b_x)]$. The same is true for the function $x \mapsto f_n \circ g_n^{-1}(-x)$ on $[-g_n(b_x); -g_n(a_x)]$. It remains to show that these functions are Lipschitz-continuous on $[0;1]$. By point 4 of Theorem 5.3.8, for all $(\alpha_1,\alpha_2) \in [a_x; b_x]^2$,

$$|g_n(\alpha_2) - g_n(\alpha_1)| \geqslant 4\left\|s\right\|^2 |\alpha_2 - \alpha_1| \geqslant 4|\alpha_2-\alpha_1|,$$

therefore g_n^{-1} is $\frac{1}{4}-$ Lipschitz on $[a_x; b_x]$ (by lemma 6.9.7). As for f_n, there are two cases.

- If $\Delta = \Delta_d$, then f_n is 1-Lipschitz on $[0;1]$ and in particular $f_n(1) \leqslant 1$, therefore $b_x \geqslant 1$ since we assumed $x \geqslant 1$. $a_x \leqslant 0$ by definition. Thus $f_n \circ g_n^{-1}$ is $\frac{1}{4}-$Lipschitz on $[0;1] \subset [a_x; b_x]$. Proposition 6.7.4 applies to $f_n \circ g_n^{-1}$ with $L = \frac{1}{4}$ which yields, for all $(y_1,y_2) \in \mathbb{R}^2$,

$$\left|\mathbb{P}\left(g_n(\hat{\alpha}_W^{\infty,x}) \leqslant y_2\right) - \mathbb{P}\left(g_n(\hat{\alpha}_W^{\infty,x}) \leqslant y_1\right)\right| \leqslant 5|y_1-y_2|^{\frac{1}{4}}.$$

- If $\Delta = \Delta_g$, then f_n is 1-Lipschitz on $[-1;0]$ and in particular $f_n(-1) \leqslant 1$, therefore $a_x \leqslant -1$ since we assumed $x \geqslant 1$. $b_x \geqslant 0$ by definition. Hence $x \mapsto f_n \circ g_n^{-1}(-x)$ is $\frac{1}{4}-$Lipschitz on $[0;1] \subset [-b_x; -a_x]$. Proposition 6.7.4 applies to $x \mapsto f_n \circ g_n^{-1}(-x)$ (which reaches its minimum at $-g_n(\hat{\alpha}_W^{\infty,x})$), with $L = \frac{1}{4}$. This yields, for all $(y_1, y_2) \in \mathbb{R}^2$,

$$\left|\mathbb{P}\left(-g_n(\hat{\alpha}_W^{\infty,x}) \leqslant y_2\right) - \mathbb{P}\left(-g_n(\hat{\alpha}_W^{\infty,x}) \leqslant y_1\right)\right| \leqslant 5|y_1 - y_2|^{\frac{1}{4}}.$$

Thus, in all cases, for all $(y_1, y_2) \in \mathbb{R}^2$,

$$\left|\mathbb{P}\left(g_n(\hat{\alpha}_W^{\infty,x}) \leqslant y_2\right) - \mathbb{P}\left(g_n(\hat{\alpha}_W^{\infty,x}) \leqslant y_1\right)\right| \leqslant 5|y_1 - y_2|^{\frac{1}{4}},$$

which proves that equation (6.50) holds true.

Let now $j \in [|a_x\Delta; b_x\Delta - 1|]$. Since g_n is increasing (hence injective),

$$\mathbb{P}\left(\hat{\alpha}_W^{\infty,x} \leqslant \tfrac{j+1}{\Delta}\right) - \mathbb{P}\left(\hat{\alpha}_W^{\infty,x} \leqslant \tfrac{j}{\Delta}\right) = \mathbb{P}\left(g_n(\hat{\alpha}_W^{\infty,x}) \leqslant g_n\left(\tfrac{j+1}{\Delta}\right)\right) - \mathbb{P}\left(g_n(\hat{\alpha}_W^{\infty,x}) \leqslant g_n\left(\tfrac{j}{\Delta}\right)\right).$$

By equation (6.50),

$$\left|\mathbb{P}\left(\hat{\alpha}_W^{\infty,x} \leqslant \tfrac{j+1}{\Delta}\right) - \mathbb{P}\left(\hat{\alpha}_W^{\infty,x} \leqslant \tfrac{j}{\Delta}\right)\right| \leqslant 5\left|g_n\left(\tfrac{j+1}{\Delta}\right) - g_n\left(\tfrac{j}{\Delta}\right)\right|^{\frac{1}{4}}. \tag{6.52}$$

By point 5 of Theorem 5.3.8,

$$\left|g_n\left(\tfrac{j+1}{\Delta}\right) - g_n\left(\tfrac{j}{\Delta}\right)\right| \leqslant 8\|s\|_\infty \left|f_n\left(\tfrac{j+1}{\Delta}\right) - f_n\left(\tfrac{j}{\Delta}\right)\right| + 12\|s\|_\infty \frac{1}{\Delta}.$$

Thus, by claim 6.9.2.1 and lemma 6.6.2,

$$\left|g_n\left(\tfrac{j+1}{\Delta}\right) - g_n\left(\tfrac{j}{\Delta}\right)\right| \leqslant 8\|s\|_\infty \kappa_3(1+x)^2 n^{-u_2} + 12\|s\|_\infty \frac{n - n_t}{n_t}.$$

By hypothesis H5 of section 6.2.2, there exists therefore a constant κ such that

$$\left|g_n\left(\tfrac{j+1}{\Delta}\right) - g_n\left(\tfrac{j}{\Delta}\right)\right| \leqslant \kappa(1+x)^2 n^{-u_2 \wedge \delta_3}.$$

Therefore, by equation (6.52),

$$\left|\mathbb{P}\left(\hat{\alpha}_W^{\infty,x} \leqslant \tfrac{j+1}{\Delta}\right) - \mathbb{P}\left(\hat{\alpha}_W^{\infty,x} \leqslant \tfrac{j}{\Delta}\right)\right| \leqslant 5\kappa^{\frac{1}{4}}\sqrt{1+x}\, n^{-\frac{u_2 \wedge \delta_3}{4}},$$

wich proves equation (6.51). ■

We can now prove claim 6.7.2.2. Since W_1, W_2 are two independent processes, for all $y \in \mathbb{R}$,

$$\mathbb{P}\left(g_n(\hat{\alpha}_{W_1}^{\infty,x} \wedge \hat{\alpha}_{W_2}^{\infty,x}) \leqslant y\right) = 1 - \mathbb{P}\left(g_n(\hat{\alpha}_{W_1}^{\infty,x}) \wedge g_n(\hat{\alpha}_{W_2}^{\infty,x}) > y\right) = 1 - \left[1 - \mathbb{P}\left(g_n(\hat{\alpha}_{W_1}^{\infty,x}) \leqslant y\right)\right]^2.$$

By the same argument, for all $y \in \mathbb{R}$,

$$\mathbb{P}\left(g_n(\hat{\alpha}^{\infty,x}_{W_1} \vee \hat{\alpha}^{\infty,x}_{W_2}) \leqslant y\right) = \mathbb{P}\left(g_n(\hat{\alpha}^{\infty,x}_{W_1}) \vee g_n(\hat{\alpha}^{\infty,x}_{W_2}) \leqslant y\right) = \mathbb{P}\left(g_n(\hat{\alpha}^{\infty,x}_{W_1}) \leqslant y\right)^2.$$

Since the functions $x \mapsto 1-(1-x)^2$ and $x \mapsto x^2$ are $2-$Lipschitz on the interval $[0;1]$, it follows from equation (6.50) of claim 6.7.4.1, that for all $(y_1, y_2) \in \mathbb{R}^2$,

$$\forall i \in \{1;2\} \left|\mathbb{P}\left(g_n(\hat{\alpha}^{\infty,x}_{W_i}) \leqslant y_2\right) - \mathbb{P}\left(g_n(\hat{\alpha}^{\infty,x}_{W_i}) \leqslant y_1\right)\right| \leqslant 5|y_1 - y_2|^{\frac{1}{4}} \tag{6.53}$$

$$\left|\mathbb{P}\left(g_n(\hat{\alpha}^{\infty,x}_{W_1} \wedge \hat{\alpha}^{\infty,x}_{W_2}) \leqslant y_2\right) - \mathbb{P}\left(g_n(\hat{\alpha}^{\infty,x}_{W_1} \wedge \hat{\alpha}^{\infty,x}_{W_2}) \leqslant y_1\right)\right| \leqslant 10|y_1 - y_2|^{\frac{1}{4}} \tag{6.54}$$

$$\left|\mathbb{P}\left(g_n(\hat{\alpha}^{\infty,x}_{W_1} \vee \hat{\alpha}^{\infty,x}_{W_2}) \leqslant y_2\right) - \mathbb{P}\left(g_n(\hat{\alpha}^{\infty,x}_{W_1} \vee \hat{\alpha}^{\infty,x}_{W_2}) \leqslant y_1\right)\right| \leqslant 10|y_1 - y_2|^{\frac{1}{4}}. \tag{6.55}$$

For all $i \in \{1;2\}$, let

$$\hat{\alpha}^{ho}_{T_i,x} = \frac{\hat{k}^{ho}_{T_i,x} - k_*}{\Delta} \in \underset{\alpha \in [a_x;b_x]}{\operatorname{argmin}} \hat{R}^{ho}(\alpha). \tag{6.56}$$

By claim 6.7.0.1,

$$\mathbb{P}\left(\left|g_n(\hat{\alpha}^{\infty,x}_{W_i}) - g_n(\hat{\alpha}^{ho}_{T_i,x})\right| \geqslant \kappa_4(1+x)^{\frac{9}{8}} n^{-\frac{u_1}{4}}\right) \leqslant \kappa_4(1+x)^{\frac{1}{4}} n^{-\frac{u_1}{6}}. \tag{6.57}$$

Since

$$\left|g_n(\hat{\alpha}^{\infty,x}_{W_1} \wedge \hat{\alpha}^{\infty,x}_{W_2}) - g_n(\hat{\alpha}^{ho}_{T_2,x} \wedge \hat{\alpha}^{ho}_{T_2,x})\right| \leqslant \max(|g_n(\hat{\alpha}^{\infty,x}_{W_1}) - g_n(\hat{\alpha}^{ho}_{T_1,x})|, |g_n(\hat{\alpha}^{\infty,x}_{W_1}) - g_n(\hat{\alpha}^{ho}_{T_1,x})|) \tag{6.58}$$

$$\left|g_n(\hat{\alpha}^{\infty,x}_{W_1} \vee \hat{\alpha}^{\infty,x}_{W_2}) - g_n(\hat{\alpha}^{ho}_{T_1,x} \vee \hat{\alpha}^{ho}_{T_2,x})\right| \leqslant \max(|g_n(\hat{\alpha}^{\infty,x}_{W_1}) - g_n(\hat{\alpha}^{ho}_{T_1,x})|, |g_n(\hat{\alpha}^{\infty,x}_{W_2}) - g_n(\hat{\alpha}^{ho}_{T_2,x})|), \tag{6.59}$$

it follows that

$$\mathbb{P}\left(\left|g_n(\hat{\alpha}^{\infty,x}_{W_1} \wedge \hat{\alpha}^{\infty,x}_{W_2}) - g_n(\hat{\alpha}^{ho}_{T_1,x} \wedge \hat{\alpha}^{ho}_{T_2,x})\right| \geqslant \kappa_4(1+x)^{\frac{9}{8}} n^{-\frac{u_1}{4}}\right) \leqslant 2\kappa_4(1+x)^{\frac{1}{4}} n^{-\frac{u_1}{6}} \tag{6.60}$$

$$\mathbb{P}\left(\left|\hat{\alpha}^{\infty,x}_{W_1} \vee \hat{\alpha}^{\infty,x}_{W_2} - \hat{\alpha}^{ho}_{T_1,x} \vee \hat{\alpha}^{ho}_{T_2,x}\right| \geqslant \kappa_4(1+x)^{\frac{9}{8}} n^{-\frac{u_1}{4}}\right) \leqslant 2\kappa_4(1+x)^{\frac{1}{4}} n^{-\frac{u_1}{6}}. \tag{6.61}$$

Let $T_{max} = T_{min} = T_1 \cap T_2$ and define the following events.

$$\begin{aligned}
\text{For any } i \in 1;2, A_i &= \left\{\left|g_n(\hat{\alpha}^{\infty,x}_{W_i}) - g_n(\hat{\alpha}^{ho}_{T_i,x})\right| \geqslant \kappa_4(1+x)^{\frac{9}{8}} n^{-\frac{u_1}{4}}\right\} \\
A_{min} &= \left\{\left|g_n(\hat{\alpha}^{\infty,x}_{W_1} \wedge \hat{\alpha}^{\infty,x}_{W_2}) - g_n(\hat{\alpha}^{ho}_{T_1,x} \wedge \hat{\alpha}^{ho}_{T_2,x})\right| \geqslant \kappa_4(1+x)^{\frac{9}{8}} n^{-\frac{u_1}{4}}\right\} \\
A_{max} &= \left\{\left|g_n(\hat{\alpha}^{\infty,x}_{W_1} \vee \hat{\alpha}^{\infty,x}_{W_2}) - g_n(\hat{\alpha}^{ho}_{T_1,x} \vee \hat{\alpha}^{ho}_{T_2,x})\right| \geqslant \kappa_4(1+x)^{\frac{9}{8}} n^{-\frac{u_1}{4}}\right\}.
\end{aligned} \tag{6.62}$$

Then for all $w \in \{0, 1, min, max\}$,

$$\sqrt{2\kappa_4}(1+x)^{\frac{1}{8}} n^{-\frac{u_1}{12}} \mathbb{P}\left(\mathbb{P}\left(A_w | D_n^{T_w}\right) \geqslant \sqrt{2\kappa_4}(1+x)^{\frac{1}{8}} n^{-\frac{u_1}{12}}\right) \leqslant \mathbb{E}\left[\mathbb{P}\left(A_w | D_n^{T_w}\right)\right] \leqslant 2\kappa_4 (1+x)^{\frac{1}{4}} n^{-\frac{u_1}{6}}.$$

It follows that, for all $w \in \{0, 1, min, max\}$,

$$\mathbb{P}\left(\mathbb{P}\left(A_w | D_n^{T_w}\right) \geqslant \sqrt{2\kappa_4}(1+x)^{\frac{1}{8}} n^{-\frac{u_1}{12}}\right) \leqslant \sqrt{2\kappa_4}(1+x)^{\frac{1}{8}} n^{-\frac{u_1}{12}}.$$

By independence of $\hat{\alpha}_{W_i}^{\infty,x}$ and $D_n^{T_i}$, for all $y \in \mathbb{R}$,

$$\mathbb{P}\left(g_n(\hat{\alpha}_{W_i}^{\infty,x}) \leqslant y | D_n^{T_i}\right) = \mathbb{P}\left(g_n(\hat{\alpha}_{W_i}^{\infty,x}) \leqslant y\right) \tag{6.63}$$

$$\mathbb{P}\left(g_n(\hat{\alpha}_{W_2}^{\infty,x}) \wedge g_n(\hat{\alpha}_{W_2}^{\infty,x}) \leqslant y | D_n^{T_1 \cap T_2}\right) = \mathbb{P}\left(g_n(\hat{\alpha}_{W_2}^{\infty,x}) \wedge g_n(\hat{\alpha}_{W_2}^{\infty,x}) \leqslant y\right) \tag{6.64}$$

$$\mathbb{P}\left(g_n(\hat{\alpha}_{W_2}^{\infty,x}) \vee g_n(\hat{\alpha}_{W_2}^{\infty,x}) \leqslant y | D_n^{T_1 \cap T_2}\right) = \mathbb{P}\left(g_n(\hat{\alpha}_{W_2}^{\infty,x}) \vee g_n(\hat{\alpha}_{W_2}^{\infty,x}) \leqslant y\right). \tag{6.65}$$

By definition of A_w (equation 6.62), equations (6.53), (6.54), (6.55) and lemma 6.7.3 applied conditionally given $D_n^{T_w}$, with $\gamma = \frac{1}{4}$ and $L = 5$ or 10, there exists a constant κ such that

- For $i \in \{1; 2\}$, on the event $\mathbb{P}\left(A_i | D_n^{T_i}\right) \leqslant \sqrt{2\kappa_4}(1+x)^{\frac{1}{8}} n^{-\frac{u_1}{12}}$ of probability greater than $1 - \sqrt{2\kappa_4}(1+x)^{\frac{1}{8}} n^{-\frac{u_1}{12}}$,

$$\sup_{y \in \mathbb{R}} \Big| \mathbb{P}\left(g_n(\hat{\alpha}_{T_i,x}^{ho}) \leqslant y | D_n^{T_i}\right) - \mathbb{P}\left(g_n(\hat{\alpha}_{W_i}^{\infty,x}) \leqslant y\right) \Big| \leqslant \kappa(1+x)^{\frac{9}{32}} n^{-\frac{u_1}{16}} + 2\sqrt{2\kappa_4}(1+x)^{\frac{1}{8}} n^{-\frac{u_1}{12}}.$$

  Hence by bijectivity of $g_n : [a_x; b_x] \to [g_n(a_x); g_n(b_x)]$,

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{y \in [a_x; b_x]} \left|\mathbb{P}\left(\hat{\alpha}_{T_i,x}^{ho} \leqslant y | D_n^{T_i}\right) - \mathbb{P}\left(\hat{\alpha}_{W_i}^{\infty,x} \leqslant y\right)\right|\Big] \\
&= \mathbb{E}\Big[\sup_{y \in [g_n(a_x); g_n(b_x)]} \left|\mathbb{P}\left(g_n(\hat{\alpha}_{T_i,x}^{ho}) \leqslant y | D_n^{T_i}\right) - \mathbb{P}\left(g_n(\hat{\alpha}_{W_i}^{\infty,x}) \leqslant y\right)\right|\Big] \\
&\leqslant \kappa(1+x)^{\frac{9}{32}} n^{-\frac{u_1}{16}} + 2\sqrt{2\kappa_4}(1+x)^{\frac{1}{8}} n^{-\frac{u_1}{12}} \\
&\quad + \mathbb{P}\left(\mathbb{P}\left(A_i | D_n^{T_i}\right) > \sqrt{2\kappa_4}(1+x)^{\frac{1}{8}} n^{-\frac{u_1}{12}}\right) \\
&\leqslant \kappa(1+x)^{\frac{9}{32}} n^{-\frac{u_1}{16}} + 3\sqrt{2\kappa_4}(1+x)^{\frac{1}{8}} n^{-\frac{u_1}{12}}.
\end{aligned}$$

- On the event $\mathbb{P}\left(A_{max} | D_n^{T_1 \cap T_2}\right) \leqslant \sqrt{2\kappa_4}(1+x)^{\frac{1}{8}} n^{-\frac{u_1}{12}}$ of probability greater than $1 - \sqrt{2\kappa_4}(1+x)^{\frac{1}{8}} n^{-\frac{u_1}{12}}$,

$$\begin{aligned}
&\sup_{y \in \mathbb{R}} \left|\mathbb{P}\left(g_n(\hat{\alpha}_{T_1,x}^{ho} \vee \hat{\alpha}_{T_2,x}^{ho}) \leqslant y | D_n^{T_1 \cap T_2}\right) - \mathbb{P}\left(g_n(\hat{\alpha}_{W_1}^{\infty,x} \vee \hat{\alpha}_{W_2}^{\infty,x}) \leqslant y\right)\right| \\
&\leqslant \kappa(1+x)^{\frac{9}{32}} n^{-\frac{u_1}{16}} + 2\sqrt{2\kappa_4}(1+x)^{\frac{1}{8}} n^{-\frac{u_1}{12}}.
\end{aligned}$$

Hence by bijectivity of $g_n : [a_x; b_x] \to [g_n(a_x); g_n(b_x)]$,

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{y\in[a_x;b_x]}\left|\mathbb{P}\left(\hat{\alpha}^{ho}_{T_1,x}\vee\hat{\alpha}^{ho}_{T_2,x}\leqslant y|D_n^{T_1\cap T_2}\right)-\mathbb{P}\left(\hat{\alpha}^{\infty,x}_{W_1}\vee\hat{\alpha}^{\infty,x}_{W_2}\leqslant y\right)\right|\Big]\\
&\quad=\mathbb{E}\Big[\sup_{y\in[g_n(a_x);g_n(b_x)]}\left|\mathbb{P}\left(g_n(\hat{\alpha}^{ho}_{T_1,x}\vee\hat{\alpha}^{ho}_{T_2,x})\leqslant y|D_n^{T_1\cap T_2}\right)-\mathbb{P}\left(g_n(\hat{\alpha}^{\infty,x}_{W_1}\vee\hat{\alpha}^{\infty,x}_{W_2})\leqslant y\right)\right|\Big]\\
&\quad\leqslant\kappa(1+x)^{\frac{9}{32}}n^{-\frac{u_1}{16}}+2\sqrt{2\kappa_4}(1+x)^{\frac{1}{8}}n^{-\frac{u_1}{12}}\\
&\qquad+\mathbb{P}\left(\mathbb{P}\left(A_{max}|D_n^{T_1\cap T_2}\right)>\sqrt{2\kappa_4}(1+x)^{\frac{1}{8}}n^{-\frac{u_1}{12}}\right)\\
&\quad\leqslant\kappa(1+x)^{\frac{9}{32}}n^{-\frac{u_1}{16}}+3\sqrt{2\kappa_4}(1+x)^{\frac{1}{8}}n^{-\frac{u_1}{12}}.
\end{aligned}$$

- The same argument yields the corresponding result for min instead of max.

Claim 6.7.2.2 is obtained by changing variables from y to $k_* + \Delta y$.

6.7.3 Proof of Theorem 6.4.3

We can now prove Theorem 6.4.3.

Definition 6.7.5 *For all $x \in \mathbb{R}_+$ and for all $t \in L^2([0;1])$, let*

$$L_x(t) = \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta}\langle t,\psi_j\rangle\psi_j.$$

For any random variable X, denote by F_X the distribution function of X. Let W be a Wiener process independent from D_n^T and satisfying equation (5.11) of Theorem 5.3.8. Let $x > 0$,

$$\begin{aligned}
\hat{\alpha}^{\infty,x}_W &= \operatorname*{argmin}_{\alpha\in[a_x;b_x]} f_n(\alpha) - W_{g_n(\alpha)},\\
\hat{k}^{\infty}_x &= k_* + \hat{\alpha}^{\infty,x}_W\Delta
\end{aligned}$$

and

$$E_x = \left\{ f_n\left(\frac{\hat{k}^{ho}_T - k_*}{\Delta}\right)\leqslant x \text{ and } k_* + b_x\Delta \leqslant n - n_t\right\}. \tag{6.66}$$

Since $\{k_*+a_x\Delta,\dots,k_*+b_x\Delta\}\subset\{0,\dots,n-n_t\}$ and $\hat{k}^{ho}_T\in\{k_*(n_t)+a_x\Delta,\dots,k_*(n_t)+b_x\Delta\}$ on E_x by definition of a_x, b_x, $\hat{k}^{ho}_T = \hat{k}^{ho}_{T,x}$ on E_x. Thus, on E_x,

$$\begin{aligned}
\left\|\hat{s}^T_{\hat{k}^{ho}_T}-s\right\|^2 &= \sum_{j=1}^{k_*+a_x\Delta}\left(\hat{\theta}^T_j-\theta_j\right)^2+\sum_{j=k_*+b_x\Delta+1}^{+\infty}\theta_j^2+\left\|L_x(\hat{s}^T_{\hat{k}^{ho}_{T,x}}-s)\right\|^2\\
&= \left\|\hat{s}^T_{k_*}-s\right\|^2+\left\|L_x(\hat{s}^T_{\hat{k}^{ho}_{T,x}}-s)\right\|^2-\left\|L_x(\hat{s}^T_{k_*}-s)\right\|^2,
\end{aligned}$$

so, taking expectations, by lemma 6.9.1

$$\mathbb{E}\left[\left\|\hat{s}^T_{\hat{k}^{ho}_T} - s\right\|^2 \mathbb{I}_{E_x}\right] = \sum_{j=1}^{k_*} \frac{\mathrm{Var}(\psi_j)}{n_t} + \sum_{j=k_*+1}^{+\infty} \theta_j^2 + \mathbb{E}\left[\left\|L_x(\hat{s}^T_{\hat{k}^{ho}_{T,x}} - s)\right\|^2 \mathbb{I}_{E_x}\right]$$
$$- \mathbb{E}\left[\left\|L_x(\hat{s}^T_{k_*} - s)\right\|^2 \mathbb{I}_{E_x}\right] \tag{6.67}$$
$$= \mathrm{or}(n_t) + \mathbb{E}\left[\left(\left\|L_x(\hat{s}^T_{\hat{k}^{ho}_{T,x}} - s)\right\|^2 - \left\|L_x(\hat{s}^T_{k_*} - s)\right\|^2\right)\mathbb{I}_{E_x}\right] \pm \frac{\|\theta\|_{\ell^1}}{n_t}. \tag{6.68}$$

Since

$$\left\|L_x(\hat{s}^T_{\hat{k}^{ho}_{T,x}} - s)\right\|^2 - \left\|L_x(\hat{s}^T_{k_*} - s)\right\|^2 \leqslant \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \theta_j^2 + \left(\hat{\theta}^T_j - \theta_j\right)^2,$$

by claim 6.9.4.1, for all $n \geqslant n_1$,

$$\mathbb{E}\left[\left(\left\|L_x(\hat{s}^T_{\hat{k}^{ho}_{T,x}} - s)\right\|^2 - \left\|L_x(\hat{s}^T_{k_*} - s)\right\|^2\right)\mathbb{I}_{E_x^c}\right] \leqslant \mathbb{P}(E_x^c) \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \theta_j^2 + \frac{3}{2}\mathbb{P}(E_x^c)(b_x - a_x)\mathcal{E}$$
$$\leqslant 7(1+x)\mathbb{P}(E_x^c) \text{ by lemma 6.6.4.}$$

Thus,

$$\left|\mathbb{E}\left[\left(\left\|L_x(\hat{s}^T_{\hat{k}^{ho}_{T,x}} - s)\right\|^2 - \left\|L_x(\hat{s}^T_{k_*} - s)\right\|^2\right)\mathbb{I}_{E_x}\right] - \mathbb{E}\left[\left\|L_x(\hat{s}^T_{\hat{k}^{ho}_{T,x}} - s)\right\|^2 - \left\|L_x(\hat{s}^T_{k_*} - s)\right\|^2\right]\right|$$
$$\leqslant 7(1+x)\mathbb{P}(E_x^c). \tag{6.69}$$

We will now approximate $\left\|L_x(\hat{s}^T_{\hat{k}^{ho}_{T,x}} - s)\right\|^2$ by $\left\|L_x(\hat{s}^T_{\hat{k}^{\infty}_{x}} - s)\right\|^2$, through the following claim.

Claim 6.7.5.1 *There exists two constants $\kappa_6 \geqslant 0, u_6 > 0$ such that for all $x > 0$,*

$$\left|\mathbb{E}\left[\left\|L_x(\hat{s}^T_{\hat{k}^{ho}_{T,x}} - s)\right\|^2\right] - \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \theta_j^2 F_{\hat{k}^{\infty}_x}(j-1) + \frac{1}{n_t}[1 - F_{\hat{k}^{\infty}_x}(j-1)]\right| \leqslant \kappa_6(1+x)^{\frac{41}{32}} n^{-u_6}\mathfrak{e}.$$

Proof

$$\left|\left\|L_x(\hat{s}^T_{\hat{k}^{ho}_{T,x}} - s)\right\|^2 - \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \theta_j^2 \mathbb{I}_{\hat{k}^{ho}_{T,x}<j} + \frac{\mathbb{I}_{\hat{k}^{ho}_{T,x}\geqslant j}}{n_t}\right| \leqslant \left|\sum_{j=k_*+a_x\Delta+1}^{\hat{k}^{ho}_{T,x}} \left(\hat{\theta}^T_j - \theta_j\right)^2 - \frac{1}{n_t}\right|$$
$$\leqslant \left|\sum_{j=k_*+a_x\Delta+1}^{b_x\Delta} \left(\hat{\theta}^T_j - \theta_j\right)^2 - \frac{1}{n_t}\right|,$$

hence by proposition 6.9.2 and lemma 6.9.4 with $Z = 1$, $\varepsilon = 0$, $\delta = 1$,

$$\mathbb{E}\left[\left|\left\|L_x(\hat{s}^T_{\hat{k}^{ho}_{T,x}} - s)\right\|^2 - \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \theta_j^2 \mathbb{I}_{\hat{k}^{ho}_{T,x}<j} + \frac{\mathbb{I}_{\hat{k}^{ho}_{T,x}\geqslant j}}{n_t}\right|\right] \leqslant \kappa_1 \mathfrak{e}(b_x - a_x) n^{-\min(\frac{1}{12}, \frac{\delta_3}{2})}$$
$$\times \int_0^{+\infty} (y + \log n)^2 e^{-y} dy$$
$$\leqslant \kappa(1+x)\log^2(n) n^{-\min(\frac{1}{12}, \frac{\delta_3}{2})}\mathfrak{e} \tag{6.70}$$

for some constant κ, by lemma 6.6.4. Furthermore, denote by $\hat{F}_{\hat{k}^{ho}_{T,x}}$ the conditional distribution function $y \mapsto \mathbb{P}\left(\hat{k}^{ho}_{T,x} \leqslant y | D^T_n\right)$. Since $\hat{k}^{ho}_{T,x}$ is an integer,

$$\mathbb{E}\left[\sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \theta_j^2 \mathbb{I}_{\hat{k}^{ho}_{T,x}<j} + \frac{\mathbb{I}_{\hat{k}^{ho}_{T,x}\geqslant j}}{n_t}\right] = \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \theta_j^2 \hat{F}_{\hat{k}^{ho}_{T,x}}(j-1) + \frac{1}{n_t}[1 - \hat{F}_{\hat{k}^{ho}_{T,x}}(j-1)], \tag{6.71}$$

therefore by equation (6.70),

$$\left|\mathbb{E}\left[\left\|L_x(\hat{s}^T_{\hat{k}^{ho}_{T,x}} - s)\right\|^2\right] - \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \theta_j^2 F_{\hat{k}^{\infty}_x}(j-1) + \frac{1}{n_t}[1 - F_{\hat{k}^{\infty}_x}(j-1)]\right|$$
$$\leqslant \kappa(1+x)\log^2 n n^{-\min(\frac{1}{12}, \frac{\delta_3}{2})}\mathfrak{e} + \mathbb{E}\left[\left|\sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} [\hat{F}_{\hat{k}^{ho}_{T,x}} - F_{\hat{k}^{\infty}_x}](j-1)\left(\theta_j^2 - \tfrac{1}{n_t}\right)\right|\right]$$
$$\leqslant \kappa(1+x)\log^2 n n^{-\min(\frac{1}{12}, \frac{\delta_3}{2})}\mathfrak{e} + \mathbb{E}\left[\left\|\hat{F}_{\hat{k}^{ho}_{T,x}} - F_{\hat{k}^{\infty}_x}\right\|_\infty\right] \mathfrak{e}[f_n(a_x) + f_n(b_x)]. \tag{6.72}$$

Furthermore, by claim 6.7.2.2,

$$\mathbb{E}\left[\left\|\hat{F}_{\hat{k}^{ho}_{T,x}} - F_{\hat{k}^{\infty}_x}\right\|_\infty\right] \leqslant \kappa_5(1+x)^{\frac{9}{32}} n^{-\frac{u_1}{16}}. \tag{6.73}$$

It follows from equation (6.72) that

$$\left|\mathbb{E}\left[\left\|L_x(\hat{s}^T_{\hat{k}^{ho}_{T,x}} - s)\right\|^2\right] - \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \theta_j^2 F_{\hat{k}_x^\infty}(j-1) + \frac{1}{n_t}[1 - F_{\hat{k}_x^\infty}(j-1)]\right|$$
$$\leqslant \kappa(1+x)\log^2 n n^{-\min(\frac{1}{12},\frac{\delta_3}{2})}\mathfrak{e} + 2\kappa_5(1+x)^{\frac{41}{32}} n^{-\frac{u_1}{16}}\mathfrak{e}.$$

This proves claim 6.7.5.1 for all $u < \min\left(\frac{1}{12}, \frac{\delta_3}{2}, \frac{u_1}{16}\right)$. ■

By lemma 6.9.1 and definition 6.7.5,

$$\begin{aligned}\mathbb{E}\left[\left\|L_x(\hat{s}^T_{k_*} - s)\right\|^2\right] &= \sum_{j=k_*+a_x\Delta+1}^{k_*} \frac{\mathrm{Var}(\psi_j)}{n_t} + \sum_{j=k_*+1}^{k_*+b_x\Delta} \theta_j^2 \\ &= \sum_{j=k_*+a_x\Delta+1}^{k_*} \frac{1}{n_t} + \sum_{j=k_*+1}^{k_*+b_x\Delta} \theta_j^2 \pm \frac{\|\theta\|_{\ell^1}}{n_t},\end{aligned}$$

it follows that

$$\begin{aligned}&\sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \theta_j^2 F_{\hat{k}_x^\infty}(j-1) + \frac{1}{n_t}[1 - F_{\hat{k}_x^\infty}(j-1)] - \mathbb{E}\left[\left\|L_x(\hat{s}^T_{k_*} - s)\right\|^2\right] \pm \frac{\|\theta\|_{\ell^1}}{n_t} \\ &= \sum_{j=k_*+a_x\Delta+1}^{k_*} F_{\hat{k}_x^\infty}(j-1)\left[\theta_j^2 - \frac{1}{n_t}\right] + \sum_{j=k_*+1}^{k_*+b_x\Delta} [1 - F_{\hat{k}_x^\infty}(j-1)]\left[\theta_j^2 - \frac{1}{n_t}\right].\end{aligned}$$

Since $\hat{k}_x^\infty = k_* + \hat{\alpha}_W^{\infty,x}\Delta$, $F_{\hat{k}_x^\infty}(j-1) = F_{\hat{\alpha}_W^{\infty,x}}\left(\frac{j-1-k_*}{\Delta}\right)$. Changing variables from $j = k_* - i + 1$ to i for $j \leqslant k_*$ and from $j = k_* + i + 1$ to i for $j > k_*$, it follows therefore that:

$$\begin{aligned}&\sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \theta_j^2 F_{\hat{k}_x^\infty}(j-1) + \frac{1}{n_t}[1 - F_{\hat{k}_x^\infty}(j-1)] - \mathbb{E}\left[\left\|L_x(\hat{s}^T_{k_*} - s)\right\|^2\right] \pm \frac{\|\theta\|_{\ell^1}}{n_t} \\ &= \sum_{i=1}^{-a_x\Delta} F_{\hat{\alpha}_W^{\infty,x}}\left(\tfrac{-i}{\Delta}\right)\left[\theta^2_{k_*-i+1} - \frac{1}{n_t}\right] + \sum_{i=0}^{b_x\Delta-1}\left[1 - F_{\hat{\alpha}_W^{\infty,x}}\left(\tfrac{i}{\Delta}\right)\right]\left[\theta^2_{k_*+i+1} - \frac{1}{n_t}\right].\end{aligned} \tag{6.74}$$

Furthermore, since $\hat{\alpha}_W^{\infty,x} \in [a_x; b_x]$,

$$\begin{aligned}
\mathbb{E}\left[f_n(\hat{\alpha}_W^{\infty,x})\right] &= \int_{a_x}^{0} f_n(\alpha) dF_{\hat{\alpha}_W^{\infty,x}}(\alpha) + \int_{0}^{b_x} f_n(\alpha) dF_{\hat{\alpha}_W^{\infty,x}}(\alpha) \\
&= \left[F_{\hat{\alpha}_W^{\infty,x}} f_n\right]_{a_x}^{0} - \int_{a_x}^{0} f_n'(\alpha) F_{\hat{\alpha}_W^{\infty,x}}(\alpha) d\alpha \\
&\quad + \left[-(1 - F_{\hat{\alpha}_W^{\infty,x}}) f_n\right]_{0}^{b_x} + \int_{0}^{b_x} f_n'(\alpha)[1 - F_{\hat{\alpha}_W^{\infty,x}}](\alpha) d\alpha \\
&= -\int_{a_x}^{0} f_n'(\alpha) F_{\hat{\alpha}_W^{\infty,x}}(\alpha) d\alpha + \int_{0}^{b_x} f_n'(\alpha)[1 - F_{\hat{\alpha}_W^{\infty,x}}](\alpha) d\alpha. \qquad (6.75)
\end{aligned}$$

On the other hand, for all $j \in [|a_x\Delta; b_x\Delta - 1|]$ and for all $\alpha \in \left[\frac{j}{\Delta}; \frac{j+1}{\Delta}\right]$, by Definition 6.4.2,

$$\mathfrak{e} f_n'(\alpha) = \Delta\left[R(k_* + j + 1) - R(k_* + j)\right] + \frac{\Delta}{n_t}$$

where $R(k) = \sum_{i=k+1}^{+\infty} \theta_i^2$. It follows that

$$\mathfrak{e} f_n'(\alpha) = \Delta\left[\frac{1}{n_t} - \theta_{k_*+j+1}^2\right],$$

therefore for all $\alpha \in \left]\frac{j}{\Delta}; \frac{j+1}{\Delta}\right[$,

$$f_n'(\alpha) = \frac{\Delta}{\mathfrak{e}}\left[\frac{1}{n_t} - \theta_{k_*+j+1}^2\right].$$

Thus, by equation (6.75):

$$\begin{aligned}
\mathfrak{e}\mathbb{E}\left[f_n(\hat{\alpha}_W^{\infty,x})\right] = &\sum_{j=0}^{b_x\Delta-1} \left[\frac{1}{n_t} - \theta_{k_*+j+1}^2\right] \Delta \int_{\frac{j}{\Delta}}^{\frac{j+1}{\Delta}} [1 - F_{\hat{\alpha}_W^{\infty,x}}](\alpha) d\alpha \\
&+ \sum_{j=1}^{-a_x\Delta} \left[\theta_{k_*-j+1}^2 - \frac{1}{n_t}\right] \Delta \int_{\frac{-j}{\Delta}}^{\frac{-j+1}{\Delta}} F_{\hat{\alpha}_W^{\infty,x}}(\alpha) d\alpha. \qquad (6.76)
\end{aligned}$$

Since for all $j \in [|a_x\Delta; b_x\Delta - 1|]$, by claim 6.7.4.1

$$\left|\Delta \int_{\frac{j}{\Delta}}^{\frac{j+1}{\Delta}} F_{\hat{\alpha}_W^{\infty,x}}(\alpha) d\alpha - F_{\hat{\alpha}_W^{\infty,x}}\left(\tfrac{j}{\Delta}\right)\right| \leqslant \left|F_{\hat{\alpha}_W^{\infty,x}}\left(\tfrac{j+1}{\Delta}\right) - F_{\hat{\alpha}_W^{\infty,x}}\left(\tfrac{j}{\Delta}\right)\right| \leqslant \kappa\sqrt{1+x} n^{-\frac{u_2 \wedge \delta_3}{4}},$$

by equation (6.74), equation (6.76) and claim 6.7.5.1, it follows that for all $x > 0$,

$$\begin{aligned}
&\left|\mathbb{E}\left[\left\|L_x(\hat{s}^T_{\hat{k}^{ho}_{T,x}} - s)\right\|^2 - \left\|L_x(\hat{s}^T_{k_*} - s)\right\|^2\right] - \mathfrak{e}\mathbb{E}\left[f_n(\hat{\alpha}^{\infty,x}_W)\right]\right| \\
&\quad\leqslant \kappa_6(1+x)^{\frac{41}{32}} n^{-u_6}\mathfrak{e} + \kappa\sqrt{1+x}\, n^{-\frac{u_2\wedge\delta_3}{4}} \sum_{j=a_x\Delta+1}^{b_x\Delta} \left|\theta^2_{k_*+j} - \frac{1}{n_t}\right| + \frac{\|\theta\|_{\ell^1}}{n_t} \\
&\quad\leqslant \kappa_6(1+x)^{\frac{41}{32}} n^{-u_6}\mathfrak{e} + \kappa\sqrt{1+x}\, n^{-\frac{u_2\wedge\delta_3}{4}} \mathfrak{e}[f_n(a_x) + f_n(b_x)] + \|\theta\|_{\ell^1} \frac{n-n_t}{n_t}\frac{1}{n-n_t} \\
&\quad\leqslant \kappa(1+x)^{\frac{3}{2}} n^{-u}\mathfrak{e},
\end{aligned}$$

where $u = \min\left(u_6, \frac{u_2}{4}, \frac{\delta_3}{4}\right)$. By equation (6.69) and equation (6.68), it follows that for all $n \geqslant n_1$ and all $x \geqslant 0$,

$$\left|\mathbb{E}\left[\left\|\hat{s}^T_{\hat{k}^{ho}_T} - s\right\|^2 \mathbb{I}_{E_x}\right] - \mathrm{or}(n_t) - \mathfrak{e}\mathbb{E}\left[f_n(\hat{\alpha}^{\infty,x}_W)\right]\right| \leqslant 7(1+x)\mathbb{P}(E_x^c) + \kappa(1+x)^{\frac{3}{2}} n^{-u}\mathfrak{e} + \frac{\|\theta\|_{\ell^1}}{n_t}. \tag{6.77}$$

Furthermore, since by definition $\hat{k}^{ho}_T \in \{1, \ldots, n-n_t\}$ and $n - n_t \leqslant n_t$,

$$\left\|\hat{s}^{T_1}_{\hat{k}^{ho}_T} - s\right\|^2 \leqslant \sum_{j=1}^{n_t} \theta_j^2 + \left(\hat{\theta}^{T_1}_j - \theta_j\right)^2.$$

Hence by claim 6.9.4.1 applied with $\alpha_1 = \frac{-k_*}{\Delta}, \alpha_2 = \frac{n_t - k_*}{\Delta}$, for all $n \geqslant n_1$,

$$\begin{aligned}
\mathbb{E}\left[\mathbb{I}_{E_x^c}\left\|\hat{s}^{T_1}_{\hat{k}^{ho}_T} - s\right\|^2\right] &\leqslant \mathbb{P}(E_x^c)\|s\|^2 + \frac{3}{2}\mathbb{P}(E_x^c)\frac{n_t}{\Delta}\mathcal{E} \\
&\leqslant \mathbb{P}(E_x^c)\left[\|s\|^2 + \tfrac{3}{2}\right].
\end{aligned}$$

By equation (6.77), there exists therefore a constant κ such that for all $n \geqslant n_1$ and all $x > 0$,

$$\left|\mathbb{E}\left[\left\|\hat{s}^T_{\hat{k}^{ho}_T} - s\right\|^2\right] - \mathrm{or}(n_t) - \mathfrak{e}\mathbb{E}\left[f_n(\hat{\alpha}^{\infty,x}_W)\right]\right| \leqslant \kappa(1+x)\mathbb{P}(E_x^c) + \kappa(1+x)^{\frac{3}{2}} n^{-u}\mathfrak{e} + \frac{\|\theta\|_{\ell^1}}{n_t}. \tag{6.78}$$

Since by definition,

$$\begin{aligned}
\hat{\alpha}^{\infty}_W &= \operatorname*{argmin}_{\alpha \in \left[\frac{-k_*(n_t)}{\Delta}; +\infty\right[} \left\{f_n(\alpha) - W_{g_n(\alpha)}\right\} \\
\hat{\alpha}^{\infty,x}_W &= \operatorname*{argmin}_{\alpha \in [a_x; b_x]} \left\{f_n(\alpha) - W_{g_n(\alpha)}\right\},
\end{aligned}$$

$a_x \geqslant \frac{-k_*}{\Delta}$ and $f_n(\alpha) \leqslant x$ for all $\alpha \in [a_x; b_x]$,

$$\mathbb{E}\left[f_n(\hat{\alpha}^{\infty}_W)\mathbb{I}_{\hat{\alpha}^{\infty}_W \in [a_x; b_x]}\right] \leqslant \mathbb{E}\left[f_n(\hat{\alpha}^{\infty,x}_W)\right] \leqslant \mathbb{E}\left[f_n(\hat{\alpha}^{\infty}_W)\right] + x\mathbb{P}\left(\hat{\alpha}^{\infty}_W \notin [a_x; b_x]\right). \tag{6.79}$$

By claim 6.9.3.1, for all $n \geqslant n_2$ and all $x \geqslant 2\kappa_7(1+\log n)$, $\mathbb{P}(\hat{\alpha}_W^\infty \notin [a_x; b_x]) \leqslant \frac{1}{n}$. Hence, by equation (6.163) of claim 6.9.3.1, lemma 6.9.4 applies with $\varepsilon = 0$, $\delta = \frac{1}{n}$ and $f : x \mapsto \kappa_7(1+x)$, which yields, for all $n \geqslant n_2$ and all $x \geqslant 2\kappa_7(1+\log n)$,

$$\begin{aligned}\mathbb{E}\left[f_n(\hat{\alpha}_W^\infty)\mathbb{I}_{\hat{\alpha}_W^\infty \notin [a_x;b_x]}\right] &\leqslant \int_{\log n}^{+\infty} \kappa_7(1+y)e^{-y}dy \\ &\leqslant \frac{\kappa_7}{n} + \left[-\kappa_7 y e^{-y}\right]_{\log n}^{+\infty} + \int_{\log n}^{+\infty} \kappa_7 e^{-y}dy \\ &\leqslant \frac{\kappa_7 + \kappa_7 + 2\kappa_7 \log n}{n}. \end{aligned} \tag{6.80}$$

Thus, by equation (6.79), for all $n \geqslant \max(2, n_2)$ and all $x \geqslant 2\kappa_7(1+\log n)$,

$$\left|\mathbb{E}\left[f_n(\hat{\alpha}_W^{\infty,x})\right] - \mathbb{E}\left[f_n(\hat{\alpha}_W^\infty)\right]\right| \leqslant 4\kappa_7 \frac{\log n}{n} + \frac{x}{n}. \tag{6.81}$$

Let $x = x_n = \max\left(2\kappa_7(1+\log n), \kappa_2 \log^2 n\right)$. By claim 6.9.2.2, for all $n \geqslant \max(n_3, \kappa_9(1+x_n)^3)$, $k_* + b_x\Delta \leqslant n - n_t$. As x_n is of order $\log^2 n$, there exists n_4 such that for all $n \geqslant n_4$, $n \geqslant \max(n_3, \kappa_9(1+x_n)^3)$, hence $k_* + b_x\Delta \leqslant n - n_t$. Finally, it follows from Theorem 6.6.5 that for all $n \geqslant \max(n_0, n_4)$, $\mathbb{P}(E_{x_n}^c) \leqslant \frac{2}{n^2}$. In conclusion, by setting $x = x_n$, equations (6.78) and (6.81) yield a constant κ such that for all $n \geqslant \max(2, n_0, n_1, n_2, n_4)$,

$$\left|\mathbb{E}\left[\left\|\hat{s}^T_{\hat{k}_T^{ho}} - s\right\|^2\right] - \mathrm{or}(n_t) - \mathfrak{e}\mathbb{E}\left[f_n(\hat{\alpha}_W^\infty)\right]\right| \leqslant \kappa \frac{\log^2 n}{n}\mathfrak{e} + \kappa \frac{\log^2(n)}{n^2} + \kappa \log^3(n) n^{-u}\mathfrak{e} + \frac{\|\theta\|_{\ell^1}}{n_t}.$$

Since $\frac{1}{n_t} = \frac{n-n_t}{n_t}\frac{1}{n-n_t} \leqslant n^{-\delta_3}\mathfrak{e}$, this proves theorem 6.4.3.

6.7.4 Proof of theorem 6.4.4

For all $i \in \{1, \ldots, V\}$ and all $x > 0$, let

$$\begin{aligned}\hat{k}_i^{ho} = \hat{k}_{T_i}^{ho} &= \min \operatorname*{argmin}_{k \in \{1,\ldots,n-n_t\}} \left\|\hat{s}_k^{T_i}\right\|^2 - 2P_n^{T_i^c}(\hat{s}_k^{T_i}) \\ \hat{k}_{i,x}^{ho} = \hat{k}_{T_i,x}^{ho} &= \min \operatorname*{argmin}_{k \in \{k_*+a_x\Delta,\ldots,k_*+b_x\Delta\}} \left\|\hat{s}_k^{T_i}\right\|^2 - 2P_n^{T_i^c}(\hat{s}_k^{T_i}).\end{aligned}$$

Thus, $\widehat{s}^{\mathrm{ho}}_{T_i} = \hat{s}^{T_i}_{\hat{k}_i^{ho}}$ for all $i \in [|1; V|]$, and $\widehat{s}^{vf}_{n_t,V} = \frac{1}{V}\sum_{i=1}^V \hat{s}^{T_i}_{\hat{k}_i^{ho}}$. Let $x > 0$ and let

$$E_x = \cap_{i=1}^V \left\{ f_n\left(\frac{\hat{k}_i^{ho} - k_*}{\Delta}\right) \leqslant x \text{ and } k_* + b_x\Delta \leqslant n - n_t \right\}. \tag{6.82}$$

By definition of a_x, b_x, on the event E_x, $\{k_* + a_x\Delta, \dots, k_* + b_x\Delta\} \subset \{1, \dots, n - n_t\}$ and for all $i \in [|1; V|]$, $\frac{\hat{k}_i^{ho} - k_*}{\Delta} \in [a_x; b_x]$, therefore $\hat{k}_i^{ho} = \hat{k}_{i,x}^{ho}$. As a consequence, for all $j \leqslant k_* + a_x\Delta$, on E_x,

$$\langle \widehat{s}^{vf}_{n_t,V}, \psi_j \rangle = \frac{1}{V} \sum_{i=1}^{V} \hat{\theta}_j^{T_i},$$

and for all $j \geqslant k_* + b_x\Delta + 1$, $\langle \widehat{s}^{vf}_{n_t,V}, \psi_j \rangle = 0$. Therefore, on the event E_x,

$$\left\| \widehat{s}^{vf}_{n_t,V} - s \right\|^2 = \sum_{j=1}^{k_* + a_x\Delta} \left(\frac{1}{V} \sum_{i=1}^{V} \hat{\theta}_j^{T_i} - \theta_j \right)^2 + \left\| L_x(\widehat{s}^{vf}_{n_t,V}) - L_x(s) \right\|^2 + \sum_{j \geqslant k_* + b_x\Delta + 1} \theta_j^2.$$

By linearity of L_x, $L_x(\widehat{s}^{vf}_{n_t,V}) = \frac{1}{V} \sum_{i=1}^{V} L_x(\widehat{s}^{\mathrm{ho}}_{T_i})$. The random variables $L_x(\widehat{s}^{\mathrm{ho}}_{T_i}) \mathbb{I}_{E_x}$ are exchangeable, therefore by lemma 6.9.5,

$$\begin{aligned} \mathbb{E}\left[\left\| \widehat{s}^{vf}_{n_t,V} - s \right\|^2 \mathbb{I}_{E_x} \right] = \mathbb{E}\left[\mathbb{I}_{E_x} \sum_{j=1}^{k_* + a_x\Delta} \left(\frac{1}{V} \sum_{i=1}^{V} \hat{\theta}_j^{T_i} - \theta_j \right)^2 \right] + \frac{1}{V} \mathbb{E}[\mathbb{I}_{E_x} \left\| L_x(\widehat{s}^{\mathrm{ho}}_{T_i} - s) \right\|^2] \\ + \frac{V-1}{V} \mathbb{E}\left[\mathbb{I}_{E_x} \langle L_x(\widehat{s}^{\mathrm{ho}}_{T_1} - s), L_x(\widehat{s}^{\mathrm{ho}}_{T_2} - s) \rangle \right]. \end{aligned}$$

Furthermore, for all $i \in [|1; V|]$, on E_x,

$$\left\| \widehat{s}^{\mathrm{ho}}_{T_i} - s \right\|^2 = \sum_{j=1}^{k_* + a_x\Delta} \left(\hat{\theta}_j^{T_i} - \theta_j \right)^2 + \left\| L_x(\widehat{s}^{\mathrm{ho}}_{T_i}) - L_x(s) \right\|^2 + \sum_{j \geqslant k_* + b_x\Delta + 1} \theta_j^2.$$

It follows that

$$\begin{aligned} \mathbb{E}\left[\mathbb{I}_{E_x} \left(\left\| \widehat{s}^{vf}_{n_t,V} - s \right\|^2 - \left\| \widehat{s}^{\mathrm{ho}}_{T_1} - s \right\|^2 \right) \right] = \mathbb{E}\left[\mathbb{I}_{E_x} \sum_{j=1}^{k_* + a_x\Delta} \left(\frac{1}{V} \sum_{i=1}^{V} \hat{\theta}_j^{T_i} - \theta_j \right)^2 - \left(\hat{\theta}_j^{T_1} - \theta_j \right)^2 \right] \\ + \frac{V-1}{V} \mathbb{E}\left[\mathbb{I}_{E_x} \langle L_x(\widehat{s}^{\mathrm{ho}}_{T_1} - s), L_x(\widehat{s}^{\mathrm{ho}}_{T_2} - s) \rangle \right] \\ - \frac{V-1}{V} \mathbb{E}\left[\mathbb{I}_{E_x} \left\| L_x(\widehat{s}^{\mathrm{ho}}_{T_1} - s) \right\|^2 \right]. \quad (6.83) \end{aligned}$$

By exchangeability of the collection $(\widehat{s}^{\mathrm{ho}}_{T_i} \mathbb{I}_{E_x^c})_{1 \leqslant i \leqslant V}$,

$$\mathbb{E}\left[\mathbb{I}_{E_x^c} \left(\left\| \widehat{s}^{vf}_{n_t,V} - s \right\|^2 - \left\| \widehat{s}^{\mathrm{ho}}_{T_1} - s \right\|^2 \right) \right] \leqslant 0, \text{ therefore}$$

$$\mathbb{E}\left[\left(\left\| \widehat{s}^{vf}_{n_t,V} - s \right\|^2 - \left\| \widehat{s}^{\mathrm{ho}}_{T_1} - s \right\|^2 \right) \right] \leqslant \mathbb{E}\left[\mathbb{I}_{E_x} \left(\left\| \widehat{s}^{vf}_{n_t,V} - s \right\|^2 - \left\| \widehat{s}^{\mathrm{ho}}_{T_1} - s \right\|^2 \right) \right]. \quad (6.84)$$

For the first term on the right of equation (6.83), the following upper bound can be proven.

Claim 6.7.5.2 *For all $n \geqslant n_1$ and all $x > 0$,*

$$\mathbb{E}\left[\mathbb{I}_{E_x}\sum_{j=1}^{k_*+a_x\Delta}\left(\frac{1}{V}\sum_{i=1}^{V}\hat{\theta}_j^{T_i}-\theta_j\right)^2-\left(\hat{\theta}_j^{T_1}-\theta_j\right)^2\right]\leqslant -\frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*}{n_t} + \kappa(1+x)n^{-\delta_3}\mathcal{E}+\frac{3}{2}\|s\|^2\mathbb{P}(E_x^c). \tag{6.85}$$

Proof For all $k \in \mathbb{N}$,

$$\begin{aligned}\frac{1}{V}\sum_{j=1}^{V}\hat{\theta}_k^{T_j} &= \frac{1}{V}\sum_{j=1}^{V}\frac{1}{n_t}\sum_{i=1}^{n}\mathbb{I}_{T_j}(i)\psi_k(X_i)\\ &= \frac{1}{n_t}\sum_{i=1}^{n}\psi_k(X_i)\times\frac{1}{V}\sum_{j=1}^{V}\mathbb{I}_{T_j}(i).\end{aligned} \tag{6.86}$$

By definition 6.3.2, the subsets $(T_j^c)_{j=1,\dots,V}$ are disjoint and have cardinality $n-n_t$. Let $B_V = \cup_{j=1}^{V}T_j^c$. If $i \in B_V$, $\sum_{j=1}^{V}\mathbb{I}_{T_j}(i) = V-1$ (there exists a unique index j such that $i \in T_j^c$). If $i \notin B_V$, then i is an element of all subsets T_j, therefore $\sum_{j=1}^{V}\mathbb{I}_{T_j}(i) = V$. Hence by equation (6.86), for all $k \in \mathbb{N}$,

$$\frac{1}{V}\sum_{i=1}^{V}\hat{\theta}_k^{T_i} = \frac{V-1}{Vn_t}\sum_{i\in B_V}\psi_k(X_i)+\frac{1}{n_t}\sum_{i\notin B_V}\psi_k(X_i).$$

Hence for all $k \in \mathbb{N}$, since the random variables X_i are i.i.d,

$$\mathbb{E}\left[\left(\frac{1}{V}\sum_{i=1}^{V}\hat{\theta}_k^{T_i}-\theta_j\right)^2\right] = \left(\frac{V-1}{Vn_t}\right)^2|B_V|\mathrm{Var}(\psi_k)+\frac{1}{n_t^2}(n-|B_V|)\mathrm{Var}(\psi_k).$$

By definition of B_V and since the subsets T_j^c are disjoint and have cardinality

$n - n_t$, $|B_V| = |\cup_{j=1}^{V} T_j^c| = V(n - n_t)$. Therefore,

$$\begin{aligned}
\frac{1}{\mathrm{Var}(\psi_k)}\mathbb{E}\left[\left(\frac{1}{V}\sum_{i=1}^{V}\hat{\theta}_k^{T_i} - \theta_j\right)^2\right] &= \left(\frac{V-1}{V}\right)^2 \frac{(n-n_t)V}{n_t^2} + \frac{n-(n-n_t)V}{n_t^2} \\
&= \left(\frac{V-1}{V}\right)^2 \frac{(n-n_t)V}{n_t^2} + \left[\frac{n}{n_t^2} - \frac{(n-n_t)V}{n_t^2}\right] \\
&= \frac{n}{n_t^2}\left[1 + \frac{n-n_t}{n}\left[\frac{(V-1)^2}{V} - V\right]\right] \\
&= \frac{n}{n_t^2}\left[1 + \frac{n-n_t}{n}\frac{-2V+1}{V}\right] \\
&= \frac{n}{n_t^2}\left[\frac{n_t}{n} - \frac{n-n_t}{n} + \frac{n-n_t}{nV}\right] \\
&= \frac{1}{n_t} - \frac{V-1}{V}\frac{n-n_t}{n_t^2}.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=1}^{k_*+a_x\Delta}\left(\frac{1}{V}\sum_{i=1}^{V}\hat{\theta}_j^{T_i} - \theta_j\right)^2 - \left(\hat{\theta}_j^{T_1} - \theta_j\right)^2\right] \\
&= -\frac{V-1}{V}\frac{n-n_t}{n_t}\sum_{j=1}^{k_*+a_x\Delta}\frac{\mathrm{Var}(\psi_j)}{n_t} \\
&\leqslant -\frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*+a_x\Delta}{n_t} + \frac{\|\theta\|_{\ell^1}}{n_t} \text{ by lemma 6.9.1} \\
&\leqslant -\frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*(n_t)}{n_t} + 2(1+x)\frac{n-n_t}{n_t}\mathcal{E} + \frac{n-n_t}{n_t}\frac{\|\theta\|_{\ell^1}}{n-n_t} \text{ by lemma 6.6.4.}
\end{aligned}$$

Setting $\kappa = 2 + \|\theta\|_{\ell^1}$, it follows from hypothesis H5 of section 6.2.2) and lemma 6.6.2 that

$$\mathbb{E}\left[\sum_{j=1}^{k_*+a_x\Delta}\left(\frac{1}{V}\sum_{i=1}^{V}\hat{\theta}_j^{T_i} - \theta_j\right)^2 - \left(\hat{\theta}_j^{T_1} - \theta_j\right)^2\right] \leqslant -\frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*}{n_t} + \kappa(1+x)n^{-\delta_3}\mathcal{E}. \tag{6.87}$$

Furthermore,

$$\mathbb{E}\left[\mathbb{I}_{E_x}\sum_{j=1}^{k_*+a_x\Delta}\left(\frac{1}{V}\sum_{i=1}^{V}\hat{\theta}_j^{T_i}-\theta_j\right)^2-\left(\hat{\theta}_j^{T_1}-\theta_j\right)^2\right]\leqslant\sum_{j=1}^{k_*+a_x\Delta}\mathbb{E}\left[\left(\frac{1}{V}\sum_{i=1}^{V}\hat{\theta}_j^{T_i}-\theta_j\right)^2\right]-\sum_{j=1}^{k_*+a_x\Delta}\mathbb{E}\left[\left(\hat{\theta}_j^{T_1}-\theta_j\right)^2\right]+\mathbb{E}\left[\mathbb{I}_{E_x^c}\sum_{j=1}^{k_*+a_x\Delta}\left(\hat{\theta}_j^{T_1}-\theta_j\right)^2\right],\tag{6.88}$$

thus by claim 6.9.4.1 and equation (6.87), for all $n\geqslant n_1$,

$$\begin{aligned}\mathbb{E}\left[\mathbb{I}_{E_x}\sum_{j=1}^{k_*+a_x\Delta}\left(\frac{1}{V}\sum_{i=1}^{V}\hat{\theta}_j^{T_i}-\theta_j\right)^2-\left(\hat{\theta}_j^{T_1}-\theta_j\right)^2\right]&\leqslant-\frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*}{n_t}+\kappa(1+x)n^{-\delta_3}\mathcal{E}\\&\quad+\frac{3}{2}\|s\|^2\mathbb{P}(E_x^c)\frac{k_*+a_x\Delta}{n_t}\\&\leqslant-\frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*}{n_t}+\kappa(1+x)n^{-\delta_3}\mathcal{E}\\&\quad+\frac{3}{2}\mathbb{P}(E_x^c)\frac{k_*}{n_t}.\end{aligned}$$

By comparison with $k=0,\hat{s}_0^T=0$, $\frac{k_*(n_t)}{n_t}\leqslant\mathrm{or}(n_t)\leqslant\|s\|^2$, which proves claim 6.7.5.2. ■

By definition 6.82, on the event E_x, for all $i\in[|1;V|]$, $\hat{k}_i^{ho}\in[|k_*+a_x\Delta;k_*+b_x\Delta|]$, therefore $\hat{k}_i^{ho}=\hat{k}_{i,x}^{ho}$ and

$$\mathbb{I}_{E_x}\langle L_x(\widehat{s}_{T_1}^{\mathrm{ho}}-s),L_x(\widehat{s}_{T_2}^{\mathrm{ho}}-s)\rangle=\mathbb{I}_{E_x}\langle L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}}-s),L_x(\hat{s}^{T_2}_{\hat{k}^{ho}_{2,x}}-s)\rangle$$

$$\mathbb{I}_{E_x}\left\|L_x(\widehat{s}_{T_1}^{\mathrm{ho}}-s)\right\|^2=\mathbb{I}_{E_x}\left\|L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}}-s)\right\|^2.$$

Furthermore, by the Cauchy-Schwarz inequality and the fact that $\mathbb{I}_{E_x^c}\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}}$ and

$\mathbb{I}_{E_x^c}\hat{s}^{T_2}_{\hat{k}^{ho}_{2,x}}$ have the same probability distribution,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{I}_{E_x^c}\left(\langle L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}}-s), L_x(\hat{s}^{T_2}_{\hat{k}^{ho}_{2,x}}-s)\rangle - \left\|L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}}-s)\right\|^2\right)\right]\\
&\geqslant -2\mathbb{E}\left[\mathbb{I}_{E_x^c}\left\|L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}}-s)\right\|^2\right]\\
&\geqslant -2\mathbb{P}(E_x^c)\sum_{j=a_x\Delta+1}^{b_x\Delta}\theta^2_{k_*+j} - 2\mathbb{E}\left[\mathbb{I}_{E_x^c}\sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta}\left(\hat{\theta}^T_j-\theta_j\right)^2\right].
\end{aligned}$$

By lemma 6.6.4 and claim 6.9.4.1, for all $n \geqslant n_1$,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{I}_{E_x^c}\left(\langle L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}}-s), L_x(\hat{s}^{T_2}_{\hat{k}^{ho}_{2,x}}-s)\rangle - \left\|L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}}-s)\right\|^2\right)\right]\\
&\geqslant -8(1+x)\mathbb{P}(E_x^c)\mathcal{E} - 3(b_x-a_x)\mathbb{P}(E_x^c)\mathcal{E}\\
&\geqslant -14(1+x)\mathbb{P}(E_x^c)\mathcal{E}.
\end{aligned}$$

Hence:

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{I}_{E_x}\left(\langle L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}}-s), L_x(\hat{s}^{T_2}_{\hat{k}^{ho}_{2,x}}-s)\rangle - \left\|L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}}-s)\right\|^2\right)\right]\\
&\leqslant \mathbb{E}\left[\langle L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}}-s), L_x(\hat{s}^{T_2}_{\hat{k}^{ho}_{2,x}}-s)\rangle - \left\|L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}}-s)\right\|^2\right] + 14(1+x)\mathbb{P}(E_x^c)\mathcal{E}.
\end{aligned} \tag{6.89}$$

To sum up, with the upper bounds obtained in equations (6.84), (6.89) and claim 6.7.5.2, equation (6.83) yields, for all $n \geqslant n_1$,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{s}^{vf}_{n_t,V}-s\right\|^2\right] - \mathbb{E}\left[\left\|\widehat{s}^{\mathrm{ho}}_{T_1}-s\right\|^2\right]\\
&\leqslant \mathbb{E}\left[\langle L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}}-s), L_x(\hat{s}^{T_2}_{\hat{k}^{ho}_{2,x}}-s)\rangle - \left\|L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}}-s)\right\|^2\right]\\
&\quad - \frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*(n_t)}{n_t} + \left[\tfrac{3}{2}\|s\|^2 + 14(1+x)\mathcal{E}\right]\mathbb{P}(E_x^c) + \kappa n^{-\delta_3}(1+x)\mathcal{E}.
\end{aligned} \tag{6.90}$$

We would now like to prove that $\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}}$ and $\hat{s}^{T_2}_{\hat{k}^{ho}_{2,x}}$ can be replaced in equation (6.90) by $\hat{s}^{T}_{\hat{k}^{ho}_{1,x}}, \hat{s}^{T}_{\hat{k}^{ho}_{2,x}}$, up to negligible error, where the common subset T is such

that D_n^T is independent of $D_n^{T_1^c}$ and $D_n^{T_2^c}$. Let $D'_{n-n_t} \sim P^{\otimes(n-n_t)}$ be an i.i.d sample of size $n-n_t$ independent from D_n, and let $D_{2n-n_t} = (D_n, D'_{n-n_t})$. For $T \not\subset \{1 \dots n\}$, $P^T_{2n-n_t}$ and $\hat{s}^T_k$ are defined using the extended dataset D_{2n-n_t}. Let $T' = [|n+1; 2n-n_t|]$ and $T = (T_1 \cap T_2) \cup T' = (T_1 \cap T_2) \cup [|n+1; 2n-n_t|]$. By the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\mathbb{E}\left[\langle L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}}) - L_x(s), L_x(\hat{s}^{T_2}_{\hat{k}^{ho}_{2,x}}) - L_x(s)\rangle\right] \\
&\quad= \mathbb{E}\Big[\langle L_x(\hat{s}^{T}_{\hat{k}^{ho}_{1,x}}) - L_x(s), L_x(\hat{s}^{T}_{\hat{k}^{ho}_{2,x}}) - L_x(s)\rangle + \langle L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}} - \hat{s}^{T}_{\hat{k}^{ho}_{1,x}}), L_x(\hat{s}^{T}_{\hat{k}^{ho}_{2,x}}) - L_x(s)\rangle \\
&\qquad+ \langle L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}}) - L_x(s), L_x(\hat{s}^{T_2}_{\hat{k}^{ho}_{2,x}} - \hat{s}^{T}_{\hat{k}^{ho}_{2,x}})\rangle\Big] \\
&\quad= \mathbb{E}\left[\langle L_x(\hat{s}^{T}_{\hat{k}^{ho}_{1,x}}) - L_x(s), L_x(\hat{s}^{T}_{\hat{k}^{ho}_{2,x}}) - L_x(s)\rangle\right] \\
&\qquad\pm \mathbb{E}\left[\left\|L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}} - \hat{s}^{T}_{\hat{k}^{ho}_{1,x}})\right\|^2\right]^{\frac{1}{2}} \mathbb{E}\left[\left\|L_x(\hat{s}^{T}_{\hat{k}^{ho}_{2,x}} - s)\right\|^2\right]^{\frac{1}{2}} \\
&\qquad\pm \mathbb{E}\left[\left\|L_x(\hat{s}^{T_2}_{\hat{k}^{ho}_{2,x}} - \hat{s}^{T}_{\hat{k}^{ho}_{2,x}})\right\|^2\right]^{\frac{1}{2}} \mathbb{E}\left[\left\|L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}} - s)\right\|^2\right]^{\frac{1}{2}}. \qquad (6.91)
\end{aligned}$$

Since T_1^c, T_2^c, T' are pairwise disjoint, $T = T_1 \cap T_2 \cup T'$ and $|T_1^c| = |T_2^c| = |T'| = n - n_t$,

$$\begin{aligned}
\left\|L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}} - \hat{s}^{T}_{\hat{k}^{ho}_{1,x}})\right\|^2 &\leqslant \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \left(P_n^T - P_n^{T_1}\right)^2(\psi_j) \\
&= \left(\frac{n-n_t}{n_t}\right)^2 \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \left(P^{T'}_{2n-n_t} - P^{T_2^c}_{2n-n_t}\right)^2(\psi_j).
\end{aligned}$$

Hence:

$$\begin{aligned}
\mathbb{E}\left[\left\|L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}} - \hat{s}^{T}_{\hat{k}^{ho}_{1,x}})\right\|^2\right] &= 2\left(\frac{n-n_t}{n_t}\right)^2 \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \frac{\mathrm{Var}(\psi_j(X_1))}{n-n_t} \\
&\leqslant 2\left(\frac{n-n_t}{n_t}\right)^2 \sqrt{2}\frac{[b_x-a_x]\Delta}{n-n_t} \text{ since } \|\psi_j\|_\infty \leqslant \sqrt{2} \\
&\leqslant 2\sqrt{2}\frac{[b_x-a_x](n-n_t)}{n_t}\mathcal{E} \\
&\leqslant 4\sqrt{2}(1+x)\frac{n-n_t}{n_t}\mathcal{E} \text{ by lemma 6.6.4} \\
&\leqslant 4\sqrt{2}(1+x)n^{-\delta_3}\mathcal{E} \text{ by hypothesis H5 of section 6.2.2.}
\end{aligned} \qquad (6.92)$$

The terms $\left\|L_x(\hat{s}^T_{\hat{k}^{ho}_{i,x}} - s)\right\|^2$ in equation (6.91) can be bounded as follows.

$$\begin{aligned}\left\|L_x(\hat{s}^T_{\hat{k}^{ho}_{2,x}} - s)\right\|^2 &= \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \left(\hat{\theta}^T_j - \theta_j\right)^2 \mathbb{I}_{j\leqslant \hat{k}^{ho}_{2,x}} + \theta_j^2 \mathbb{I}_{j\geqslant \hat{k}^{ho}_{2,x}+1} \\ &\leqslant \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \left(\hat{\theta}^T_j - \theta_j\right)^2 + \theta_j^2.\end{aligned}$$

Hence, by lemma 6.6.4,

$$\begin{aligned}\mathbb{E}\left[\left\|L_x(\hat{s}^T_{\hat{k}^{ho}_{2,x}} - s)\right\|^2\right] &\leqslant \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \frac{\mathrm{Var}(\psi_j(X_1))}{n_t} + \theta_j^2 \\ &\leqslant \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \frac{\sqrt{2}}{n_t} + \theta_j^2 \\ &\leqslant \sqrt{2}(b_x - a_x)\mathcal{E} + 4(1+x)\mathcal{E} \\ &\leqslant (4+2\sqrt{2})(1+x)\mathcal{E}. \end{aligned} \tag{6.93}$$

In conclusion, by equations (6.91), (6.93) and (6.92):

$$\left|\mathbb{E}\left[\langle L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}} - s), L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{2,x}} - s)\rangle - \langle L_x(\hat{s}^T_{\hat{k}^{ho}_{1,x}} - s), L_x(\hat{s}^T_{\hat{k}^{ho}_{2,x}} - s)\rangle\right]\right| \leqslant 16(1+x)n^{-\frac{\delta_3}{2}}\mathcal{E}. \tag{6.94}$$

Let $(W_i)_{[|1;V|]}$ be independent Wiener processes, depending on D_{2n-n_t} only through $D_n^{T_i^c}$ and satisfying equation (5.11) of Theorem 5.3.8. For all $i \in [|1;V|]$ and all $x > 0$, let

$$\hat{\alpha}_i = \underset{\alpha\in\left[\frac{-k_*}{\Delta};+\infty\right[}{\mathrm{argmin}} f_n(\alpha) - W_i(g_n(\alpha)) \tag{6.95}$$

$$\hat{\alpha}_i^{\infty,x} = \underset{\alpha\in[a_x;b_x]}{\mathrm{argmin}} f_n(\alpha) - W_i(g_n(\alpha)) \tag{6.96}$$

$$\hat{k}_i^{\infty} = \lceil k_* + \hat{\alpha}_i\Delta\rceil \tag{6.97}$$

$$\hat{k}_{i,x}^{\infty} = \lceil k_* + \hat{\alpha}_i^{\infty,x}\Delta\rceil. \tag{6.98}$$

It follows that $\hat{\alpha}_1, \hat{\alpha}_1^{\infty,x}, \hat{k}_{1,x}^{\infty}, \hat{k}_1^{\infty}$ are independent from $\hat{\alpha}_2, \hat{\alpha}_2^{\infty,x}, \hat{k}_{2,x}^{\infty}, \hat{k}_2^{\infty}$. Furthermore,

$$\begin{aligned}\langle L_x(\hat{s}^T_{\hat{k}^{ho}_{1,x}} - s), L_x(\hat{s}^T_{\hat{k}^{ho}_{2,x}} - s)\rangle = &\sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \left(\hat{\theta}^T_j - \theta_j\right)^2 \mathbb{I}_{j\leqslant \hat{k}^{ho}_{1,x}\wedge\hat{k}^{ho}_{2,x}} + \theta_j^2 \mathbb{I}_{j\geqslant \hat{k}^{ho}_{1,x}\vee\hat{k}^{ho}_{2,x}+1} \\ &- \theta_j\left(\hat{\theta}^T_j - \theta_j\right)\mathbb{I}_{\hat{k}^{ho}_{1,x}\wedge\hat{k}^{ho}_{2,x}+1\leqslant j\leqslant \hat{k}^{ho}_{1,x}\vee\hat{k}^{ho}_{2,x}},\end{aligned} \tag{6.99}$$

while by the same argument,

$$\langle L_x(\hat{s}^T_{\hat{k}^\infty_{1,x}} - s), L_x(\hat{s}^T_{\hat{k}^\infty_{2,x}} - s)\rangle = \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \left(\hat{\theta}^T_j - \theta_j\right)^2 \mathbb{I}_{j\leqslant \hat{k}^\infty_{1,x}\wedge\hat{k}^\infty_{2,x}} + \theta_j^2 \mathbb{I}_{j\geqslant \hat{k}^\infty_{1,x}\vee\hat{k}^\infty_{2,x}+1} - \theta_j\left(\hat{\theta}^T_j - \theta_j\right)\mathbb{I}_{\hat{k}^\infty_{1,x}\wedge\hat{k}^\infty_{2,x}+1\leqslant j\leqslant \hat{k}^\infty_{1,x}\vee\hat{k}^\infty_{2,x}}. \tag{6.100}$$

For any real random variable $\hat{t}$, introduce the notation $\hat{F}_{\hat{t}}(t) = \mathbb{P}(\hat{t}\leqslant t|D_n^{T_1\cap T_2})$ and $F_{\hat{t}}(t) = \mathbb{P}(\hat{t}\leqslant t)$.

Recall that $T = (T_1\cap T_2)\cup T'$ where $T'\cap\{1\dots n\} = \emptyset$, therefore the pair $(\hat{k}^{ho}_{1,x}, \hat{k}^{ho}_{2,x})$ is conditionally independent of $D^T_{2n-n_t}$ given $D_n^{T_1\cap T_2}$. Taking the conditional expectation of equation (6.99) given $D^T_{2n-n_t}$, it follows that:

$$\mathbb{E}\left[\langle L_x(\hat{s}^T_{\hat{k}^{ho}_{1,x}} - s), L_x(\hat{s}^T_{\hat{k}^{ho}_{2,x}} - s)\rangle|D^T_{2n-n_t}\right] = \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \left[1 - \hat{F}_{\hat{k}^{ho}_{1,x}\wedge\hat{k}^{ho}_{2,x}}(j-1)\right]\left(\hat{\theta}^T_j - \theta_j\right)^2 - \left[\hat{F}_{\hat{k}^{ho}_{1,x}\wedge\hat{k}^{ho}_{2,x}}(j-1) - \hat{F}_{\hat{k}^{ho}_{1,x}\vee\hat{k}^{ho}_{2,x}}(j-1)\right]\theta_j\left(\hat{\theta}^T_j - \theta_j\right) + \hat{F}_{\hat{k}^{ho}_{1,x}\vee\hat{k}^{ho}_{2,x}}(j-1)\theta_j^2. \tag{6.101}$$

By construction (equation (6.98)), the pair $(\hat{k}^\infty_{1,x}, \hat{k}^\infty_{2,x})$ is independent from $(D^T_{2n-n_t}, D^{T'}_{2n-n_t})$ therefore from $D^T_{2n-n_t}$. Thus, taking the conditional expectation of equation (6.100) given $D^T_{2n-n_t}$ yields

$$\mathbb{E}\left[\langle L_x(\hat{s}^T_{\hat{k}^\infty_{1,x}} - s), L_x(\hat{s}^T_{\hat{k}^\infty_{2,x}} - s)\rangle|D^T_{2n-n_t}\right] = \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \left[1 - F_{\hat{k}^\infty_{1,x}\wedge\hat{k}^\infty_{2,x}}(j-1)\right]\left(\hat{\theta}^T_j - \theta_j\right)^2 - \left[F_{\hat{k}^\infty_{1,x}\wedge\hat{k}^\infty_{2,x}}(j-1) - F_{\hat{k}^\infty_{1,x}\vee\hat{k}^\infty_{2,x}}(j-1)\right]\theta_j\left(\hat{\theta}^T_j - \theta_j\right) + F_{\hat{k}^\infty_{1,x}\vee\hat{k}^\infty_{2,x}}(j-1)\theta_j^2. \tag{6.102}$$

Therefore, by equations (6.101) and (6.102),

$$\left|\mathbb{E}\left[\langle L_x(\hat{s}^T_{\hat{k}^{ho}_{1,x}} - s), L_x(\hat{s}^T_{\hat{k}^{ho}_{2,x}} - s)\rangle - \langle L_x(\hat{s}^T_{\hat{k}^\infty_{1,x}} - s), L_x(\hat{s}^T_{\hat{k}^\infty_{2,x}} - s)\rangle|D^T_{2n-n_t}\right]\right| \leqslant 2\max\left\{\left\|\hat{F}_{\hat{k}^{ho}_{1,x}\wedge\hat{k}^{ho}_{2,x}} - F_{\hat{k}^\infty_{1,x}\wedge\hat{k}^\infty_{2,x}}\right\|_\infty, \left\|\hat{F}_{\hat{k}^{ho}_{1,x}\vee\hat{k}^{ho}_{2,x}} - F_{\hat{k}^\infty_{1,x}\vee\hat{k}^\infty_{2,x}}\right\|_\infty\right\} \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta}\left(\hat{\theta}^T_j - \theta_j\right)^2 + \theta_j^2. \tag{6.103}$$

Since by definition, $\hat{k}^{\infty}_{i,x} = \lceil k_* + \hat{\alpha}^{\infty,x}_{W_i} \rceil$, for all $j \in \mathbb{N}$, $\hat{k}^{\infty}_{i,x} \leqslant j \iff k_* + \hat{\alpha}^{\infty,x}_{W_i} \leqslant j$, therefore by claim 6.7.2.2,

$$\begin{aligned}
&\sqrt{\kappa_5}(1+x)^{\frac{9}{64}} n^{-\frac{u_1}{32}} \mathbb{P}\Big(\max\Big\{ \Big\| \hat{F}_{\hat{k}^{ho}_{1,x} \wedge \hat{k}^{ho}_{2,x}} - F_{\hat{k}^{\infty}_{1,x} \wedge \hat{k}^{\infty}_{2,x}} \Big\|_{\infty}, \\
&\qquad\qquad \Big\| \hat{F}_{\hat{k}^{ho}_{1,x} \vee \hat{k}^{ho}_{2,x}} - F_{\hat{k}^{\infty}_{1,x} \vee \hat{k}^{\infty}_{2,x}} \Big\|_{\infty} \Big\} \geqslant \sqrt{\kappa_5}(1+x)^{\frac{9}{64}} n^{-\frac{u_1}{32}} \Big) \\
&\leqslant \mathbb{E}\left[\max\left\{ \Big\| \hat{F}_{\hat{k}^{ho}_{1,x} \wedge \hat{k}^{ho}_{2,x}} - F_{\hat{k}^{\infty}_{1,x} \wedge \hat{k}^{\infty}_{2,x}} \Big\|_{\infty}, \Big\| \hat{F}_{\hat{k}^{ho}_{1,x} \vee \hat{k}^{ho}_{2,x}} - F_{\hat{k}^{\infty}_{1,x} \vee \hat{k}^{\infty}_{2,x}} \Big\|_{\infty} \right\} \right] \\
&\leqslant 2\kappa_5 (1+x)^{\frac{9}{32}} n^{-\frac{u_1}{16}}.
\end{aligned} \tag{6.104}$$

Hence by equation (6.103) and claim 6.9.4.1, for all $n \geqslant n_1$,

$$\begin{aligned}
&\left| \mathbb{E}\left[\langle L_x(\hat{s}^T_{\hat{k}^{ho}_{1,x}} - s), L_x(\hat{s}^T_{\hat{k}^{ho}_{2,x}} - s) \rangle - \langle L_x(\hat{s}^T_{\hat{k}^{\infty}_{1,x}} - s), L_x(\hat{s}^T_{\hat{k}^{\infty}_{2,x}} - s) \rangle \right] \right| \\
&\leqslant (b_x - a_x)[5\sqrt{\kappa_5}(1+x)^{\frac{9}{64}} n^{-\frac{u_1}{32}}]\mathcal{E} + 2\kappa_5 (1+x)^{\frac{9}{32}} n^{-\frac{u_1}{16}} \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \theta_j^2 \\
&\leqslant (10\sqrt{\kappa_5} + 8\kappa_5)(1+x)^{\frac{41}{32}} n^{-\frac{u_1}{32}} \mathcal{E} \text{ by lemma 6.6.4.}
\end{aligned} \tag{6.105}$$

By construction (equation (6.98)), $\hat{k}^{\infty}_{1,x}$ and $\hat{k}^{\infty}_{2,x}$ are i.i.d, in particular

$$\forall y \in \mathbb{R}, F_{\hat{k}^{\infty}_{1,x} \vee \hat{k}^{\infty}_{2,x}}(y) = F_{\hat{k}^{\infty}_{1,x}}(y)^2 \text{ and } \left[1 - F_{\hat{k}^{\infty}_{1,x} \wedge \hat{k}^{\infty}_{2,x}}\right](y) = \left[1 - F_{\hat{k}^{\infty}_{1,x}}\right]^2 (y).$$

It follows by equation (6.102) and lemma 6.9.1 that

$$\begin{aligned}
&\mathbb{E}\left[\langle L_x(\hat{s}^T_{\hat{k}^{\infty}_{1,x}} - s), L_x(\hat{s}^T_{\hat{k}^{\infty}_{2,x}} - s) \rangle \right] \\
&= \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \left[1 - F_{\hat{k}^{\infty}_{1,x}}(j-1)\right]^2 \frac{\mathrm{Var}(\psi_j(X_1))}{n_t} + F_{\hat{k}^{\infty}_{1,x}}(j-1)^2 \theta_j^2 \\
&= \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} \left[1 - F_{\hat{k}^{\infty}_{1,x}}(j-1)\right]^2 \frac{1}{n_t} + F_{\hat{k}^{\infty}_{1,x}}(j-1)^2 \theta_j^2 \pm \frac{\|\theta\|_{\ell^1}}{n_t}
\end{aligned}$$

By equation (6.105) and claim 6.7.5.1, there exists therefore a constant κ such that, for all $n \geqslant n_1$ and all $x > 0$,

$$\begin{aligned}
&\mathbb{E}\left[\langle L_x(\hat{s}^T_{\hat{k}^{ho}_{1,x}} - s), L_x(\hat{s}^T_{\hat{k}^{ho}_{2,x}} - s) \rangle - \left\| L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}} - s) \right\|^2 \right] \\
&\leqslant - \sum_{j=k_*+a_x\Delta+1}^{k_*+b_x\Delta} [F_{\hat{k}^{\infty}_{1,x}}(1 - F_{\hat{k}^{\infty}_{1,x}})](j-1) \left[\frac{1}{n_t} + \theta_j^2 \right] + \kappa (1+x)^{\frac{41}{32}} n^{-\frac{u_1}{32}} \mathcal{E} \\
&\quad + \kappa_6 (1+x)^{\frac{41}{32}} n^{-u_6} \mathfrak{e} + \frac{\|\theta\|_{\ell^1}}{n_t}.
\end{aligned} \tag{6.106}$$

Let $\kappa = \kappa + \kappa_6 + \|\theta\|_{\ell^1}$ and $u = \min(\frac{u_1}{32}, u_6, \delta_3)$. For any $j \leqslant 0$, $\theta^2_{k_*+j} \geqslant \frac{1}{n_t}$ while for any $j \geqslant 0$, $\theta^2_j \geqslant 0$, thus for all $n \geqslant n_1$,

$$\begin{aligned}
&\mathbb{E}\left[\langle L_x(\hat{s}^T_{\hat{k}^{ho}_{1,x}} - s), L_x(\hat{s}^T_{\hat{k}^{ho}_{2,x}} - s)\rangle - \left\|L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}} - s)\right\|^2\right] \\
&\quad \leqslant -\frac{2\Delta}{n_t}\frac{1}{\Delta}\sum_{j=a_x\Delta+1}^{0}\left[F_{\hat{k}^{\infty}_{1,x}}(1 - F_{\hat{k}^{\infty}_{1,x}})\right](k_* + j - 1) - \frac{\Delta}{n_t}\sum_{j=1}^{b_x\Delta}\left[F_{\hat{k}^{\infty}_{1,x}}(1 - F_{\hat{k}^{\infty}_{1,x}})\right](k_* + j - 1) \\
&\quad + \kappa(1+x)^{\frac{41}{32}} n^{-u}\mathcal{E}.
\end{aligned}$$

Remark also that for all $j \in \mathbb{Z}$,

$$\hat{k}^{\infty}_{1,x} \leqslant j \iff \lceil k_* + \hat{\alpha}^{\infty,x}_1 \Delta\rceil \leqslant j \iff k_* + \hat{\alpha}^{\infty,x}_1\Delta \leqslant j \iff \hat{\alpha}^{\infty,x}_1 \leqslant \frac{j - k_*}{\Delta}.$$

It follows that for all $n \geqslant n_1$,

$$\begin{aligned}
&\mathbb{E}\left[\langle L_x(\hat{s}^T_{\hat{k}^{ho}_{1,x}} - s), L_x(\hat{s}^T_{\hat{k}^{ho}_{2,x}} - s)\rangle - \left\|L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}} - s)\right\|^2\right] \\
&\quad \leqslant -2\mathcal{E}\frac{1}{\Delta}\sum_{j=a_x\Delta}^{-1}\left[F_{\hat{\alpha}^{\infty,x}_1}(1 - F_{\hat{\alpha}^{\infty,x}_1})\right]\left(\frac{j}{\Delta}\right) - \mathcal{E}\frac{1}{\Delta}\sum_{j=0}^{b_x\Delta-1}\left[F_{\hat{\alpha}^{\infty,x}_1}(1 - F_{\hat{\alpha}^{\infty,x}_1})\right]\left(\frac{j}{\Delta}\right) \\
&\qquad + \kappa(1+x)^{\frac{41}{32}} n^{-u}\mathcal{E}.
\end{aligned}$$

On the other hand by claim 6.7.4.1, for all $j \in [|a_x\Delta; b_x\Delta - 1|]$,

$$\begin{aligned}
\left|\left[F_{\hat{\alpha}^{\infty,x}_1}(1 - F_{\hat{\alpha}^{\infty,x}_1})\right]\left(\tfrac{j}{\Delta}\right) - \Delta\int_{\frac{j}{\Delta}}^{\frac{j+1}{\Delta}}\left[F_{\hat{\alpha}^{\infty,x}_1}(1 - F_{\hat{\alpha}^{\infty,x}_1})\right](t)dt\right| &\leqslant F_{\hat{\alpha}^{\infty,x}_1}\left(\tfrac{j+1}{\Delta}\right) - F_{\hat{\alpha}^{\infty,x}_1}\left(\tfrac{j}{\Delta}\right) \\
&\leqslant \kappa\sqrt{1+x}n^{-\frac{u_2\wedge\delta_3}{4}},
\end{aligned}$$

therefore

$$\begin{aligned}
\mathbb{E}\left[\langle L_x(\hat{s}^T_{\hat{k}^{ho}_{1,x}} - s), L_x(\hat{s}^T_{\hat{k}^{ho}_{2,x}} - s)\rangle - \left\|L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}} - s)\right\|^2\right] &\leqslant -2\mathcal{E}\int_{a_x}^{0}\left[F_{\hat{\alpha}^{\infty,x}_1}(1 - F_{\hat{\alpha}^{\infty,x}_1})\right](t)dt \\
&\quad - \mathcal{E}\int_{0}^{b_x}\left[F_{\hat{\alpha}^{\infty,x}_1}(1 - F_{\hat{\alpha}^{\infty,x}_1})\right](t)dt \\
&\quad + 2\kappa(b_x - a_x)\sqrt{1+x}n^{-\frac{u_2\wedge\delta_3}{4}}\mathcal{E} \\
&\quad + \kappa(1+x)^{\frac{41}{32}} n^{-u}\mathcal{E}.
\end{aligned}$$

Since $b_x - a_x \leqslant 2(1+x)$ by lemma 6.6.4, setting $u = \min(u, \frac{u_2}{4}, \frac{\delta_3}{4})$, it follows that for some constant $\kappa \geqslant 0$ and all $n \geqslant n_1$,

$$\begin{aligned}\mathbb{E}\left[\langle L_x(\hat{s}^T_{\hat{k}^{ho}_{1,x}} - s), L_x(\hat{s}^T_{\hat{k}^{ho}_{2,x}} - s)\rangle - \left\|L_x(\hat{s}^{T_1}_{\hat{k}^{ho}_{1,x}} - s)\right\|^2\right] &\leqslant -2\mathcal{E}\int_{a_x}^{0}[F_{\hat{\alpha}_1^{\infty,x}}(1-F_{\hat{\alpha}_1^{\infty,x}})](t)dt \\ &\quad - \mathcal{E}\int_{0}^{b_x}[F_{\hat{\alpha}_1^{\infty,x}}(1-F_{\hat{\alpha}_1^{\infty,x}})](t)dt \\ &\quad + \kappa(1+x)^{\frac{3}{2}}n^{-u}\mathcal{E}.\end{aligned}$$

By equation (6.90), this yields

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{s}^{vf}_{n_t,V} - s\right\|^2\right] &\leqslant -\frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*}{n_t} - \frac{V-1}{V}\mathcal{E}\Big[2\int_{a_x}^{0}[F_{\hat{\alpha}_1^{\infty,x}}(1-F_{\hat{\alpha}_1^{\infty,x}})](t)dt \\ &\quad + \int_{0}^{b_x}[F_{\hat{\alpha}_1^{\infty,x}}(1-F_{\hat{\alpha}_1^{\infty,x}})](t)dt\Big] \\ &\quad + \left[\tfrac{3}{2}\|s\|^2 + 14(1+x)\mathcal{E}\right]\mathbb{P}(E_x^c) + \kappa n^{-\delta_3}(1+x)\mathcal{E} + \kappa(1+x)^{\frac{3}{2}}n^{-u}\mathcal{E}.\end{aligned} \tag{6.107}$$

Furthermore, since $\hat{\alpha}_1^{\infty,x} \in [a_x; b_x]$,

$$\begin{aligned}&\int_{-\infty}^{+\infty}\left|F_{\hat{\alpha}_1^{\infty}}[1-F_{\hat{\alpha}_1^{\infty}}] - F_{\hat{\alpha}_1^{\infty,x}}[1-F_{\hat{\alpha}_1^{\infty,x}}]\right|(t)dt \\ &\quad\leqslant \int_{a_x}^{b_x}\left|F_{\hat{\alpha}_1^{\infty}}(t) - F_{\hat{\alpha}_1^{\infty,x}}(t)\right|dt + \int_{-\infty}^{a_x}F_{\hat{\alpha}_1^{\infty}}(t)dt + \int_{b_x}^{+\infty}[1-F_{\hat{\alpha}_1^{\infty}}](t)dt \\ &\quad\leqslant (b_x-a_x)\left\|F_{\hat{\alpha}_1^{\infty}} - F_{\hat{\alpha}_1^{\infty,x}}\right\|_\infty + b_x\mathbb{P}\left(\hat{\alpha}_1^{\infty} \geqslant b_x\right) + \int_{b_x}^{+\infty}[1-F_{\hat{\alpha}_1^{\infty}}](t)dt \\ &\qquad - a_x\mathbb{P}\left(\hat{\alpha}_1^{\infty} \leqslant a_x\right) + \int_{-\infty}^{a_x}F_{\hat{\alpha}_1^{\infty}}(t)dt \\ &\quad\leqslant (b_x-a_x)\mathbb{P}\left(\hat{\alpha}_1^{\infty} \neq \hat{\alpha}_1^{\infty,x}\right) + \int_0^{+\infty}\mathbb{P}(-\hat{\alpha}_1^{\infty}\mathbb{I}_{\hat{\alpha}_1^{\infty}\leqslant a_x} \geqslant t)dt + \int_0^{+\infty}\mathbb{P}(\hat{\alpha}_1^{\infty}\mathbb{I}_{\hat{\alpha}_1^{\infty}\geqslant b_x} \geqslant t)dt \\ &\quad\leqslant (b_x-a_x)\mathbb{P}(\hat{\alpha}_1^{\infty} \notin [a_x; b_x]) + \mathbb{E}\left[|\hat{\alpha}_1^{\infty}|\mathbb{I}_{\hat{\alpha}_1^{\infty}\notin[a_x;b_x]}\right].\end{aligned}$$

By lemma 6.6.3, for all x, $f_n(x) \geqslant |x| - 1$, hence

$$\begin{aligned}&\int_{-\infty}^{+\infty}\left|F_{\hat{\alpha}_1^{\infty}}[1-F_{\hat{\alpha}_1^{\infty}}] - F_{\hat{\alpha}_1^{\infty,x}}[1-F_{\hat{\alpha}_1^{\infty,x}}]\right|(t)dt \\ &\quad\leqslant (b_x-a_x)\mathbb{P}(\hat{\alpha}_1^{\infty} \notin [a_x; b_x]) + \mathbb{E}\left[(1+f_n(\hat{\alpha}_1^{\infty}))\mathbb{I}_{\hat{\alpha}_1^{\infty}\notin[a_x;b_x]}\right].\end{aligned}$$

By claim 6.9.3.1 and lemma 6.9.4, for all $n \geqslant n_2$ and all $x \geqslant 2\kappa_7(1+\log n)$, $\mathbb{P}(\hat{\alpha}_1^\infty \notin [a_x; b_x]) \leqslant \frac{1}{n}$ and

$$\begin{aligned}
&\int_{-\infty}^{+\infty} \left|F_{\hat{\alpha}_1^\infty}[1-F_{\hat{\alpha}_1^\infty}] - F_{\hat{\alpha}_1^{\infty,x}}[1-F_{\hat{\alpha}_1^{\infty,x}}]\right|(t)dt \\
&\quad\leqslant \frac{b_x - a_x}{n} + \frac{1}{n} + \int_{\log n}^{+\infty} \kappa_7(1+x)e^{-x}dx \\
&\quad\leqslant \frac{1+b_x-a_x}{n} + \frac{\kappa_7}{n} + \left([xe^{-x}]_{\log n}^{+\infty} + \int_{\log n}^{+\infty} e^{-x}dx\right) \\
&\quad\leqslant (1+x)\frac{3}{n} + \kappa_7\frac{2+\log n}{n}. \qquad (6.108)
\end{aligned}$$

Set $x = x_n = \max\big(2\kappa_7(1+\log n), \kappa_2 \log^2 n\big)$. By claim 6.9.2.2, there exists a constant n_4 such that for all $n \geqslant n_4$, $k_* + b_{x_n}\Delta \leqslant n - n_t$. It follows by Theorem 6.6.5 that

$$\begin{aligned}
\mathbb{P}(E_{x_n}^c) &\leqslant \sum_{i=1}^{V} \mathbb{P}\left(f_n\Big(\frac{\hat{k}_i^{ho} - k_*}{\Delta}\Big) \leqslant \kappa_2 \log^2 n\right) \\
&\leqslant \frac{2V}{n^2} \\
&\leqslant \frac{2}{n} \\
&\leqslant \frac{2}{n-n_t}\frac{n-n_t}{n_t} \\
&\leqslant 2n^{-\delta_3}\mathcal{E} \text{ by lemma 6.6.2 and section 6.2.2 , hypothesis } H5.
\end{aligned}$$

Thus, by equation (6.107) and equation (6.108), there exists two constants $\kappa, u > 0$ such that for all $n \geqslant \max(n_1, n_2)$,

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{s}_{n_t,V}^{vf} - s\right\|^2\right] \leqslant -\frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*}{n_t} - \frac{V-1}{V}\mathcal{E}\Big[2\int_{-\infty}^{0}[F_{\hat{\alpha}_1^\infty}(1-F_{\hat{\alpha}_1^\infty})](t)dt \\
+ \int_0^{+\infty}[F_{\hat{\alpha}_1^\infty}(1-F_{\hat{\alpha}_1^\infty})](t)dt\Big] + \kappa n^{-u}\mathcal{E}.
\end{aligned}$$

This proves equation (6.8) of Theorem 6.4.4.

It remains to examine the case of $\widehat{s}_{n_t,V}^{mc}$. Let $(T_j)_{1\leqslant j\leqslant V}$ be the subsets which appear in Definition 6.3.3. The $\widehat{s}_{T_j}^{\mathrm{ho}}$ estimators are then exchangeable, so by lemma 6.9.5,

$$\mathbb{E}\left[\left\|\widehat{s}_{n_t,V}^{mc} - s\right\|^2\right] = \frac{1}{V}\mathbb{E}\left[\left\|\widehat{s}_{T_1}^{\mathrm{ho}} - s\right\|^2\right] + \frac{V-1}{V}\mathbb{E}\left[\langle \widehat{s}_{T_1}^{\mathrm{ho}} - s, \widehat{s}_{T_2}^{\mathrm{ho}} - s\rangle\right].$$

Since T_1, T_2 are i.i.d and independent of D_n,

$$\mathbb{E}\left[\left\|\widehat{s}^{mc}_{n_t,V} - s\right\|^2\right] = \frac{1}{V}\mathbb{E}\left[\left\|\widehat{s}^{\text{ho}}_{T_1} - s\right\|^2\right] + \frac{V-1}{V}\mathbb{E}\left[\left\|\mathbb{E}\left[\widehat{s}^{\text{ho}}_{T_1}|D_n\right] - s\right\|^2\right]. \quad (6.109)$$

Moreover,

$$\mathbb{E}\left[\widehat{s}^{\text{ho}}_{T_1}|D_n\right] = \frac{1}{\binom{n}{n_t}} \sum_{T\subset\{1\dots n\},|T|=n_t} \widehat{s}^{\text{ho}}_T.$$

Let V' be an integer such that $V' \leqslant \frac{n}{n-n_t}$ and let $(T'_j)_{1\leqslant j\leqslant V'}$ be a collection of subsets satisfying the assumptions of Definition 6.3.2. Notice that

$$\frac{1}{n!}\sum_{\sigma\in\mathfrak{S}_n}\frac{1}{V'}\sum_{j=1}^{V'}\widehat{s}^{\text{ho}}_{\sigma(T'_j)} = \frac{1}{\binom{n}{n_t}}\sum_{T\subset\{1\dots n\},|T|=n_t}\widehat{s}^{\text{ho}}_T = \mathbb{E}\left[\widehat{s}^{\text{ho}}_{T_1}|D_n\right].$$

For any permutation σ, the collection $\sigma(T'_j)$ satisfies the same assumptions as the original T'_j, hence by Jensen's inequality,

$$\mathbb{E}\left[\left\|\mathbb{E}\left[\widehat{s}^{\text{ho}}_{T_1}|D_n\right] - s\right\|^2\right] \leqslant \mathbb{E}\left[\left\|\widehat{s}^{vf}_{n_t,V'} - s\right\|^2\right].$$

Therefore, by equation (6.109),

$$\mathbb{E}\left[\left\|\widehat{s}^{mc}_{n_t,V} - s\right\|^2 - \left\|\widehat{s}^{\text{ho}}_{T_1} - s\right\|^2\right] \leqslant \frac{V-1}{V}\left(\mathbb{E}\left[\left\|\widehat{s}^{vf}_{n_t,V'} - s\right\|^2 - \left\|\widehat{s}^{\text{ho}}_{T_1} - s\right\|^2\right]\right). \quad (6.110)$$

By equation (6.8) of theorem 6.4.4, which we have already proved,

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{s}^{vf}_{n_t,V'} - s\right\|^2 - \left\|\widehat{s}^{\text{ho}}_{T_1} - s\right\|^2\right] \leqslant -\frac{V'-1}{V'}\frac{n-n_t}{n_t}\frac{k_*(n_t)}{n_t} - \frac{V'-1}{V'}\Big[2\int_{-\infty}^{0}[F(1-F)](t)dt \\ + \int_0^{+\infty}[F(1-F)](t)dt\Big]\mathcal{E} + o(\mathcal{E}).\end{aligned} \quad (6.111)$$

Choosing $V' = \max\{j\in\mathbb{N} : \frac{1}{j} \geqslant \frac{n-n_t}{n}\}$ yields $\frac{1}{V'} \geqslant \frac{n-n_t}{n} \geqslant \frac{1}{V'+1}$, in particular $\frac{1}{V'} \sim \frac{n-n_t}{n}$ since $\frac{n-n_t}{n} \to 0$. It follows that

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{s}^{vf}_{n_t,V'} - s\right\|^2 - \left\|\widehat{s}^{\text{ho}}_{T_1} - s\right\|^2\right] \leqslant -\frac{n-n_t}{n_t}\frac{k_*(n_t)}{n_t} + [1+o(1)]\left(\frac{n-n_t}{n}\right)^2\frac{k_*(n_t)}{n_t} + o(\mathcal{E}) \\ - \mathcal{E}\left[2\int_{-\infty}^{0}[F(1-F)](t)dt + \int_0^{+\infty}[F(1-F)](t)dt\right].\end{aligned} \quad (6.112)$$

In conclusion, by equation (6.110),

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{s}^{mc}_{n_t,V}-s\right\|^2-\left\|\widehat{s}^{\mathrm{ho}}_{T_1}-s\right\|^2\right]\\
&\quad\leqslant -\frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*(n_t)}{n_t}+\frac{V-1}{V}[1+o(1)]\left(\frac{n-n_t}{n}\right)^2\frac{k_*(n_t)}{n_t}\\
&\qquad-\frac{V-1}{V}\mathcal{E}\left[2\int_{-\infty}^{0}[F(1-F)](t)dt+\int_0^{+\infty}[F(1-F)](t)dt\right]+o(\mathcal{E}).
\end{aligned}\tag{6.113}$$

6.7.5 A lower bound on $\int_{-\infty}^{+\infty}[F_{\hat{\alpha}_1}(1-F_{\hat{\alpha}_1})](t)dt$

In this section, we will state and prove the following result, which, together with equation (6.8) of Theorem 6.4.4(proved in the previous section), directly implies equation (6.10) of Theorem 6.4.4.

Claim 6.7.5.3 *Let* $\hat{\alpha}=\operatorname{argmin}_{\alpha\in\left[\frac{-k_*(n_t)}{\Delta};0\right[}\{f_n(\alpha)-W_{g_n(\alpha)}\}$. *There exist constants* $\varepsilon(\|s\|_\infty,\|s\|^2)>0$, $\delta(\|s\|_\infty)>0$, *such that for all* $n\in\mathbb{N}$, *there exists* $(\alpha_g,\alpha_d)\in\mathbb{R}^2$ *such that*

$$\alpha_d-\alpha_g\geqslant\delta \tag{6.114}$$

$$\forall\alpha\in[\alpha_g;\alpha_d],\varepsilon\leqslant F_{\hat{\alpha}}(\alpha)\leqslant 1-\varepsilon. \tag{6.115}$$

As a result, equation (6.10) *of Theorem 6.4.4 follows from equation* (6.8), *with* $\kappa_{ag}=\delta\varepsilon(1-\varepsilon)$.

We now prove claim 6.7.5.3. First, lemma 6.7.6 below implies that, for a two-sided brownian motion W and a function f reaching a unique minimum at 0, a bound on the distribution function of $\operatorname{argmin} f-W$ can be deduced from bounds on $|f(x)-f(y)|$.

Lemma 6.7.6 *Let* $I\subset I_0$ *be intervals of non-zero length containing* 0. *Let* $h,h_-:I_0\to\mathbb{R}$ *and* $h_+:I\to\mathbb{R}$ *be functions. Assume that* h,h_-,h_+ *are non-increasing on* $\mathbb{R}_-$, *non-decreasing on* $\mathbb{R}_+$ *and that* $h_+(0)=h_-(0)=0$. *Let* Z *be a continuous stochastic process on* I_0 *such that* h_--Z, h_+-Z *and* $h-Z$ *almost surely reach a unique minimum on* I_0 *or* I. *Assume also that:*

$$\forall x,y\in I_0, xy\geqslant 0\implies|h_-(x)-h_-(y)|\leqslant|h(x)-h(y)|, \tag{6.116}$$

$$\forall x,y\in I, xy\geqslant 0\implies|h(x)-h(y)|\leqslant|h_+(x)-h_+(y)|. \tag{6.117}$$

Then the following bound holds

$$\mathbb{P}\left(\operatorname*{argmin}_{t\in I\cup(I_0\cap\mathbb{R}_+)}\{h_+(t)\mathbb{I}_{t\leqslant 0}+h_-(t)\mathbb{I}_{t\geqslant 0}-W_t\}\leqslant z\right)\leqslant\mathbb{P}\left(\operatorname*{argmin}_{t\in I_0}\{h(t)-W_t\}\leqslant z\right) \tag{6.118}$$

$$\mathbb{P}\left(\operatorname*{argmin}_{t\in I_0}\{h(t)-W_t\}\leqslant z\right)\leqslant\mathbb{P}\left(\operatorname*{argmin}_{t\in I\cup(I_0\cap\mathbb{R}_-)}\{h_-(t)\mathbb{I}_{t\leqslant 0}+h_+(t)\mathbb{I}_{t\geqslant 0}-W_t\}\leqslant z\right). \tag{6.119}$$

Proof For any function $\phi : I_0 \to \mathbb{R}\cup\{+\infty\}$, let $A_{t,t'}(\phi)$ be the event $\{\phi(t)-Z_t\leqslant\phi(t')-Z_{t'}\}$.

By abuse of notation, we will also denote by h_+ the function equal to h_+ on the interval I and equal to $+\infty$ on $\mathbb{R}\backslash I$.

First, we will prove equation (6.118). For any function $\phi : I_0 \to \mathbb{R}\cup\{+\infty\}$ and all intervals J, J_0 such that $J\subset J_0\subset I_0$,

$$\begin{aligned}\left\{\operatorname*{argmin}_{t\in J_0}\phi(t)-Z_t\leqslant z\right\}&=\left\{\min_{t\in J_0,t\leqslant z}\phi(t)-Z_t\leqslant\min_{t'\in J_0,t\geqslant z}\phi(t')-Z_{t'}\right\}\\&=\bigcap_{t'\in J_0,t'\geqslant z}\left\{\min_{t\in J_0,t\leqslant z}\phi(t)-Z_t\leqslant\phi(t')-Z_{t'}\right\}\\&=\bigcup_{t\in J_0,t\leqslant z}\bigcap_{t'\in J_0\geqslant z}A_{t,t'}(\phi) \qquad (6.120)\\&\supset\bigcup_{\{t\in J:t\leqslant z\}}\bigcap_{t'\in J_0,t\geqslant z}A_{t,t'}(\phi). \qquad (6.121)\end{aligned}$$

Let $t\in I\cup(I_0\cap\mathbb{R}_+)$ and $t'\in I_0\cap[t;+\infty[$. Three cases are possible.

- If $t\leqslant t'\leqslant 0$, since $t\in I$ and $0\in I$ by assumption, $t'\in I$. Since h, h_+ are non-increasing on $\mathbb{R}_-\cap I$, by equation (6.117),

  $$h(t')-h(t)=-|h(t')-h(t)|\geqslant-|h_+(t')-h_+(t)|=h_+(t')-h_+(t).$$

  It follows that

  $$h_+(t)-Z_t\leqslant h_+(t')-Z_{t'}\iff Z_{t'}-Z_t\leqslant h_+(t')-h_+(t)\implies Z_{t'}-Z_t\leqslant h(t')-h(t),$$

  which yields

  $$A_{t,t'}\left(h_+\mathbb{I}_{\mathbb{R}_-}+h_-\mathbb{I}_{\mathbb{R}_+}\right)\subset A_{t,t'}(h).$$

- If $t\leqslant 0\leqslant t'$, $h_-(t')\leqslant h(t')$ and $h(t)\leqslant h_+(t)$, therefore $h_-(t')-h_+(t)\leqslant h(t')-h(t)$, which yields

  $$Z_{t'}-Z_t\leqslant h_-(t')-h_+(t)\implies Z_{t'}-Z_t\leqslant h(t')-h(t)$$

  and $A_{t,t'}\left(h_+\mathbb{I}_{\mathbb{R}_-}+h_-\mathbb{I}_{\mathbb{R}_+}\right)\subset A_{t,t'}(h)$.

- If $0 < t < t'$, since h_-, h are non-decreasing on $\mathbb{R}_+$, equation (6.116) implies that
$$h_-(t') - h_-(t) \leqslant h(t') - h(t),$$
which yields
$$Z_{t'} - Z_t \leqslant h_-(t') - h_-(t) \implies Z_{t'} - Z_t \leqslant h(t') - h(t)$$
and finally $A_{t,t'}\left(h_+\mathbb{I}_{\mathbb{R}_-} + h_-\mathbb{I}_{\mathbb{R}_+}\right) \subset A_{t,t'}(h)$.

We have proven that for all t, t' such that $t \in I \cup (I_0 \cap \mathbb{R}_+)$, $t' \in I_0$ and $t \leqslant t'$,
$$A_{t,t'}\left(h_+\mathbb{I}_{\mathbb{R}_-} + h_-\mathbb{I}_{\mathbb{R}_+}\right) \subset A_{t,t'}(h).$$
Hence by equation (6.121),
$$\left\{\operatorname*{argmin}_{t\in I_0} h(t) - Z_t \leqslant z\right\} \supset \bigcup_{\{t\in I\cup(I_0\cap\mathbb{R}_+):t\leqslant z\}} \bigcap_{t'\in I_0:t'\geqslant z} A_{t,t'}\left(h_+\mathbb{I}_{\mathbb{R}_-} + h_-\mathbb{I}_{\mathbb{R}_+}\right).$$
On the other hand, for all $t \in I\cup(I_0\cap\mathbb{R}_+)$ and all $t' \notin I\cup(I_0\cap\mathbb{R}_+)$, $A_{t,t'}\left(h_+\mathbb{I}_{\mathbb{R}_-} + h_-\mathbb{I}_{\mathbb{R}_+}\right)$ is the certain event (since $h_+(t') = +\infty$), therefore
$$\left\{\operatorname*{argmin}_{t\in\mathbb{R}} h(t) - Z_t \leqslant z\right\} \supset \bigcup_{\{t\in I\cup(I_0\cap\mathbb{R}_+):t\leqslant z\}} \bigcap_{t'\in I\cup(I_0\cap\mathbb{R}_+),t'\geqslant z} A_{t,t'}\left(h_+\mathbb{I}_{\mathbb{R}_-} + h_-\mathbb{I}_{\mathbb{R}_+}\right).$$
It follows from equation (6.120) that
$$\left\{\operatorname*{argmin}_{t\in I_0} h(t) - Z_t \leqslant z\right\} \supset \left\{\operatorname*{argmin}_{t\in I\cup(I_0\cap\mathbb{R}_+)} h_+(t)\mathbb{I}_{t<0} + h_-(t)\mathbb{I}_{t\geqslant 0} - Z_t \leqslant z\right\}.$$
This proves equation (6.118).

We will now prove equation (6.119). Let $(t, t') \in I_0^2$ be two real numbers such that $t \leqslant t'$. Three cases are again possible.

- If $t \leqslant t' \leqslant 0$, $h(t')-h(t) = -|h(t')-h(t)| \leqslant -|h_-(t')-h_-(t)| = h_-(t')-h_-(t)$ by equation (6.116), since h_-, h are non-increasing on $\mathbb{R}_-$. It follows that
$$Z_{t'}-Z_t \leqslant h(t')-h(t) \implies Z_{t'}-Z_t \leqslant h_-(t')-h_-(t) \implies h_-(t)-Z_t \leqslant h_-(t')-Z_{t'},$$
which yields
$$A_{t,t'}(h) \subset A_{t,t'}\left(h_-\mathbb{I}_{\mathbb{R}_-} + h_+\mathbb{I}_{\mathbb{R}_+}\right).$$

- If $t \leqslant 0 \leqslant t'$, $h_-(t') \leqslant h(t') \leqslant h_+(t')$ and $h_-(t) \leqslant h(t) \leqslant h_+(t)$, therefore $h_-(t') - h_+(t) \leqslant h(t') - h(t) \leqslant h_+(t') - h_-(t)$, which yields

$$Z_{t'} - Z_t \leqslant h(t') - h(t) \implies Z_{t'} - Z_t \leqslant h_+(t') - h_-(t)$$

  and $A_{t,t'}(h) \subset A_{t,t'}\left(h_-\mathbb{I}_{\mathbb{R}_-} + h_+\mathbb{I}_{\mathbb{R}_+}\right)$.

- If $0 < t < t'$, assume that $h(t) - Z_t \leqslant h(t') - Z_{t'}$, or equivalently $Z_{t'} - Z_t \leqslant h(t') - h(t)$. Then:
  - If $t' \notin I$, $h_+(t') = +\infty$ therefore $h_+(t) - Z_t \leqslant h_+(t') - Z_{t'}$.
  - If $t' \in I$, since $0 \in I$ by assumption, $t \in I$ therefore by equation (6.117),

$$Z_{t'} - Z_t \leqslant h(t') - h(t) \leqslant h_+(t') - h_+(t).$$

  This proves that $h(t) - Z_t \leqslant h(t') - Z_{t'} \implies h_+(t) - Z_t \leqslant h_+(t') - Z_{t'}$, so

$$A_{t,t'}(h) \subset A_{t,t'}\left(h_-\mathbb{I}_{\mathbb{R}_-} + h_+\mathbb{I}_{\mathbb{R}_+}\right).$$

In all cases, we have proven that for all $t \leqslant t'$,

$$A_{t,t'}(h) \subset A_{t,t'}\left(h_-\mathbb{I}_{\mathbb{R}_-} + h_+\mathbb{I}_{\mathbb{R}_+}\right).$$

By equation (6.120), it follows that for all $z \in \mathbb{R}$,

$$\left\{\operatorname*{argmin}_{t\in I_0} h(t) - Z_t \leqslant z\right\} \subset \left\{\operatorname*{argmin}_{t\in I_0} h_-(t)\mathbb{I}_{t\leqslant 0} + h_+(t)\mathbb{I}_{t\geqslant 0} - Z_t \leqslant z\right\}.$$

This proves equation (6.119). ■

The idea is now to apply lemma 6.7.6 to $Z = W$ and $h = f_n \circ g_n^{-1}$. To this end, it is necessary to bound $f_n \circ g_n^{-1}(y_2) - f_n \circ g_n^{-1}(y_1)$ for all y_1, y_2, uniformly in n. This is the purpose of lemma 6.7.7 below.

Lemma 6.7.7 *For all* $(\alpha_1, \alpha_2) \in \left[g_n\left(\frac{-k_*}{\Delta}\right); +\infty\right[^2$ *with the same sign,*

$$|f_n \circ g_n^{-1}(\alpha_2) - f_n \circ g_n^{-1}(\alpha_1)| \geqslant \left|\frac{1}{20\left\|s\right\|_\infty}(|\alpha_2| - 20\left\|s\right\|_\infty)_+ - \frac{1}{20\left\|s\right\|_\infty}(|\alpha_1| - 20\left\|s\right\|_\infty)_+\right|. \tag{6.122}$$

Furthermore,

- *If* $\Delta = \Delta_d$, *then for all* $\alpha_1, \alpha_2 \in \left[0; 4\left\|s\right\|^2\right]$,

$$|f_n \circ g_n^{-1}(\alpha_2) - f_n \circ g_n^{-1}(\alpha_1)| \leqslant \frac{|\alpha_2 - \alpha_1|}{4\left\|s\right\|^2}. \tag{6.123}$$

- *If $\Delta = \Delta_g$, then for all $\alpha_1, \alpha_2 \in \left[-4\|s\|^2; 0\right]$,*

$$|f_n \circ g_n^{-1}(\alpha_2) - f_n \circ g_n^{-1}(\alpha_1)| \leqslant \frac{|\alpha_2 - \alpha_1|}{4\|s\|^2}. \tag{6.124}$$

Proof Let $(\alpha_1, \alpha_2) \in \left[\frac{-k_*}{\Delta}; +\infty\right[^2$ be such that $\alpha_1\alpha_2 \geqslant 0$ and $\alpha_1 \leqslant \alpha_2$. If $0 < \alpha_1 < \alpha_2$, since f_n is non-decreasing on $\mathbb{R}_+$, $f_n(\alpha_2) - f_n(\alpha_1) \geqslant 0$, therefore by point 5 of Theorem 5.3.8,

$$g_n(\alpha_2) - g_n(\alpha_1) \leqslant \frac{8\|s\|_\infty}{(n-n_t)\mathfrak{e}}[f_n(\alpha_1) - f_n(\alpha_2)] + 12\|s\|_\infty [\alpha_2 - \alpha_1] \leqslant 12\|s\|_\infty [\alpha_2 - \alpha_1].$$

Thus, by lemma 6.9.7, for all $\alpha_1, \alpha_2 \in \mathbb{R}_+$ such that $\alpha_1 \leqslant \alpha_2$,

$$g_n^{-1}(\alpha_2) - g_n^{-1}(\alpha_1) \geqslant \frac{[\alpha_2 - \alpha_1]}{12\|s\|_\infty}.$$

Let α_1, α_2 be two real numbers such that $0 \leqslant \alpha_1 < \alpha_2$. Since $\alpha_1 > 0$, by the above equation, $g_n^{-1}(\alpha_1) \leqslant \frac{\alpha_1}{12\|s\|_\infty}$. Since f_n is convex and non-decreasing on $\mathbb{R}_+$, by lemma 6.9.6,

$$f_n(g_n^{-1}(\alpha_2)) - f_n(g_n^{-1}(\alpha_1)) \geqslant f_n\left(\frac{\alpha_2}{12\|s\|_\infty}\right) - f_n\left(\frac{\alpha_1}{12\|s\|_\infty}\right).$$

By the properties of f_n (lemma 6.6.3),

$$f_n(g_n^{-1}(\alpha_2)) - f_n(g_n^{-1}(\alpha_1)) \geqslant \left(\frac{\alpha_2}{12\|s\|_\infty} - 1\right)_+ - \left(\frac{\alpha_1}{12\|s\|_\infty} - 1\right)_+. \tag{6.125}$$

Now consider the case where $g_n\left(\frac{-k_*}{\Delta}\right) \leqslant \alpha_1 < \alpha_2 \leqslant 0$. For any $c \geqslant 0$, let $x_c \in \left[\frac{-k_*}{\Delta}; 0\right]$ be such that

$$\forall x \in \left[\frac{-k_*}{\Delta}; 0\right], x_c > x \implies \frac{f_n(x_c) - f_n(x)}{x - x_c} \leqslant c$$
$$x < x_c \implies \frac{f_n(x) - f_n(x_c)}{x_c - x} \geqslant c.$$

x_c exists by convexity of f_n.

If $g_n(x_c) < \alpha_1 < \alpha_2 \leqslant 0$, let $x_1 = g_n^{-1}(\alpha_1), x_2 = g_n^{-1}(\alpha_2)$, then $x_c < x_1 < x_2 \leqslant 0$, therefore by point 5 of Theorem 5.3.8 and by convexity of f_n,

$$\begin{aligned} g_n(x_2) - g_n(x_1) &\leqslant \frac{8\|s\|_\infty}{(n-n_t)\mathfrak{e}}[f_n(x_1) - f_n(x_2)] + 12\|s\|_\infty [x_2 - x_1] \\ &\leqslant 8\|s\|_\infty [f_n(x_1) - f_n(x_2)] + 12\|s\|_\infty [x_2 - x_1] (\text{ by lemma 6.6.2}) \\ &\leqslant 8c\|s\|_\infty [x_2 - x_1] + 12\|s\|_\infty [x_2 - x_1] \\ &\leqslant [8c + 12]\|s\|_\infty [x_2 - x_1]. \end{aligned} \tag{6.126}$$

Let $r(c) = [8c + 12] \|s\|_\infty$. By definition of x_1, x_2, equation (6.126) yields

$$g_n^{-1}(\alpha_2) - g_n^{-1}(\alpha_1) \geqslant \frac{\alpha_2 - \alpha_1}{r(c)}.$$

By the same argument, since $\alpha_1 < 0$, $g_n^{-1}(\alpha_1) \leqslant \frac{\alpha_1}{r(c)}$. Since $-f_n$ is non-decreasing and concave on $\left[\frac{-k_*}{\Delta}; 0\right]$, by lemma 6.9.6,

$$f_n(g_n^{-1}(\alpha_1)) - f_n(g_n^{-1}(\alpha_2)) \geqslant f_n\left(\frac{\alpha_1}{r(c)}\right) - f_n\left(\frac{\alpha_2}{r(c)}\right).$$

In conclusion, by the properties of f_n, for all α_1, α_2 such that $g_n(x_c) \leqslant \alpha_1 < \alpha_2 \leqslant 0$,

$$f_n(g_n^{-1}(\alpha_1)) - f_n(g_n^{-1}(\alpha_2)) \geqslant \left(\frac{\alpha_2}{r(c)} + 1\right)_+ - \left(\frac{\alpha_1}{r(c)} + 1\right)_+. \tag{6.127}$$

Let now

$$h_n = -8 \|s\|_\infty f_n + 12 \|s\|_\infty \mathrm{Id} \tag{6.128}$$

(where all functions are restricted to $\left[\frac{-k_*}{\Delta}; 0\right]$). By definition of h_n,

$$\begin{aligned} -f_n \circ h_n^{-1} &= \frac{1}{8 \|s\|_\infty} \left[-8 \|s\|_\infty f_n + 12 \|s\|_\infty \mathrm{Id}\right] \circ h_n^{-1} - \frac{12 \|s\|_\infty}{8 \|s\|_\infty} \mathrm{Id} \circ h_n^{-1} \\ &= \frac{1}{8 \|s\|_\infty} \mathrm{Id} - \frac{3}{2} h_n^{-1}. \end{aligned} \tag{6.129}$$

By definition of x_c and h_n, for any u_1, u_2 such that $-\frac{k_*}{\Delta} \leqslant u_1 < u_2 < x_c$,

$$h_n(u_2) - h_n(u_1) \geqslant 8c \|s\|_\infty [u_2 - u_1] + 12 \|s\|_\infty [u_2 - u_1] = r(c)[u_2 - u_1].$$

Hence by lemma 6.9.7, for any $(v_1, v_2) \in \left]h_n\left(-\frac{k_*}{\Delta}\right); h_n(x_c)\right[^2$ such that $v_1 < v_2$,

$$h_n^{-1}(v_2) - h_n^{-1}(v_1) \leqslant \frac{v_2 - v_1}{r(c)}.$$

Therefore, by equation (6.129), for any $(v_1, v_2) \in \left]h_n\left(-\frac{k_*}{\Delta}\right); h_n(x_c)\right[^2$ such that $v_1 \leqslant v_2$,

$$-f_n \circ h_n^{-1}(v_2) + f_n \circ h_n^{-1}(v_1) \geqslant \left[\frac{1}{8 \|s\|_\infty} - \frac{3}{2r(c)}\right] [v_2 - v_1].$$

If now (α_1, α_2) are such that $g_n\left(\frac{-k_*}{\Delta}\right) < \alpha_1 < \alpha_2 \leqslant g_n(x_c)$. Denote also $y_1 = g_n^{-1}(\alpha_1)$ and $y_2 = g_n^{-1}(\alpha_2)$. Since g_n is non-decreasing and $g_n(0) = 0$, it is also the

case that $\frac{-k_*}{\Delta} \leqslant y_1 < y_2 \leqslant x_c$, therefore

$$\begin{aligned} f_n \circ g_n^{-1}(\alpha_1) - f_n \circ g_n^{-1}(\alpha_2) &= f_n(y_1) - f_n(y_2) \\ &= -f_n \circ h_n^{-1}(h_n(y_2)) + f_n \circ h_n^{-1}(h_n(y_1)) \\ &\geqslant \left[\frac{1}{8\left\|s\right\|_\infty} - \frac{3}{2r(c)}\right][h_n(y_2) - h_n(y_1)] \\ &\geqslant \left[\frac{1}{8\left\|s\right\|_\infty} - \frac{3}{2r(c)}\right][g_n(y_2) - g_n(y_1)], \end{aligned}$$

by point 4. of Theorem 5.3.8, definition (6.128) of h_n and the inequality $(n-n_t)\mathfrak{e} \leqslant 1$ (which follows from lemma 6.6.2). By definition of y_1, y_2, this yields

$$f_n \circ g_n^{-1}(\alpha_1) - f_n \circ g_n^{-1}(\alpha_2) \geqslant \left[\frac{1}{8\left\|s\right\|_\infty} - \frac{3}{2r(c)}\right][\alpha_2 - \alpha_1]. \tag{6.130}$$

To sum up, let $c = 1$, then $r(1) = 20\left\|s\right\|_\infty$ and the following holds for all α_1, α_2 such that $g_n\left(\frac{-k_*}{\Delta}\right) \leqslant \alpha_1 \leqslant \alpha_2 \leqslant 0$.

- If $g_n(x_1) \leqslant \alpha_1 < \alpha_2 \leqslant -20\left\|s\right\|_\infty$, by equation (6.127),

$$f_n(g_n^{-1}(\alpha_1)) - f_n(g_n^{-1}(\alpha_2)) \geqslant \frac{\alpha_2 - \alpha_1}{20\left\|s\right\|_\infty}.$$

- If $\alpha_1 < \alpha_2 \leqslant g_n(x_1)$, by equation (6.130),

$$f_n(g_n^{-1}(\alpha_1)) - f_n(g_n^{-1}(\alpha_2)) \geqslant \frac{5-3}{40\left\|s\right\|_\infty}[\alpha_2 - \alpha_1] \geqslant \frac{\alpha_2 - \alpha_1}{20\left\|s\right\|_\infty}.$$

- If $\alpha_1 \leqslant g_n(x_1) \leqslant \alpha_2 \leqslant -20\left\|s\right\|_\infty$,

$$\begin{aligned} f_n(g_n^{-1}(\alpha_1)) - f_n(g_n^{-1}(\alpha_2)) &\geqslant f_n(g_n^{-1}(\alpha_1)) - f_n(g_n^{-1}(g_n(x_1)) \\ &\quad + f_n(g_n^{-1}(g_n(x_1))) - f_n(g_n^{-1}(\alpha_2)) \\ &\geqslant \frac{\alpha_2 - \alpha_1}{20\left\|s\right\|_\infty}, \end{aligned}$$

  by the two previous cases.

- If $\alpha_1 \leqslant -20\left\|s\right\|_\infty \leqslant \alpha_2$, since $f_n \circ g_n^{-1}$ is non-increasing on $\mathbb{R}_-$,

$$f_n(g_n^{-1}(\alpha_1)) - f_n(g_n^{-1}(\alpha_2)) \geqslant f_n(g_n^{-1}(\alpha_1)) - f_n(g_n^{-1}(-20\left\|s\right\|_\infty)) \geqslant \frac{-20\left\|s\right\|_\infty - \alpha_1}{20\left\|s\right\|_\infty}.$$

Thus, equation (6.122) holds for all (α_1, α_2) such that $g_n\left(\frac{-k_*}{\Delta}\right) \leqslant \alpha_1 \leqslant \alpha_2 \leqslant 0$. By equation (6.125) and since $\frac{1}{12\|s\|_\infty} \geqslant \frac{1}{20\|s\|_\infty}$, equation (6.122) is also true if $0 \leqslant \alpha_1 < \alpha_2$.

Let us now prove the upper bound (equations (6.124) and (6.123)). By point 4 of Theorem 5.3.8,

$$\forall(\alpha_1, \alpha_2) \in \left[\frac{-k_*}{\Delta}; +\infty\right[^2, g_n(\alpha_2) - g_n(\alpha_1) \geqslant 4\|s\|^2 [\alpha_2 - \alpha_1].$$

By lemma 6.9.7, this implies that

$$\forall(\alpha_1, \alpha_2) \in \left[\frac{-k_*}{\Delta}; +\infty\right[^2, |g_n^{-1}(\alpha_2) - g_n^{-1}(\alpha_1)| \leqslant \frac{1}{4\|s\|^2}[\alpha_2 - \alpha_1]. \tag{6.131}$$

As for f_n, there are two cases.

- If $\Delta = \Delta_d$, for all α_1, α_2 such that $0 \leqslant \alpha_1 \leqslant \alpha_2 \leqslant 1$, by lemma 6.6.3,

  $$|f_n(\alpha_2) - f_n(\alpha_1)| \leqslant \alpha_2 - \alpha_1. \tag{6.132}$$

  By equation (6.131), since g_n^{-1} increases and $g_n(0) = 0$, for all (α_1, α_2) such that $0 \leqslant \alpha_1 \leqslant \alpha_2 \leqslant 4\|s\|^2$,

  $$0 \leqslant g_n^{-1}(\alpha_1) \leqslant g_n^{-1}(\alpha_2) \leqslant 1.$$

  Hence by equation (6.132), for all (α_1, α_2) such that $0 \leqslant \alpha_1 \leqslant \alpha_2 \leqslant 4\|s\|^2$,

  $$\begin{aligned}|f_n \circ g_n^{-1}(\alpha_2) - f_n \circ g_n^{-1}(\alpha_1)| &\leqslant |g_n^{-1}(\alpha_2) - g_n^{-1}(\alpha_1)| \\ &\leqslant \frac{1}{4\|s\|^2}|\alpha_2 - \alpha_1| \text{ by equation (6.131).}\end{aligned}$$

  This proves equation (6.123) of lemma 6.7.7.

- If $\Delta = \Delta_g$, then $\frac{-k_*}{\Delta} \leqslant -1$ and for all (α_1, α_2) such that $-1 < \alpha_1 < \alpha_2 < 0$, by lemma 6.6.3,

  $$|f_n(\alpha_2) - f_n(\alpha_1)| \leqslant \alpha_2 - \alpha_1. \tag{6.133}$$

  By equation (6.131), since g_n^{-1} is increasing and $g_n(0) = 0$, for all (α_1, α_2) such that $-4\|s\|^2 \leqslant \alpha_1 \leqslant \alpha_2 \leqslant 0$, g_n^{-1} is defined and

  $$-1 \leqslant g_n^{-1}(\alpha_1) \leqslant g_n^{-1}(\alpha_2) \leqslant 0.$$

  Hence by equation (6.132), for all (α_1, α_2) such that $-4\|s\|^2 \leqslant \alpha_1 \leqslant \alpha_2 \leqslant 0$,

  $$\begin{aligned}|f_n \circ g_n^{-1}(\alpha_2) - f_n \circ g_n^{-1}(\alpha_1)| &\leqslant |g_n^{-1}(\alpha_2) - g_n^{-1}(\alpha_1)| \\ &\leqslant \frac{1}{4\|s\|^2}|\alpha_2 - \alpha_1| \text{ by equation (6.131).}\end{aligned}$$

  This proves equation (6.124) of lemma 6.7.7.

■

We can now prove claim 6.7.5.3. Remark first that

$$g_n(\hat{\alpha}) = \underset{u\in\left[g_n\left(\frac{-k_*}{\Delta}\right);+\infty\right[}{\operatorname{argmin}} f_n(g_n^{-1}(u)) - W_u.$$

There are two cases, both of which allow for the application of lemma 6.7.6.

- If $\Delta = \Delta_d$, by lemma 6.7.7, lemma 6.7.6 applies to $Z = W$, $h = f_n \circ g_n^{-1}$ with $h_- : x \mapsto \frac{1}{20\|s\|_\infty}(|x| - 20\|s\|_\infty)_+$, $I = [0; 4\|s\|^2]$, $I_0 = \left[g_n\left(\frac{-k_*}{\Delta}\right); +\infty\right[$ and $h_+ : x \mapsto \frac{x}{4\|s\|^2}$. It follows that for all $y \in \mathbb{R}$:

$$\mathbb{P}\left(\underset{u\in[0;+\infty[}{\operatorname{argmin}} \left\{\frac{1}{20\|s\|_\infty}(u - 20\|s\|_\infty)_+ - W_u\right\} \leqslant y\right) \leqslant \mathbb{P}(g_n(\hat{\alpha}) \leqslant y) \tag{6.134}$$

$$\mathbb{P}(g_n(\hat{\alpha}) \leqslant y) \leqslant \mathbb{P}\Big(\underset{u\in\left]g_n\left(\frac{-k_*}{\Delta}\right);4\|s\|^2\right]}{\operatorname{argmin}} \left\{\frac{1}{20\|s\|_\infty}(-u - 20\|s\|_\infty)_+ + \frac{(u)_+}{4\|s\|^2} - W_u\right\} \leqslant y\Big). \tag{6.135}$$

Let

$$\varepsilon(\|s\|_\infty, \|s\|^2) = \min\Big(\mathbb{P}\big(\underset{u\in[0;+\infty[}{\operatorname{argmin}}\big\{\tfrac{1}{20\|s\|_\infty}(u - 20\|s\|_\infty)_+ - W_u\big\} \leqslant 0, 2\big),$$
$$\mathbb{P}\big(\underset{u\in]-\infty;4\|s\|^2]}{\operatorname{argmin}} \big\{\tfrac{1}{20\|s\|_\infty}(-u - 20\|s\|_\infty)_+ + \frac{(u)_+}{4\|s\|^2} - W_u\big\} > 0.8\big)\Big). \tag{6.136}$$

By lemma 6.9.8, $\varepsilon(\|s\|_\infty, \|s\|^2) > 0$. By equations (6.135) and (6.134), for all $\alpha \in [g_n^{-1}(0.2); g_n^{-1}(0.8)]$:

$$\varepsilon \leqslant \mathbb{P}(g_n(\hat{\alpha}) \leqslant g_n(\alpha)) = \mathbb{P}(\hat{\alpha} \leqslant \alpha) \leqslant 1 - \varepsilon.$$

Set $\alpha_g = g_n^{-1}(0, 2)$ and $\alpha_d = g_n^{-1}(0, 8)$. Since f_n is non-decreasing on $\mathbb{R}_+$, by point 5 of Theorem 5.3.8, for all $(\alpha_1, \alpha_2) \in \mathbb{R}_+^2$,

$$|g_n(\alpha_2) - g_n(\alpha_1)| \leqslant 12\|s\|_\infty [\alpha_2 - \alpha_1].$$

Hence by lemma 6.9.7, for all $(u_1, u_2) \in \mathbb{R}_+^2$

$$|g_n^{-1}(u_2) - g_n^{-1}(u_1)| \geqslant \frac{|u_2 - u_1|}{12\|s\|_\infty}.$$

In particular :

$$\alpha_d - \alpha_g = g_n^{-1}(0, 8) - g_n^{-1}(0, 2) \geqslant \frac{0, 6}{12\|s\|_\infty} = \frac{0, 05}{\|s\|_\infty} > 0.$$

Moreover, by equation (6.131), $|\alpha_g| \leqslant |\alpha_d| \leqslant \frac{1}{4\|s\|^2}$.

- If $\Delta = \Delta_g$, by lemma 6.7.7, lemma 6.7.6 applies to $Z = W$, $h = f_n \circ g_n^{-1}$ with $h_- : x \mapsto \frac{1}{20\|s\|_\infty} (|x| - 20\|s\|_\infty)_+$, $I = [-4\|s\|^2; 0]$, $I_0 = \left[g_n\left(\frac{-k_*}{\Delta}\right); +\infty\right[$ and $h_+ : x \mapsto -\frac{x}{4\|s\|^2}$. It follows that for all $y \in \mathbb{R}$:

$$\mathbb{P}\left(\underset{u\in]-4\|s\|^2;+\infty]}{\operatorname{argmin}} \left\{\frac{1}{20\|s\|_\infty}(u - 20\|s\|_\infty)_+ + \frac{(u)_-}{4\|s\|^2} - W_u\right\} \leqslant y\right) \leqslant \mathbb{P}\left(g_n(\hat{\alpha}) \leqslant y\right) \tag{6.137}$$

$$\mathbb{P}\left(g_n(\hat{\alpha}) \leqslant y\right) \leqslant \mathbb{P}\Big(\underset{u\in\left]g_n\left(\frac{-k_*}{\Delta}\right);0\right]}{\operatorname{argmin}} \left\{\tfrac{1}{20\|s\|_\infty}(-u - 20\|s\|_\infty)_+ - W_u\right\} \leqslant y\Big). \tag{6.138}$$

Let

$$\begin{aligned}\varepsilon_d(\|s\|_\infty, \|s\|^2) = \\ \min\Big(&\mathbb{P}\big(\underset{u\in]-4\|s\|^2;+\infty]}{\operatorname{argmin}} \left\{\tfrac{1}{20\|s\|_\infty}(u - 20\|s\|_\infty)_+ + \frac{(u)_-}{4\|s\|^2} - W_u\right\} \leqslant -0,8\big),\\ &\mathbb{P}\big(\underset{u\in]-\infty;0]}{\operatorname{argmin}} \left\{\tfrac{1}{20\|s\|_\infty}(-u - 20\|s\|_\infty)_+ - W_u\right\} > -0.2\big)\Big).\end{aligned} \tag{6.139}$$

By distributional symmetry of W with respect to the map $x \mapsto -x$, one can see by comparison with (6.136) that $\varepsilon_d = \varepsilon > 0$. Therefore, by equations (6.138) and (6.137), for all $\alpha \in [g_n^{-1}(-0.8); g_n^{-1}(-0.2)]$:

$$\varepsilon \leqslant \mathbb{P}(g_n(\hat{\alpha}) \leqslant g_n(\alpha)) = \mathbb{P}(\hat{\alpha} \leqslant \alpha) \leqslant 1 - \varepsilon.$$

Let $\alpha_g = g_n^{-1}(-0,8)$ and $\alpha_d = g_n^{-1}(-0,2)$. Since $\Delta = \Delta_g$, for all $(\alpha_1, \alpha_2) \in [-1; 0]^2$ such that $\alpha_1 \leqslant \alpha_2$, $f_n(\alpha_1) - f_n(\alpha_2) \leqslant \alpha_2 - \alpha_1$. Hence for all $(\alpha_1, \alpha_2) \in [-1; 0]^2$, by point 5 of Theorem 5.3.8,

$$|g_n(\alpha_2) - g_n(\alpha_1)| \leqslant 8\|s\|_\infty |f_n(\alpha_1) - f_n(\alpha_2)| + 12\|s\|_\infty [\alpha_2 - \alpha_1] \leqslant 20\|s\|_\infty |\alpha_2 - \alpha_1|,$$

therefore by lemma 6.9.7, for all $(\alpha_1, \alpha_2) \in [g_n(-1); 0]^2$,

$$|g_n^{-1}(\alpha_2) - g_n^{-1}(\alpha_1)| \geqslant \frac{|\alpha_2 - \alpha_1|}{20\|s\|_\infty}.$$

Furthermore, by point 4 of Theorem 5.3.8, $g_n(-1) \leqslant -4\|s\|^2 < -1$, therefore in particular, $\alpha_d - \alpha_g = g_n^{-1}(-0,2) - g_n^{-1}(-0,8)| \geqslant \frac{0.6}{20\|s\|_\infty} = \frac{0,03}{\|s\|_\infty}$.

Thus, in all cases, $\alpha_d - \alpha_g \geqslant \frac{0.03}{\|s\|_\infty} > 0$. This proves claim 6.7.5.3.

6.8 Proof of the results of section 6.4.3

6.8.1 Proof of lemma 6.4.5

$\theta_j^2 \sim Lj^{-2\beta-1}$, there exists therefore an integer k_0 such that for all $k \geqslant k_0$,

$$\sum_{j=k+1}^{+\infty} \theta_j^2 \leqslant 2L \sum_{j=k+1}^{+\infty} \frac{1}{j^{2\beta+1}} \leqslant 2L \int_k^{+\infty} \frac{dx}{x^{2\beta+1}} = \frac{L}{\beta k^{2\beta+1}}.$$

In particular, $\mathrm{or}(n) = \inf_{k\in\mathbb{N}} \sum_{j=k+1}^{+\infty} \theta_j^2 + \frac{k}{n} \to 0$, which implies that $k_*(n) \to +\infty$.

For all $\varepsilon > 0$, let $k_1(\varepsilon)$ such that for all $k \geqslant k_1(\varepsilon)$,

$$(1-\varepsilon)Lk^{-2\beta-1} \leqslant \theta_k^2 \leqslant (1+\varepsilon)Lk^{-2\beta-1}.$$

Let $n_1(\varepsilon)$ such that $n \geqslant n_1(\varepsilon) \implies k_*(n) \geqslant \max(k_1(\varepsilon), \frac{1}{\varepsilon})$. Let $x_*(n) = \frac{k_*(n)}{(Ln)^{\frac{1}{2\beta+1}}}$, where by definition, $k_*(n)$ is the greatest integer such that:

$$\theta^2_{(k_*(n)} \geqslant \frac{1}{n}.$$

Therefore, for all $n \geqslant n_1(\varepsilon)$, $(1+\varepsilon)Lk_*(n)^{-2\beta-1} \geqslant \theta^2_{(k_*(n)} \geqslant \frac{1}{n}$, or in other words,

$$(1+\varepsilon)x_*(n)^{-2\beta-1}\frac{1}{n} \geqslant \frac{1}{n} \iff x_*(n) \leqslant (1+\varepsilon)^{\frac{1}{2\beta+1}}.$$

Furthermore, by definition $\theta^2_{k_*(n)+1} \leqslant \frac{1}{n}$. Since $n_1 \geqslant \frac{1}{\varepsilon}$, $(k_*(n)+1) \leqslant (1+\varepsilon)k_*(n)$ for all $n \geqslant n_1$,

$$n \geqslant n_1(\varepsilon) \implies (1-\varepsilon)L[(1+\varepsilon)k_*(n)]^{-2\beta-1} \leqslant \theta^2_{k_*(n)+1} \leqslant \frac{1}{n},$$

in other words,

$$\frac{1-\varepsilon}{(1+\varepsilon)^{2\beta+1}} x_*(n)^{-2\beta-1}\frac{1}{n} \leqslant \frac{1}{n} \iff x_*(n) \geqslant \frac{(1-\varepsilon)^{\frac{1}{2\beta+1}}}{1+\varepsilon}.$$

It follows that $x_*(n) \to 1$, or equivalently,

$$k_*(n) \sim (Ln)^{\frac{1}{2\beta+1}}, \tag{6.140}$$

which proves equation (6.13) of lemma 6.4.5. Furthermore, for all $n \geqslant n_1(\varepsilon)$,

$$\begin{aligned}
\frac{(1-\varepsilon)}{(1+\varepsilon)^{2\beta}} \frac{L}{2\beta} k_*(n)^{-2\beta} &\leqslant (1-\varepsilon)\frac{L}{2\beta}[k_*(n)+1]^{-2\beta} \\
&= (1-\varepsilon)L \int_{k_*(n)+1}^{+\infty} \frac{dx}{x^{2\beta+1}} \\
&\leqslant \sum_{j=k_*(n)+1}^{+\infty} (1-\varepsilon)Lj^{-2\beta-1} \\
&\leqslant \sum_{j=k_*(n)+1}^{+\infty} \theta_j^2 \\
&\leqslant \sum_{j=k_*(n)+1}^{+\infty} (1+\varepsilon)Lj^{-2\beta-1} \\
&\leqslant (1+\varepsilon)L \int_{k_*(n)}^{+\infty} \frac{dx}{x^{2\beta+1}} \\
&\leqslant (1+\varepsilon)\frac{L}{2\beta} k_*(n)^{-2\beta}.
\end{aligned}$$

As a consequence,

$$\sum_{j=k_*(n)+1}^{+\infty} \theta_j^2 \sim \frac{L}{2\beta} k_*(n)^{-2\beta} \sim \frac{L^{\frac{1}{2\beta+1}}}{2\beta n^{\frac{2\beta}{2\beta+1}}} \tag{6.141}$$

$$\frac{k_*(n)}{n} \sim \frac{L^{\frac{1}{2\beta+1}}}{n^{\frac{1}{2\beta+1}}} \tag{6.142}$$

$$\mathrm{or}(n) \sim \frac{2\beta+1}{2\beta} \frac{L^{\frac{1}{2\beta+1}}}{n^{\frac{1}{2\beta+1}}}, \tag{6.143}$$

which proves equation (6.14) of lemma 6.4.5. We now turn our attention to $\mathrm{or}(n_t) - \mathrm{or}(n)$. Since the sequence θ_j^2 is non-increasing, for all $j \geqslant k_*(n_t)$, $\theta_j^2 \geqslant \theta_{k_*(n_t)}^2 \geqslant \frac{1}{n_t} \geqslant \frac{1}{n}$, therefore $k_*(n) \geqslant k_*(n_t)$. Hence

$$\begin{aligned}
\mathrm{or}(n_t) - \mathrm{or}(n) &= \frac{k_*(n_t)}{n_t} + \sum_{j=k_*(n_t)+1}^{+\infty} \theta_j^2 - \frac{k_*(n)}{n} - \sum_{j=k_*(n)+1}^{+\infty} \theta_j^2 \\
&= k_*(n)\left[\frac{1}{n_t} - \frac{1}{n}\right] - \frac{k_*(n) - k_*(n_t)}{n_t} + \sum_{j=k_*(n_t)+1}^{k_*(n)} \theta_j^2.
\end{aligned}$$

By definition, for all $j \in [|k_*(n_t)+1; k_*(n)|]$, $\frac{1}{n} \leqslant \theta_j^2 < \frac{1}{n_t}$, therefore there exists $\eta \in [0;1]$ such that:

$$\begin{aligned}\mathrm{or}(n_t) - \mathrm{or}(n) &= \frac{k_*(n)}{n}\frac{n-n_t}{n_t} - \frac{k_*(n)-k_*(n_t)}{n_t} + \frac{k_*(n)-k_*(n_t)}{n_t} \\ &\quad - \eta[k_*(n)-k_*(n_t)]\left[\frac{1}{n_t}-\frac{1}{n}\right] \\ &= \frac{k_*(n)}{n}\frac{n-n_t}{n_t} - \eta\frac{k_*(n)-k_*(n_t)}{n}\frac{n-n_t}{n_t}.\end{aligned}$$

If $\frac{n_t(n)}{n} \to 1$, then by equation (6.140), $k_*(n) - k_*(n_t) = o(k_*(n))$, therefore

$$\mathrm{or}(n_t) - \mathrm{or}(n) \sim \frac{n-n_t}{n}\frac{k_*(n)}{n} \sim \frac{n-n_t}{n}\frac{L^{\frac{1}{2\beta+1}}}{n^{\frac{1}{2\beta+1}}}, \tag{6.144}$$

which proves equation (6.15) of lemma 6.4.5. We now prove equation (6.16). Let $\varepsilon > 0$. Assume that

$$\Delta_d(n) \geqslant \varepsilon k_*(n). \tag{6.145}$$

We now show that for excessively large values of ε, equation (6.145) is in contradiction with the hypothesis that $\frac{n_t}{n-n_t} \leqslant \eta_0 k_*(n)$ (equation (6.16) of lemma 6.4.5). Indeed, if $\frac{n_t}{n-n_t} \leqslant \eta_0 k_*(n)$, by equation (6.145),

$$\sqrt{\frac{n_t}{(n-n_t)\Delta_d}} \leqslant \sqrt{\frac{\eta_0}{\varepsilon}}, \text{ therefore } \theta^2_{k_*(n_t)+\Delta_d} \geqslant \left[1-\sqrt{\frac{\eta_0}{\varepsilon}}\right]\theta^2_{k_*(n_t)+1}. \tag{6.146}$$

Since $k_*(n) \geqslant k_*(n_t)$ by definition, equation (6.145) implies that $\Delta_d(n) \geqslant \varepsilon k_*(n_t)$. Hence, by hypothesis (6.12) and equation (6.145), for all $\delta > 0$ and for all sufficiently large n,

$$\frac{1+\delta}{1-\delta}(1+\varepsilon)^{-2\beta-1}\theta^2_{k_*(n_t)+1} \geqslant [1+\delta]L(1+\varepsilon)^{-2\beta-1}[k_*(n_t)+1]^{-2\beta-1} \geqslant \theta^2_{\lfloor(1+\varepsilon)k_*(n_t)\rfloor} \geqslant \theta^2_{k_*(n_t)+\Delta_d}. \tag{6.147}$$

Let δ be such that $\frac{1+\delta}{1-\delta} = 1+\varepsilon$. By equations (6.146) and (6.147), there exists an integer $n_1(\varepsilon)$, such that for all $n \geqslant n_1(\varepsilon)$,

$$(\text{equation(6.145)}) \wedge \left(\frac{n_t}{n-n_t} \leqslant \eta_0 k_*(n)\right) \implies \frac{1}{(1+\varepsilon)^{2\beta}} \geqslant 1-\sqrt{\frac{\eta_0}{\varepsilon}}. \tag{6.148}$$

The convexity of the function $x \mapsto \frac{1}{(1+x)^{2\beta+1}}$ on $[0;1]$ and the fact that $\beta > 1$ imply that

$$1-\frac{3\varepsilon}{4} \geqslant 1-\left[1-\frac{1}{2^{2\beta}}\right]\varepsilon \geqslant \frac{1}{(1+\varepsilon)^{2\beta}},$$

hence by equation (6.148), for all $n \geqslant n_1(\varepsilon)$,

$$\begin{aligned}(\text{equation(6.145)}) \wedge \left(\frac{n_t}{n-n_t} \leqslant \eta_0 k_*(n)\right) &\implies 1 - \frac{3\varepsilon}{4} \geqslant 1 - \sqrt{\frac{\eta_0}{\varepsilon}} \\ &\implies \varepsilon \leqslant \left(\tfrac{16}{9}\eta_0\right)^{\frac{1}{3}}.\end{aligned}$$

In conclusion, by equation (6.145) defining ε,

$$\forall n \geqslant n_1\left(\left(\tfrac{16}{9}\eta_0\right)^{\frac{1}{3}}\right), \frac{n_t}{n-n_t} \leqslant \eta_0 k_*(n_t) \implies \frac{\Delta_d(n)}{k_*(n_t)} \leqslant \left(\tfrac{16}{9}\eta_0\right)^{\frac{1}{3}} = \varepsilon_1. \tag{6.149}$$

Let $\delta > 0$. Since $k_*(n_t) \to +\infty$, for all n large enough,

$$\theta^2_{k_*(n_t)+\Delta_d+1} \geqslant \theta^2_{\lfloor(1+\frac{\varepsilon_1}{2})k_*(n_t)\rfloor+\Delta_d}.$$

Since $\frac{n_t(n)}{n} \to 1$ by assumption, $k_*(n_t) = k_*(n_t) \sim k_*(n)$ by equation (6.140). Therefore, by equation (6.149), for n large enough, if $\frac{n_t}{n-n_t} \leqslant \eta_0 k_*(n_t)$,

$$\theta^2_{k_*(n_t)+\Delta_d+1} \geqslant \theta^2_{\lfloor(1+2\varepsilon_1)k_*(n_t)\rfloor} \geqslant (1-\delta)\frac{\theta^2_{k_*(n_t)}}{(1+2\varepsilon_1)^{2\beta+1}} \geqslant (1+2\varepsilon_1)^{-2\beta-1}\frac{1-\delta}{n_t}.$$

It follows by definition of Δ_d that

$$\left(1 - \sqrt{\frac{n_t}{(n-n_t)\Delta_d}}\right) \geqslant n_t\theta^2_{k_*(n_t)+\Delta_d+1} \geqslant \frac{1-\delta}{(1+2\varepsilon_1)^{2\beta+1}},$$

hence

$$\begin{aligned}1 - \sqrt{\frac{n_t}{(n-n_t)\Delta_d}} &\geqslant (1-\delta)(1-2(2\beta+1)\varepsilon_1) \\ &\geqslant 1 - \delta - 2(2\beta+1)\varepsilon_1 \\ \sqrt{\frac{n_t}{(n-n_t)\Delta_d}} &\leqslant \delta + 2(2\beta+1)\varepsilon_1.\end{aligned}$$

Since $\varepsilon_1 = \left(\frac{16}{9}\eta_0\right)^{\frac{1}{3}}$, there exists an integer $n_2(\beta, \eta_0)$ such that for all $n \geqslant n_2(\beta, \eta_0)$,

$$\frac{n_t}{n-n_t} \leqslant \eta_0 k_*(n_t) \implies \sqrt{\frac{n_t}{(n-n_t)\Delta_d}} \leqslant 2(2\beta+1)(2\eta_0)^{\frac{1}{3}},$$

hence:

$$\forall n \geqslant n_2(\beta,\eta_0), \frac{n_t}{n-n_t} \leqslant \eta_0 k_*(n_t) \implies \frac{n-n_t}{n_t}\Delta_d(n) \geqslant \frac{(2\eta_0)^{-\frac{2}{3}}}{4(2\beta+1)^2}. \tag{6.150}$$

To conclude, remark that

$$\begin{aligned}\frac{\mathcal{E}}{\mathfrak{e}} &= \mathcal{E} \times \sqrt{\frac{n-n_t}{\mathcal{E}}} \\ &= \sqrt{(n-n_t)\mathcal{E}} \\ &= \sqrt{(n-n_t)\frac{\Delta}{n_t}} \\ &\geqslant \sqrt{\frac{n-n_t}{n_t}\Delta_d(n)},\end{aligned}$$

which yields the following by equation (6.150).

$$\forall n \geqslant n_2(\eta_0), \frac{n_t}{n-n_t} \leqslant \eta_0 k_*(n_t) \implies \frac{\mathcal{E}(n)}{\mathfrak{e}(n)} \geqslant \frac{(2\eta_0)^{-\frac{1}{3}}}{2(2\beta+1)}. \tag{6.151}$$

This proves equation (6.16) of lemma 6.4.5.

Now, for all $\varepsilon_0 > 0$,

$$\begin{aligned}\frac{(n-n_t)^2}{n_t^2} \leqslant \frac{1}{\varepsilon_0 k_*(n_t)} &\implies \frac{(n-n_t)^2}{nn_t} \leqslant \frac{n_t}{\varepsilon_0 n k_*(n_t)} \\ &\implies \varepsilon_0 \frac{(n-n_t)^2}{nn_t}\frac{k_*(n_t)}{n_t} \leqslant \frac{1}{n} \\ &\implies \varepsilon_0 \frac{n-n_t}{n_t}\frac{k_*(n_t)}{n_t} \leqslant \frac{1}{n-n_t} \leqslant \mathcal{E}(n) \text{ by lemma 6.6.2.}\end{aligned}$$

Since by definition $k_*(n) \geqslant k_*(n_t)$, this yields

$$\frac{n-n_t}{n_t} \leqslant \frac{1}{\sqrt{\varepsilon_0 k_*(n)}} \implies \frac{n-n_t}{n_t} \leqslant \frac{1}{\sqrt{\varepsilon_0 k_*(n_t)}} \implies \mathcal{E}(n) \geqslant \varepsilon_0 \frac{n-n_t}{n_t}\frac{k_*(n_t)}{n_t}, \tag{6.152}$$

which proves equation (6.17) of lemma 6.4.5.

Let now $\eta \geqslant 0$ be such that $\frac{n_t}{n-n_t} \geqslant \eta k_*(n) \geqslant \eta k_*(n_t)$. By definition of Δ_d (Definition 6.4.1),

$$\sqrt{\frac{n_t}{(n-n_t)\Delta_d}} \leqslant 1,$$

therefore

$$\Delta \geqslant \Delta_d \geqslant \frac{n_t}{n-n_t} \geqslant \eta k_*(n_t).$$

It follows that

$$\frac{n_t}{n-n_t} \geqslant \eta k_*(n) \implies \mathcal{E}(n) \geqslant \eta \frac{k_*(n_t)}{n_t},$$

hence by equations (6.142) and (6.143),

$$\frac{n_t}{n-n_t} \geqslant \eta k_*(n) \implies \mathcal{E}(n) \geqslant [1-o(1)]\eta\frac{2\beta}{2\beta+1}\mathrm{or}(n_t). \tag{6.153}$$

This proves equation (6.18) of lemma 6.4.5.

6.8.2 Proof of corollary 6.4.6

By theorem 6.4.4,

$$\mathbb{E}\left[\left\|\widehat{s}^{vf}_{n_t,V}-s\right\|^2\right]-\mathrm{or}(n) \leqslant \mathbb{E}\left[\left\|\widehat{s}^{\mathrm{ho}}_{T_1}-s\right\|^2\right]-\mathrm{or}(n_t)+\mathrm{or}(n_t)-\mathrm{or}(n) \\ -\frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*(n_t)}{n_t}-\frac{V-1}{V}\kappa_{ag}\mathcal{E}+o(\mathcal{E}).$$

It follows by Theorem 6.4.3 that

$$\mathbb{E}\left[\left\|\widehat{s}^{vf}_{n_t,V}-s\right\|^2\right]-\mathrm{or}(n) \leqslant \kappa_{ho}\mathfrak{e}+\mathrm{or}(n_t)-\mathrm{or}(n)-\frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*(n_t)}{n_t}-\frac{V-1}{V}\kappa_{ag}\mathcal{E}+o(\mathcal{E}). \tag{6.154}$$

Let κ_{ag} be as in Theorem 6.4.4. By equation (6.15) of lemma 6.4.5, for all large enough n,

$$\mathrm{or}(n_t)-\mathrm{or}(n) \leqslant \frac{n-n_t}{n}\frac{k_*(n_t)}{n_t}+\kappa_{ag}\frac{\varepsilon_0}{12}\frac{n-n_t}{n_t},$$

therefore

$$\mathrm{or}(n_t)-\mathrm{or}(n)-\frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*(n_t)}{n_t} \leqslant \left[\frac{n}{n_tV}+\frac{\kappa_{ag}\varepsilon_0 n}{12n_t}\right]\frac{n-n_t}{n}\frac{k_*(n_t)}{n_t},$$

hence for all $V \geqslant \lceil\frac{12}{\kappa_{ag}\varepsilon_0}\rceil$,

$$\mathrm{or}(n_t)-\mathrm{or}(n)-\frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*(n_t)}{n_t} \leqslant \kappa_{ag}\frac{\varepsilon_0}{6}\frac{n-n_t}{n_t}\frac{k_*(n_t)}{n_t}.$$

It follows by equation (6.17) of lemma 6.4.5 that for all $V \geqslant \lceil\frac{12}{\kappa_{ag}\varepsilon_0}\rceil$ and all n large enough,

$$\mathrm{or}(n_t)-\mathrm{or}(n)-\frac{V-1}{V}\frac{n-n_t}{n_t}\frac{k_*(n_t)}{n_t} \leqslant \frac{\kappa_{ag}}{5}\mathcal{E}.$$

Thus by equation (6.154), for all $V \geqslant \max\left(\lceil\frac{12}{\kappa_{ag}\varepsilon_0}\rceil, 5\right)$ and all n large enough,

$$\mathbb{E}\left[\left\|\widehat{s}^{vf}_{n_t,V}-s\right\|^2\right]-\mathrm{or}(n) \leqslant \kappa_{ho}\mathfrak{e}+\frac{\kappa_{ag}}{5}\mathcal{E}-\kappa_{ag}\mathcal{E}+\frac{\kappa_{ag}}{V}\mathcal{E}+o(\mathcal{E}) \\ \leqslant \kappa_{ho}\mathfrak{e}-\frac{2\kappa_{ag}}{5}\mathcal{E}. \tag{6.155}$$

Let:

$$\eta_0 = \sup\left\{\eta > 0 : \frac{(2\eta)^{-\frac{1}{3}}}{4\beta+2} \geqslant 5\frac{\kappa_{ho}}{\kappa_{ag}}\right\}.$$

By equation (6.18) of lemma 6.4.5, for all sufficiently large n, $\frac{n_t}{n-n_t} \leqslant \eta_0 k_*(n) \implies \mathfrak{e} \leqslant \frac{\kappa_{ag}}{5\kappa_{ho}}\mathcal{E}$, therefore by equation (6.155), for all sufficiently large n, if $V \geqslant \max\left(\lceil\frac{12}{\kappa_{ag}\varepsilon_0}\rceil, 5\right)$, then

$$\frac{n_t}{n-n_t} \leqslant \eta_0 k_*(n) \implies \mathbb{E}\left[\left\|\widehat{s}^{vf}_{n_t,V} - s\right\|^2\right] - \mathrm{or}(n) \leqslant -\frac{\kappa_{ag}}{5}\mathcal{E} < 0. \tag{6.156}$$

This proves equation (6.20) of corollary 6.4.6. Let $\eta \in]0;\eta_0]$. If $\eta k_*(n) \leqslant \frac{n_t}{n-n_t}$, then à fortiori $\sqrt{\varepsilon_0 k_*(n)} \leqslant \frac{n_t}{n-n_t}$ for all $\varepsilon_0 > 0$ and all sufficiently large n. In conclusion, by equation (6.156) above and equation (6.18) of lemma 6.4.5, for all $V \geqslant 5$ and all sufficiently large n (depending on η_0, η),

$$\eta k_*(n) \leqslant \frac{n_t}{n-n_t} \leqslant \eta_0 k_*(n) \implies \mathbb{E}\left[\left\|\widehat{s}^{vf}_{n_t,V} - s\right\|^2\right] \leqslant \left(1 - \frac{\kappa_{ag}\eta}{5}\frac{2\beta}{2\beta+1} + o(1)\right)\mathrm{or}(n).$$

This proves equation (6.21) of corollary 6.4.6.

6.9 Appendix

6.9.1 Results proven in the previous chapter

Lemma 6.9.1 *(Lemma 5.5.1) Let X be a random variable belonging to $[-1;1]$, with pdf s. For all $j \in \mathbb{N}$, let $\theta_j = \langle s, \psi_j\rangle$. Then*

$$\forall k_0 \leqslant k, \sum_{j=k_0}^{k} |\,Var(\psi_j) - 1| \leqslant \|\theta\|_{\ell^1} = \sum_{j=0}^{+\infty} |\langle s, \psi_j\rangle|.$$

Proposition 6.9.2 *(Proposition 5.5.3) Under the hypotheses of section 6.2.2, for all integers $k_0 < k$, with probability greater than $1 - e^{-y}$:*

$$\left|\sum_{j=k_0+1}^{k} (\hat{\theta}_j^T - \theta_j)^2 - \frac{|k-k_0|}{n_t}\right| \leqslant \frac{\|\theta\|_{\ell^1}}{n_t} + C\sqrt{y+\log n}\left[\frac{\sqrt{|k-k_0|}}{n_t} + \frac{|k-k_0|}{n_t^{\frac{5}{4}}}\right].$$

In particular , there exists a constant $\kappa_1 = \kappa_1(\|s\|_\infty, c_1, \|\theta\|_{\ell^1})$ such that for all α_1, α_2 such that $(\alpha_1\Delta, \alpha_2\Delta) \in \mathbb{N}^2$ and $\alpha_1 < \alpha_2$, with probability greater than $1 - e^{-y}$,

$$\left|\sum_{j=k_*+\alpha_1\Delta}^{k_*+\alpha_2\Delta} (\hat{\theta}_j^T - \theta_j)^2 - \frac{|k-k_0|}{n_t}\right| \leqslant \kappa_1(\alpha_2-\alpha_1)[\log n + y]^2 \times n^{-\min(\frac{1}{12}, \frac{\delta_3}{2})}\mathfrak{e}(n). \tag{6.157}$$

Claim 6.9.2.1 *(Claim 5.5.5.1) Let* $u_2 = \min\left(\frac{2\delta_2}{3\rho_1}, \delta_3\right)$. *Let* x *be a non-negative real number. Let* a_x, b_x *be such that* $a_x \leqslant 0 \leqslant b_x$ *and* $\max(f_n(a_x), f_n(b_x)) \leqslant x$. *Assume also that* $a_x\Delta - 1 \geqslant \frac{-k_*}{\Delta}$. *There exists a constant* $\kappa_3 \geqslant 0$ *such that for all* $j \in [a_x\Delta; b_x\Delta + 1]$,

$$\left|f_n\left(\tfrac{j}{\Delta}\right) - f_n\left(\tfrac{j-1}{\Delta}\right)\right| \leqslant \kappa_3(1+x)^2 n^{-u_2} \tag{6.158}$$

$$\theta^2_{k_*+j} \leqslant \kappa_3(1+x)^2 n^{-u_2}\mathfrak{e}. \tag{6.159}$$

6.9.2 Technical arguments

Claim 6.9.2.2 *There exist constants* n_3, κ_9 *such that for all* $x > 0$,

$$n > \max(n_3, \kappa_9(1+x)^{\frac{3}{2}}) \implies a_x \geqslant \frac{-k_*(n_t)+1}{\Delta}, b_x \leqslant \frac{n - n_t - k_*(n_t) - 1}{\Delta}$$

In particular,

$$\forall n \geqslant \max(n_3, \kappa_9(1+x)^3), a_x = \inf\left\{\frac{j}{\Delta} : j \in \mathbb{Z} \cap \left[\frac{-k_*(n_t)}{\Delta}; +\infty\right[, f_n(\tfrac{j}{\Delta}) \leqslant x\right\}$$

$$b_x = \sup\left\{\frac{j}{\Delta} : j \in \mathbb{Z} \cap \left[\frac{-k_*(n_t)}{\Delta}; +\infty\right[, f_n(\tfrac{j}{\Delta}) \leqslant x\right\}.$$

Proof By definition, $\mathrm{or}(n_t) = \frac{k_*(n_t)}{n_t} + \sum_{j=k_*(n_t)+1}^{+\infty} \theta_j^2$, therefore $k_*(n_t) \leqslant n_t \mathrm{or}(n_t)$. Moreover, by assumption,

$$\mathrm{or}(n_t) \leqslant \min_{k\in\mathbb{N}} \frac{k}{n_t} + \frac{c_1}{k^2} \leqslant \frac{2+c_1}{n_t^{\frac{2}{3}}}, \tag{6.160}$$

thus $k_*(n_t) \leqslant [2+c_1]n_t^{\frac{1}{3}}$. By hypothesis H5 of section 6.2.2, $n - n_t \geqslant n^{\frac{2}{3}}$, so It follows that

$$n - n_t - k_*(n_t) - 1 \geqslant n^{\frac{2}{3}} - [2+c_1]n_t^{\frac{1}{3}} - 1.$$

Hence, there exists a constant $n_3(c_1)$ such that for all $n \geqslant n_3$,

$$n - n_t - k_*(n_t) - 1 \geqslant \frac{1}{2}n^{\frac{2}{3}}.$$

Furthermore, by lemma 6.6.2 and equation (6.160),

$$\Delta(n) = n_t\mathcal{E}(n) \leqslant 2n_t\mathrm{or}(n_t) + \frac{n_t}{n - n_t} \leqslant 2[2+c_1]n_t^{\frac{1}{3}} + n_t^{\frac{1}{3}} \leqslant [5+2c_1]n^{\frac{1}{3}}.$$

It follows that for all $n \geqslant n_3(c_1)$, $\frac{n-n_t-k_*-1}{\Delta} \geqslant \frac{n^{\frac{1}{3}}}{10+4c_1}$. As a result, for any $n \geqslant n_3$ and any $x > 0$, by lemma 6.6.3,

$$\frac{n^{\frac{1}{3}}}{10+4c_1} > (1+x) \implies f_n\left(\tfrac{n-n_t-k_*-1}{\Delta}\right) > x \implies b_x < \frac{n-n_t-k_*-1}{\Delta}.$$

Thus,

$$n > \max(n_3, [10+4c_1]^3(1+x)^3) \implies b_x < \frac{n-n_t-k_*-1}{\Delta}.$$

As for a_x, by definition of f_n,

$$\begin{aligned} f_n\left(\tfrac{-k_*}{\Delta}\right) &= \frac{1}{\mathfrak{e}} \sum_{j=0}^{k_*(n_t)} \left[\theta_j^2 - \frac{1}{n_t}\right] \\ &\geqslant \frac{\theta_0^2 - \frac{1}{n_t}}{\mathfrak{e}} \\ &\geqslant (n-n_t)\left[1 - \frac{1}{n_t}\right] \text{ by lemma 6.6.2} \\ &\geqslant n^{\frac{2}{3}}\left[1 - \frac{2}{n}\right] \text{ by assumptions (H6)}\textit{and}\text{(H5)}. \end{aligned}$$

It follows that for all $n > \max\left(4, (2x)^{\frac{3}{2}}\right), f_n(-\frac{k_*}{\Delta}) > x$. Since $\Delta a_x \in \mathbb{Z}$ and $f_n(a_x) \leqslant x$, this implies that $a_x \geqslant \frac{-k_*+1}{\Delta}$. This proves the first part of the claim. The rest follows by definition of a_x, b_x. ■

Proposition 6.9.3 *Let $I \subset \mathbb{R}$ be a closed interval such that $0 \in I$. Let f be a continuous function that is non-increasing on $I \cap \mathbb{R}_-$, non-decreasing on $I \cap \mathbb{R}_+$, g be a non-decreasing function on I, and suppose that $f(0) = g(0) = 0$. Assume that there exists two non-negative constants a, b such that for all $\alpha \in I$,*

$$|g(\alpha)| \leqslant \max\{af(\alpha), b\}.$$

Let $\hat{\alpha} = \operatorname{argmin}_{\alpha \in I} f(\alpha) - W_{g(\alpha)}$ Then for all $x \geqslant 0$,

$$\mathbb{P}\left(f(\hat{\alpha}) \geqslant \frac{b}{a} + 3ax\right) \leqslant \sqrt{2}\frac{e^{\frac{1}{3}}}{1-e^{\frac{-1}{3}}}e^{-x}.$$

In particular ,

$$\mathbb{E}\left[f(\hat{\alpha})\right] \leqslant \frac{b}{a} + 3a + 2 + \frac{3}{2}\log 2 + 3\log 3.$$

Proof

$$\mathbb{P}\left(f(\hat{\alpha}) \geqslant \frac{b}{a} + 3ax\right) = \mathbb{P}\left(\left(f(\hat{\alpha}) \geqslant \frac{b}{a} + 3ax\right) \wedge \hat{\alpha} \geqslant 0\right) + \mathbb{P}\left(\left(f(\hat{\alpha}) \geqslant \frac{b}{a} + 3ax\right) \wedge \hat{\alpha} \leqslant 0\right)$$

Consider without loss of generality the event $\hat{\alpha} \in \mathbb{R}_+$. Assume first that there exists $x \in I, x \geqslant 0$ such that $f(x) \geqslant \frac{b}{a} + 3a$. Let

$$i_M = \sup\left\{i \in \mathbb{N} : \frac{b}{a} + ai \leqslant \sup_{x \in I \cap \mathbb{R}_+} f(x)\right\}.$$

(in particular, $i_M = +\infty$ if f is unbounded). By assumption, $i_M \geqslant 3$. For all $i \in \mathbb{N}$ s.t. $i \leqslant i_M - 1$, let $\alpha_i > 0$ be such that $f(\alpha_i) = \frac{b}{a} + ai$ (α_i exists by continuity of f). Fix some integer i such that $2 \leqslant i \leqslant i_M - 1$. Then

$$\begin{aligned}
\mathbb{P}\left(\hat{\alpha} \in [\alpha_i; \alpha_{i+1}]\right) &\leqslant P\left(\min_{\alpha \in [\alpha_i; \alpha_{i+1}]} f(\alpha) - W_{g(\alpha)} \leqslant 0\right) \\
&\leqslant P\left(f(\alpha_i) + \min_{\alpha \in [\alpha_i; \alpha_{i+1}]} -W_{g(\alpha)} \leqslant 0\right) \\
&\leqslant P\left(f(\alpha_i) + \min_{u \in [g(\alpha_i); g(\alpha_{i+1})]} -W_u \leqslant 0\right) \\
&\leqslant P\left(f(\alpha_i) + \min_{u \in [0; g(\alpha_{i+1})]} -W_u \leqslant 0\right) \\
&\leqslant P\left(f(\alpha_i) - \sqrt{g(\alpha_{i+1})}|Z| \leqslant 0\right) \\
&\leqslant P\left(|Z| \geqslant \frac{f(\alpha_i)}{\sqrt{g(\alpha_{i+1})}}\right), \qquad (6.161)
\end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$. On the other hand, for all $x > 0$,

$$\begin{aligned}
\mathbb{P}\left(|Z| \geqslant x\right) &= \frac{2}{\sqrt{2\pi}} \int_x^{+\infty} \frac{1}{t} \times t e^{-\frac{t^2}{2}} dt \\
&= \left[\frac{2e^{-\frac{t^2}{2}}}{\sqrt{2\pi}t}\right]_x^{+\infty} - \int_x^{+\infty} \left(-\frac{1}{t^2}\right)\left(-\frac{2e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\right) \\
&= \frac{2e^{-\frac{x^2}{2}}}{\sqrt{2\pi}x} - \int_x^{+\infty} \frac{2e^{-\frac{t^2}{2}}}{\sqrt{2\pi}t^2} \\
&\leqslant \frac{2e^{-\frac{x^2}{2}}}{\sqrt{2\pi}x}. \qquad (6.162)
\end{aligned}$$

Since $f(\alpha_i) = \frac{b}{a} + ai \geqslant \frac{b}{a}$, $g(\alpha_i) \leqslant af(\alpha_i)$, therefore

$$\begin{aligned}\frac{f(\alpha_i)}{\sqrt{g(\alpha_{i+1})}} &\geqslant \frac{\frac{b}{a} + ai}{\sqrt{b + a^2 i + a^2}} \\ &\geqslant \frac{\frac{b}{a} + ai}{\sqrt{\frac{3}{2}(b + a^2 i)}} \\ &\geqslant \sqrt{\frac{2}{3a}}\sqrt{\frac{b}{a} + ai} \\ &\geqslant \sqrt{\frac{2i}{3}}.\end{aligned}$$

Thus, by equations (6.161) and (6.162),

$$\begin{aligned}\mathbb{P}\left(\hat{\alpha} \in [\alpha_i; \alpha_{i+1}]\right) &\leqslant \frac{2}{\sqrt{2\pi}}\sqrt{\frac{3}{2i}}e^{-\frac{i}{3}} \\ &\leqslant \frac{1}{\sqrt{2}}e^{-\frac{i}{3}}.\end{aligned}$$

Now, if $i_M < +\infty$ (in other words if f is bounded on $I \cap \mathbb{R}_+$), then for all $\alpha \geqslant 0$, $f(\alpha) \leqslant \frac{b}{a} + a(i_M + 1)$, hence $g(\alpha) \leqslant b + a^2(i_M + 1)$. It follows that

$$\begin{aligned}\mathbb{P}(\hat{\alpha} \geqslant \alpha_{i_M}) &\leqslant \mathbb{P}\left(f(\alpha_{i_M}) - \sup_{\alpha \geqslant \alpha_{i_M}} W_{g(\alpha)} \leqslant 0\right) \\ &\leqslant \mathbb{P}\left(\frac{b}{a} + ai_M \leqslant \sup_{u \leqslant b + a^2(i_M+1)} W_u\right) \\ &\leqslant \mathbb{P}\left(|Z| \geqslant \frac{\frac{b}{a} + ai_M}{b + a^2 i_M + a^2}\right) \\ &\leqslant \frac{1}{\sqrt{2}}e^{-\frac{i_M}{3}} \text{ by the same arguments as above.}\end{aligned}$$

Therefore in all cases, for all $x \geqslant 2$, using the convention $\alpha_{i_M} = +\infty$ if $i_M = +\infty$,

whenever $i_M \geqslant 3$,

$$\begin{aligned}\mathbb{P}\left(\hat{\alpha} \geqslant 0, f(\hat{\alpha}) \geqslant \frac{b}{a} + ax\right) &\leqslant \sum_{i=\lfloor x \rfloor}^{i_M - 1} \mathbb{P}\left(\hat{\alpha} \in [\alpha_i; \alpha_{i+1}]\right) + \mathbb{P}(\hat{\alpha} \geqslant \alpha_{i_M}) \\ &\leqslant \frac{1}{\sqrt{2}} \sum_{i=\lfloor x \rfloor}^{+\infty} e^{-\frac{i}{3}} \\ &\leqslant \frac{1}{\sqrt{2}} \frac{1}{1 - e^{-\frac{1}{3}}} e^{-\frac{\lfloor x \rfloor}{3}} \\ &\leqslant \frac{1}{\sqrt{2}} \frac{e^{\frac{1}{3}}}{1 - e^{-\frac{1}{3}}} e^{-\frac{x}{3}}.\end{aligned}$$

If now $\sup_{x \in I \cap \mathbb{R}_+} f(x) < \frac{b}{a} + 3a$, then trivially, for all $x \geqslant 1$, $\mathbb{P}\left(f(\hat{\alpha}) \geqslant \frac{b}{a} + 3ax\right) = 0$. Conclude by remarking that for any $x \leqslant 2$,

$$\frac{1}{\sqrt{2}} \frac{e^{\frac{1-x}{3}}}{1 - e^{-\frac{1}{3}}} \geqslant 1 \geqslant \mathbb{P}\left(f(\hat{\alpha}) \geqslant \frac{b}{a} + 3ax\right).$$

The same argument applies symmetrically to the case $\hat{\alpha} \leqslant 0$, hence the result. ■

Claim 6.9.3.1 *Let $\hat{\alpha} \in \operatorname{argmin}_{\alpha \in \left[\frac{-k_*(n_t)}{\Delta}; +\infty\right[} f_n(\alpha) - W_{g_n(\alpha)}$, where g_n satisfies properties 1-5 of Theorem 5.3.8. There exists a constant $\kappa_7 = \kappa_7(\|s\|_\infty)$ such that for all $x > 0$,*

$$\mathbb{P}\left(f_n(\hat{\alpha}) \geqslant \kappa_7(1 + x)\right) \leqslant e^{-x}. \tag{6.163}$$

Furthermore, there exists a constant $n_2 \in \mathbb{N}$ such that, for all $n \geqslant n_2$ and all $x \geqslant 2\kappa_7(1 + \log n)$,

$$\mathbb{P}\left(\hat{\alpha} \notin [a_x; b_x]\right) \leqslant \frac{1}{n}.$$

Proof By point 5 of theorem 5.3.8, since $(n - n_t)\mathfrak{e} \leqslant 1$, for all $\alpha \in \left[\frac{-k_*(n_t)}{\Delta}; +\infty\right[$,

- If $|\alpha| \leqslant 2$,

$$|g_n(\alpha)| \leqslant 8 \|s\|_\infty f_n(\alpha) + 12 \|s\|_\infty \times 2 \leqslant \max\left(16 \|s\|_\infty f_n(\alpha), 48 \|s\|_\infty\right)$$

- If $|\alpha| \geqslant 2$, $\frac{|\alpha|}{2} \leqslant |\alpha| - 1 \leqslant f_n(\alpha)$ by lemma 6.6.3, therefore

$$|g_n(\alpha)| \leqslant 8 \|s\|_\infty f_n(\alpha) + 12 \|s\|_\infty \alpha \leqslant 32 \|s\|_\infty f_n(\alpha).$$

In all cases,

$$|g_n(\alpha)| \leqslant \max\left(32\,\|s\|_\infty f_n(\alpha), 48\,\|s\|_\infty\right).$$

By proposition 6.9.3 applied to $f = f_n$, $g = g_n$, for all $y > 0$,

$$\mathbb{P}\left(f_n(\hat{\alpha}) \geqslant \frac{3}{2} + 96\,\|s\|_\infty y\right) \leqslant \sqrt{2}\frac{e^{\frac{2}{3}}}{e^{\frac{1}{3}}-1}e^{-y}.$$

This proves equation (6.163). Let $x > 0$. Assume first that $a_x\Delta > -k_*$, then by claim 6.9.2.1,

$$f_n(a_x) + \kappa_3(1+x)^2 n^{-u_2} \geqslant f_n(a_x - \tfrac{1}{\Delta})$$
$$f_n(b_x) + \kappa_3(1+x)^2 n^{-u_2} \geqslant f_n(b_x + \tfrac{1}{\Delta}).$$

On the other hand, by Definition 6.6.1 of a_x, b_x , necessarily

$$\min\left(f_n(a_x - \tfrac{1}{\Delta}), f_n(b_x + \tfrac{1}{\Delta})\right) \geqslant x.$$

It follows that

$$\min\left(f_n(a_x), f_n(b_x)\right) \geqslant x - \kappa_3(1+x)^2 n^{-u_2}.$$

Since f_n is non-increasing on $\mathbb{R}_-$ and non-decreasing on $\mathbb{R}_+$, for all $x \geqslant 2\kappa_7(1+\log n)$,

$$\min\left(f_n(a_x), f_n(b_x)\right) \geqslant 2\kappa_7(1+\log n) - \kappa_3(1+2\kappa_7)^2(1+\log n)^2 n^{-u_2}.$$

There exists therefore a constant $n_2 \in \mathbb{N}$ such that, for all $n \geqslant n_2$ and all $x \geqslant 2\kappa_7(1+\log n)$,

$$\min_{\alpha \notin [a_x; b_x]} f_n(\alpha) = \min\left(f_n(a_x), f_n(b_x)\right) \geqslant \kappa_7(1+\log n).$$

By equation (6.163) which we already proved, this implies that $\mathbb{P}\left(\hat{\alpha} \notin [a_x; b_x]\right) \leqslant \frac{1}{n}$. Now if $a_x\Delta = -k_*$, obviously $\mathbb{P}(\hat{\alpha} < a_x) = 0$, and $\mathbb{P}(\hat{\alpha} > b_x) \leqslant \frac{1}{n}$ by the same argument as above. In conclusion,

$$\forall n \geqslant n_2, \forall x \geqslant 2\kappa_7(1+\log n), \mathbb{P}\left(\hat{\alpha} \notin [a_x; b_x]\right) \leqslant \frac{1}{n}.$$

This proves the result. ■

Lemma 6.9.4 *Let X be a continuous, non-negative random variable and let $f \in L^1(e^{-x}dx)$ be a continuous, positive and non-decreasing function, such that for all $x \in \mathbb{R}$, $\mathbb{P}\left(X \geqslant f(x)\right) \leqslant e^{-x}$. Let Z be a $[0;1]-$valued random variable, such that $\mathbb{P}(Z \leqslant \varepsilon) \geqslant 1-\delta$, for two real numbers $\varepsilon, \delta \in (0;1)$. Then:*

$$\mathbb{E}\left[ZX\right] \leqslant \varepsilon\mathbb{E}[X] + \int_{-\log(\delta)}^{+\infty} f(x)e^{-x}dx.$$

Proof Let E be the event $\{Z > \varepsilon\}$. By its definition,

$$\begin{aligned}\mathbb{E}\left[ZX\right] &\leqslant \varepsilon\mathbb{E}[X\mathbb{I}_{E^c}] + \mathbb{E}[ZX\mathbb{I}_E] \\ &\leqslant \varepsilon\mathbb{E}[X] + \mathbb{E}[X\mathbb{I}_E].\end{aligned}$$

Let $a \geqslant 0$ be a real number such that $\mathbb{P}(X \geqslant a) = \mathbb{P}(E)$ (a exists since the distribution of X is atomless). Let I_a denote the event $\{X \in [a; +\infty[\}$. Then

$$\begin{aligned}\mathbb{E}[X\mathbb{I}_E] &= \mathbb{E}[X\mathbb{I}_{E\cap I_a}] + \mathbb{E}[X\mathbb{I}_{E\cap I_a^c}] \\ &\leqslant \mathbb{E}[X\mathbb{I}_{E\cap I_a}] + a\mathbb{P}[E \cap I_a^c] \\ &\leqslant \mathbb{E}[X\mathbb{I}_{E\cap I_a}] + a\left(\mathbb{P}(E) - \mathbb{P}(E \cap I_a)\right) \\ &= \mathbb{E}[X\mathbb{I}_{E\cap I_a}] + a\left(\mathbb{P}(I_a) - \mathbb{P}(E \cap I_a)\right) \\ &\leqslant \mathbb{E}[X\mathbb{I}_{E\cap I_a}] + \mathbb{E}[X\mathbb{I}_{I_a\cap E^c}] \\ &= \mathbb{E}[X\mathbb{I}_{X\geqslant a}] \\ &= \int_0^{+\infty} \mathbb{P}\left(X\mathbb{I}_{X\geqslant a} \geqslant x\right) dx \\ &= \int_0^{+\infty} \min\left(\mathbb{P}(X \geqslant a), \mathbb{P}(X \geqslant x)\right) dx \\ &\leqslant \int_0^{+\infty} \delta \wedge \mathbb{P}(X \geqslant x) dx \\ &\leqslant \int_0^{+\infty} [\delta \wedge \mathbb{P}(X \geqslant f(x))] f'(x) dx \\ &\leqslant \int_0^{+\infty} [\delta \wedge e^{-x}] f'(x) dx \text{ since } f' \geqslant 0.\end{aligned}$$

By integration by parts and since $\delta \leqslant 1$, it follows that

$$\begin{aligned}\mathbb{E}[X\mathbb{I}_E] &\leqslant \delta \int_0^{-\log\delta} f'(x)dx + \int_{-\log\delta}^{+\infty} e^{-x} f'(x)dx \\ &\leqslant \delta[f(-\log\delta) - f(0)] + \left[e^{-x}f(x)\right]_{-\log\delta}^{+\infty} + \int_{-\log\delta}^{+\infty} e^{-x}f(x)dx \\ &\leqslant \int_{-\log\delta}^{+\infty} e^{-x}f(x)dx \text{ since } f \text{ is non-negative.}\end{aligned}$$

■

Claim 6.9.4.1 *Let Z be a $[0;1]-$valued random variable such that $\mathbb{P}(Z \leqslant \varepsilon) \geqslant 1-\delta$. There exists an integer $n_1 = n_1\left(\|s\|_\infty, c_1, \|\theta\|_{\ell^1}, \delta_3\right)$ such that for all $n \geqslant n_1$*

and all $(\alpha_1, \alpha_2) \in \left[-\frac{k_*}{\Delta}; \frac{n-n_t-k_*}{\Delta}\right]^2$ *such that* $\alpha_1 < \alpha_2$,

$$\mathbb{E}\left[Z \sum_{j=k_*+\alpha_1\Delta+1}^{k_*+\alpha_2\Delta} \left(\hat{\theta}_j^T - \theta_j\right)^2\right] \leqslant \tfrac{3}{2}(\alpha_2-\alpha_1)\left[\varepsilon+\delta\right]\mathcal{E}.$$

Proof By proposition 6.9.2, $\mathbb{P}\left(\sum_{j=k_*+\alpha_1\Delta+1}^{k_*+\alpha_2\Delta} \left(\hat{\theta}_j^T - \theta_j\right)^2 \geqslant h(y)\right) \leqslant 1 - e^{-y}$, where:

$$h(y) = (\alpha_2-\alpha_1)\mathcal{E} + \kappa_1(\alpha_2-\alpha_1)(y+\log n)^2 n^{-\min(\frac{1}{12},\frac{\delta_3}{2})}\mathfrak{e}.$$

Furthermore $\mathbb{E}\left[\sum_{j=k_*+\alpha_1\Delta+1}^{k_*+\alpha_2\Delta} \left(\hat{\theta}_j^T - \theta_j\right)^2\right] \leqslant (\alpha_2-\alpha_1)\frac{\Delta}{n_t} + \frac{\|\theta\|_{\ell^1}}{n_t}$, so by lemma 6.9.4,

$$\mathbb{E}\left[Z \sum_{j=k_*+\alpha_1\Delta+1}^{k_*+\alpha_2\Delta} \left(\hat{\theta}_j^T - \theta_j\right)^2\right] \leqslant \varepsilon(\alpha_2-\alpha_1)\mathcal{E} + \varepsilon\frac{\|\theta\|_{\ell^1}}{n_t} + \int_{-\log\delta}^{+\infty} h(y)e^{-y}dy. \quad (6.164)$$

The integral can be calculated as follows.

$$\int_{-\log\delta}^{+\infty} h(y)e^{-y}dy = \delta(\alpha_2-\alpha_1)\mathcal{E} + \kappa_1(\alpha_2-\alpha_1)n^{-\frac{1}{12}\wedge\frac{\delta_3}{2}}\mathfrak{e}\int_{-\log\delta}^{+\infty}(y+\log n)^2 e^{-y}dy,$$

with

$$\begin{aligned}
\int_{-\log\delta}^{+\infty}(y+\log n)^2 e^{-y}dy &= n\int_{-\log\delta+\log n}^{+\infty} u^2 e^{-u}du \\
&= \delta\left(-\log\delta+\log n\right)^2 + 2n\int_{-\log\delta+\log n}^{+\infty} ue^{-u}du \\
&= \delta\left(-\log\delta+\log n\right)^2 + 2\delta\left(-\log\delta+\log n\right) + 2n\int_{-\log\delta+\log n} e^{-u}du \\
&= \delta\left(-\log\delta+\log n\right)^2 + 2\delta\left(-\log\delta+\log n\right) + 2\delta \\
&\leqslant 5\delta\log\left(\frac{n}{\delta}\right)^2.
\end{aligned}$$

Therefore:

$$\int_{-\log\delta}^{+\infty} h(y)e^{-y}dy \leqslant \delta(\alpha_2-\alpha_1)\mathcal{E} + 5\kappa_1(\alpha_2-\alpha_1)\delta\log\left(\frac{n}{\delta}\right)^2 n^{-\frac{1}{12}\wedge\frac{\delta_3}{2}}\mathfrak{e}.$$

For any $\delta \geqslant \frac{1}{n}$, $\log\left(\frac{n}{\delta}\right)^2 n^{-\frac{1}{12}\wedge\frac{\delta_3}{2}} \leqslant 4\log^2 n n^{-\frac{1}{12}\wedge\frac{\delta_3}{2}} \leqslant 4\log^2 n n^{-\delta_3}$ and $\mathfrak{e} \leqslant \mathcal{E}$ by lemma 6.6.2, therefore there exists a constant n_1' such that for all $n \geqslant n_1'$,

$$\int_{-\log\delta}^{+\infty} h(y)e^{-y}dy \leqslant \delta(\alpha_2-\alpha_1)\mathcal{E} + \delta(\alpha_2-\alpha_1)\frac{\mathfrak{e}}{2} \leqslant \delta(\alpha_2-\alpha_1)\frac{3\mathcal{E}}{2},$$

thus by equation 6.164,

$$\mathbb{E}\left[Z\sum_{j=k_*+\alpha_1\Delta+1}^{k_*+\alpha_2\Delta}\left(\hat{\theta}_j^T-\theta_j\right)^2\right]\leqslant\varepsilon(\alpha_2-\alpha_1)\mathcal{E}+\varepsilon\frac{\|\theta\|_{\ell^1}}{n_t}+\delta(\alpha_2-\alpha_1)\frac{3\mathcal{E}}{2}.$$

Since by lemma 6.6.2 and hypothesis H5 of section 6.2.2, $\frac{\|\theta\|_{\ell^1}}{n_t}=\frac{\|\theta\|_{\ell^1}}{n-n_t}\frac{n-n_t}{n_t}\leqslant\|\theta\|_{\ell^1}\mathcal{E}n^{-\delta_3}$, there exists a constant n_1 such that for all $n\geqslant n_1$,

$$\mathbb{E}\left[Z\sum_{j=k_*+\alpha_1\Delta+1}^{k_*+\alpha_2\Delta}\left(\hat{\theta}_j^T-\theta_j\right)^2\right]\leqslant\tfrac{3}{2}\left[\varepsilon+\delta\right](\alpha_2-\alpha_1)\mathcal{E}.$$

■

Lemma 6.9.5 *Let $(Z_i)_{1\leqslant i\leqslant V}$ be a finite family of exchangeable random variables taking values in a Hilbert space H. Assume that $\mathbb{E}\left[\|Z_1\|^2\right]<+\infty$. Then for all $x\in H$,*

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{V}\sum_{i=1}^{V}Z_i-x\right\|^2&=\frac{1}{V}\mathbb{E}\left[\|Z_1-x\|^2\right]+\frac{V-1}{V}\mathbb{E}\left[\langle Z_1-x,Z_2-x\rangle\right]\\&=\|\mathbb{E}Z_1-x\|^2+\frac{1}{V}\mathbb{E}\|Z_1-\mathbb{E}Z_1\|^2+\frac{V-1}{V}\mathbb{E}\left[\langle Z_1-\mathbb{E}Z_1,Z_2-\mathbb{E}Z_2\rangle\right].\end{aligned}$$

Proof

$$\begin{aligned}\left\|\frac{1}{V}\sum_{i=1}^{V}Z_i-x\right\|^2&=\frac{1}{V^2}\sum_{i=1}^{V}\sum_{j=1}^{V}\langle Z_i-x;Z_j-x\rangle\\&=\frac{1}{V^2}\sum_{i=1}^{V}\|Z_i-x\|^2+\frac{1}{V^2}\sum_{i=1}^{V}\sum_{j=1,j\neq i}^{V}\langle Z_i-x;Z_j-x\rangle.\end{aligned}$$

Since the variables Z_i are exchangeable,

$$\mathbb{E}\left[\left\|\frac{1}{V}\sum_{i=1}^{V}Z_i-x\right\|^2\right]=\frac{1}{V}\mathbb{E}\left[\|Z_1-x\|^2\right]+\frac{V-1}{V}\mathbb{E}\left[\langle Z_1-x,Z_2-x\rangle\right],$$

which yields the first equation. Moreover, for all random variables $U,V\in H$ which have a finite second moment,

$$\begin{aligned}\mathbb{E}\left[\langle U-x,V-x\rangle\right]&=\mathbb{E}\left[\langle U-\mathbb{E}[U]+\mathbb{E}[U]-x,\langle V-\mathbb{E}[V]+\mathbb{E}[V]-x\rangle\right]\\&=\mathbb{E}\left[\langle U-\mathbb{E}[U],V-\mathbb{E}[V]\rangle\right]+\langle\mathbb{E}[U]-x,\mathbb{E}[V]-x\rangle,\end{aligned}$$

which allows to derive the second equation. ■

Lemma 6.9.6 *Let f be a non-decreasing function. If f is convex and $(u_1, u_2), (v_1, v_2)$ are such that $u_1 \leqslant v_1 \leqslant u_2 \leqslant v_2$,*

$$v_2 - v_1 \geqslant u_2 - u_1 \implies f(v_2) - f(v_1) \geqslant f(u_2) - f(u_1).$$

If now f is concave and $v_1 \leqslant u_1$, then

$$v_2 - v_1 \geqslant u_2 - u_1 \geqslant 0 \implies f(v_2) - f(v_1) \geqslant f(u_2) - f(u_1).$$

Proof Assume first that $u_1 \leqslant v_1 \leqslant u_2 \leqslant v_2$ and that f is convex. Then the function $h : x \mapsto f(x + v_2 - v_1) - f(x)$ is non-decreasing by convexity of f and non-negativity of $v_2 - v_1$. Therefore, $h(v_1) \geqslant h(u_1)$, or equivalently:

$$f(v_2) - f(v_1) \geqslant f(u_1 + v_2 - v_1) - f(u_1).$$

Since f is non-decreasing, if $v_2 - v_1 \geqslant u_2 - u_1$, then $f(u_1 + v_2 - v_1) \geqslant f(u_2)$, therefore

$$f(v_2) - f(v_1) \geqslant f(u_2) - f(u_1).$$

If now f is concave non-decreasing and $v_1 \leqslant u_1$, $h : x \mapsto f(x + v_2 - v_1) - f(x)$ is non-increasing, therefore $h(v_1) \geqslant h(u_1)$, that is

$$f(v_2) - f(v_1) \geqslant f(u_1 + v_2 - v_1) - f(v_1).$$

Since f is non-decreasing, if $v_2 - v_1 \geqslant u_2 - u_1$, then $f(u_1 + v_2 - v_1) \geqslant f(u_2)$, hence

$$f(v_2) - f(v_1) \geqslant f(u_2) - f(u_1).$$

■

Lemma 6.9.7 *Let $g : I \to J$ be a bijection with domain an interval I. If there exists a constant $\mu > 0$ such that*

$$\forall (x, y) \in I^2, |g(y) - g(x)| \geqslant \mu |y - x|,$$

then

$$\forall (u, v) \in J^2, |g^{-1}(v) - g^{-1}(u)| \leqslant \frac{|y - x|}{\mu}.$$

Conversely, if there exists a constant $L > 0$ such that:

$$\forall (x, y) \in J^2, |g(y) - g(x)| \leqslant L |y - x|,$$

then

$$\forall (u, v) \in J^2, |g^{-1}(v) - g^{-1}(u)| \geqslant \frac{|y - x|}{L}.$$

Proof Let $(y_1, y_2) \in J^2$ and $(x_1, x_2) \in I^2$ be such that $g(x_1) = y_1, g(x_2) = y_2$. The equation $|g(x_1) - g(x_2)| \geqslant \mu|x_1 - x_2|$ can be rewritten as $|y_1 - y_2| \geqslant \mu|g^{-1}(y_2) - g^{-1}(y_1)|$. By the same argument, equation $|g(x_1) - g(x_2)| \leqslant L|x_1 - x_2|$ can be rewritten as $|y_1 - y_2| \leqslant L|g^{-1}(y_2) - g^{-1}(y_1)|$. ■

Lemma 6.9.8 *Let $J \subset \mathbb{R}$ be a non-empty interval and let f be a continuous function on J. Let $(W_t)_{t\in\mathbb{R}}$ be a two-sided brownian motion such that $W_0 = 0$. Assume that the process $f - W$ reaches almost surely a unique minimum on J, and define:*

$$\hat{t} = \operatorname*{argmin}_{t\in J} f(t) - W_t.$$

Then for any $a \in \mathbb{R}$ and any interval $I = [a; +\infty[\cap J$ or $I =]-\infty; a]\cap J$ of non-zero length,

$$\mathbb{P}(\hat{t} \in I) > 0.$$

Proof By distributional symmetry of W, assume without loss of generality that $I = [a; +\infty[\cap J$. Then by uniqueness of the minimum:

$$\mathbb{P}(\hat{t} \in I) = \mathbb{P}\left(\inf_{t\in J; t\geqslant a} f(t) - W_t + W_a \leqslant \inf_{t\in J; t<a} f(t) - W_t + W_a\right).$$

Let $b > a$ such that $[a; b] \subset I$, which exists since I has non-zero length by assumption. Then

$$\mathbb{P}(\hat{t} \in I) \geqslant \mathbb{P}\left(\inf_{t\in[a;b]} f(t) - W_t + W_a \leqslant \inf_{t\in J; t<a} f(t) - W_t + W_a\right).$$

The continuous function f is bounded on the closed interval $[a; b]$ by a constant R, therefore

$$\mathbb{P}(\hat{t} \in I) \geqslant \mathbb{P}\left(\inf_{t\in[a;b]} -W_t + W_a \leqslant -R + \inf_{t\in J; t<a} f(t) - W_t + W_a\right).$$

By the Markov property, both sides of the above inequality are independent. Furthermore, by the reflexion principle, $\min_{t\in[a;b]} -W_t + W_a \sim -\sqrt{b-a}|Z_0|$, where $Z_0 \sim \mathcal{N}(0, 1)$. Therefore,

$$\mathbb{P}(\hat{t} \in I) \geqslant \mathbb{E}\left[2\bar{\Phi}\left(\tfrac{R}{\sqrt{b-a}} - \tfrac{1}{\sqrt{b-a}} \inf_{t\in J; t<a} f(t) - W_t + W_a\right)\right],$$

where $\bar{\Phi} : x \mapsto \mathbb{P}(Z_0 \geqslant x)$. By assumption, $\inf_{t\in J; t<a} f(t) - W_t + W_a \geqslant f(\hat{t}) - W_{\hat{t}} + W_a > -\infty$ almost surely, therefore by positivity of $\bar{\Phi}$,

$$\mathbb{P}(\hat{t} \in I) > 0.$$

■

Chapter 7

Conclusion and Perspectives

This thesis has proven that Agghoo and the hold-out satisfy a general oracle inequality (Theorem 3.7.3). This general theorem was applied in two settings, kernel methods and sparse regression, where it yields novel bounds on the risk of the hold-out (and Agghoo). Finally, a precise study of Agghoo was conducted in the case of least-squares density estimation using Fourier series estimators. This study shows that depending on the choice of its parameters, Agghoo can simply improve on the remainder term in the oracle inequality for the hold-out, perform better than the oracle trained on n_t data and even beat the oracle trained on the full sample, by as much as a constant factor.

7.1 Oracle inequalities for the hold-out

The first part of this thesis focused on improving theoretical guarantees for the hold-out to allow for unbounded loss functions. A new oracle inequality was proved, which allows for unbounded losses and more general margin hypotheses (Theorem 3.7.3). This Theorem was applied to the cases of kernel methods (Chapter 3) and sparse linear regression (Chapter 4). Theorem 3.7.3 has as its main assumption a *margin hypothesis* of the form

$$P\left(\gamma(t_1) - \gamma(t_2) - c_{t_1,t_2}\right)^2 \leqslant \left(w(\sqrt{\ell(s,t_1)}) + w(\sqrt{\ell(s,t_2)})\right)^2, \tag{7.1}$$

where t_1, t_2 are any two of the given estimators (trained on a training sample), c_{t_1,t_2} is a constant and w is a *subquadratic* function, more precisely, a function such that $\frac{w(x)}{x^2}$ is non-increasing. Theorem 3.7.3 also requires a similar inequality to hold for the higher moments of $\gamma(t_1) - \gamma(t_2) - c_{t_1,t_2}$, possibly involving a different function w'. Two different strategies were used to derive the margin assumption required by Theorem 3.7.3. The $L^\infty - L^2$ norm inequalities of Chapter 4 are more

natural for estimators $\hat{s}_m$ which belong to a finite-dimensional linear model m, as equivalence of norms holds over finite dimensional linear spaces. Penalized empirical risk minimizers which do not belong to a finite-dimensional model, such as kernel methods, require different arguments. A lower bound on the regularization parameter, as used in Chapter 3, is a much more natural and easily verified assumption in this context. The methods developed in Chapters 3 and 4 to relax boundedness assumptions for the hold-out are likely to be applicable in many other settings involving *empirical risk minimization*. This applies both to the general arguments used to prove Theorem 3.7.2 and to the hypotheses used to derive margin assumptions, especially the $L^\infty - L^2$ inequalities of chapter 4.

7.1.1 Norm inequalities

Theorem 4.3.2 states an oracle inequality for the hold-out in sparse linear regression. The assumptions of Theorem 4.3.2 are exactly those required to prove that the candidate estimators $(\hat{s}_k : x \mapsto \hat{q}_k + \langle \hat{\theta}_k, \cdot \rangle)_{1 \leqslant k \leqslant n}$

- Satisfy a weak boundedness assumption of the form $\|\hat{s}_k\|_\infty \leqslant Ln^\alpha$
- Satisfy an $L^\infty - L^2$ norm-inequality, $\|\hat{s}_k - \hat{s}_{k'}\|_\infty \leqslant \kappa(n, n_v) \|\hat{s}_k - \hat{s}_{k'}\|_{L^2(X)}$,

in the setting of Chapter 4. Those are virtually the only properties of the estimators used in the proof. Thus, the $L^\infty - L^2$ norm inequality of Chapter 4 seems to be a general strategy to derive margin assumptions, which applies to any collection of estimators. Moreover, the arguments of Chapter 4 are not specific to the Huber loss function: they use only the Lipschitz-continuity of the Huber loss and the fact that the associated risk $t \mapsto \ell(s,t)$ is *locally strongly convex* in some L^∞ ball around its optimum s. This leads to the following conjecture.

Open Problem 7.1.1 *For a loss function in regression which is Lipschitz continuous and* locally strongly convex *in the sense that for two constants $r, \mu > 0$, and any predictor t,*

$$\|t - s\|_\infty \leqslant r \implies \ell(s,t) \geqslant \mu \|t - s\|^2_{L^2(X)},$$

the hold-out satisfies an oracle inequality under the two assumptions

- $\forall m \in \mathcal{M}, \|\hat{s}_m\|_\infty \leqslant Ln^\alpha$
- $\forall (m, m') \in \mathcal{M}^2 \|\hat{s}_m - \hat{s}_{m'}\|_\infty \leqslant \kappa(n_v, |\mathcal{M}|) \|\hat{s}_m - \hat{s}_{m'}\|_{L^2(X)}$,

where $\kappa(n_v, |\mathcal{M}|)$ may grow as fast as $\sqrt{\frac{n_v}{\log(|\mathcal{M}|)}}$.

In *least-squares regression*, the risk is obviously strongly convex, but the loss-function is no longer Lipschitz continuous, so a direct analogy with Chapter 4 is impossible. We nonetheless conjecture that $L^\infty - L^2$ norm inequalities are sufficient conditions to prove oracle inequalities for the hold-out in this setting.

Open Problem 7.1.2 *Under an $L^\infty - L^2$ norm inequality of the form*

$$\forall (m, m') \in \mathcal{M}^2, \|\hat{s}_m - \hat{s}_{m'}\|_\infty \leqslant \kappa(n_v, |\mathcal{M}|) \|\hat{s}_m - \hat{s}_{m'}\|_{L^2(X)}$$

similar to that of Chapter 4, Theorem 4.3.2, and a weak polynomial upper bound on the estimators, of the form $\|\hat{s}_m\|_\infty \leqslant Ln^\alpha$, the hold-out and Agghoo satisfy an oracle inequality in least-squares regression.

Motivation for this conjecture comes from the following calculation.

$$\begin{aligned}(\gamma(t_1,(x,y)) - \gamma(t_2,(x,y)))^2 &= \left((y - t_1(x))^2 - (y - t_2(x))^2\right)^2 \\ &= (t_1(x) - t_2(x))^2 (2y - t_1(x) - t_2(x))^2 \\ &\leqslant 8(y - s(x))^2 (t_1 - t_2)^2(x) + 2[(t_1 - t_2)^2 (2s - t_1 - t_2)^2](x).\end{aligned}$$

If $\|t_1 - t_2\|_\infty \leqslant \kappa \|t_1 - t_2\|_{L^2(X)}$, and the noise $Y - s(X)$ is such that $\mathbb{E}[(Y - s(X))^2 | X] \leqslant \sigma^2$ a.s., then

$$P\left(\gamma(t_1) - \gamma(t_2)\right)^2 \leqslant 8\sigma^2 \|t_1 - t_2\|^2_{L^2(X)} + 2\kappa^2 \|t_1 - t_2\|^2_{L^2(X)} \left[\|t_1 - s\|_{L^2(X)} + \|t_2 - s\|_{L^2(X)}\right]^2.$$

This shows that hypothesis (7.1) holds for a quadratic function w depending on σ^2 and κ.

Least-squares density estimation is another setting where Theorem 3.7.3 can be used to prove oracle inequalities under an $L^\infty - L^2$ norm-equivalence. In Chapter 6, Theorem 3.7.3 was used to provide a first bound on the risk of the hold-out for Fourier series estimators. It seems likely that more general oracle inequalities are possible in least-squares density estimation. Justification for this claim comes from the fact that, in least-squares density estimation,

$$\gamma(t_1, z) - \gamma(t_2, z) = 2(t_2(z) - t_1(z)) + \|t_1\|^2 - \|t_2\|^2.$$

As $\|t_1\|$ and $\|t_2\|$ are constants, setting $c_{t_1,t_2} = \|t_1\|^2 - \|t_2\|^2$ yields

$$\begin{aligned}P\left(\gamma(t_1) - \gamma(t_2) - c_{t_1,t_2}\right)^2 &= 4 \int (t_1 - t_2)^2 s \leqslant 4 \|s\|_\infty \|t_1 - t_2\|^2 \\ \|\gamma(t_1) - \gamma(t_2) - c_{t_1,t_2}\|_\infty &\leqslant 2 \|t_1 - t_2\|_\infty.\end{aligned}$$

Thus, provided that $\|s\|_\infty < +\infty$, a margin assumption (7.1) is satisfied, and under the norm-inequality $\|t_1 - t_2\|_\infty \leqslant \kappa \|t_1 - t_2\|$, an equation similar to (7.1)

also holds for the higher moments of $\gamma(t_1) - \gamma(t_2) - c_{t_1,t_2}$. Moreover, under similar assumptions, with a constant κ of order $\sqrt{n}$ in the norm inequality, Arlot and Lerasle [5] obtain oracle inequalities for cross-validation with leading constant $C > 1$ and remainder term $r_n = \mathcal{O}\left(\frac{\log|\mathcal{M}|}{n}\right)$. This leads to the following conjecture.

Open Problem 7.1.3 *Under an $L^\infty - L^2$ norm-inequality with constant $\kappa = \kappa(n_v) \leqslant \sqrt{n_v}$ and for uniformly bounded densities ($\|s\|_\infty < +\infty$), the hold-out satisfies an oracle inequality in least-squares density estimation with remainder term $\mathcal{O}\left(\frac{\log|\mathcal{M}|}{n_v}\right)$, where n_v denotes the size of the validation sample and $|\mathcal{M}|$ is the number of estimators in the given collection.*

7.1.2 Penalized empirical risk minimization

For *penalized empirical risk minimization* methods which do not belong to a fixed, finite dimensional model, such as kernel methods, $L^\infty - L^2$ norm inequalities may not hold or may be inadequate to prove an optimal oracle inequality. In chapter 3, an oracle inequality is proved for the hold-out applied to selecting the regularization parameter of kernel methods, when the loss function is Lipschitz and under the assumption of a lower bound on the regularization parameter. The key properties of the kernel penalty $\Omega(t) = \|t\|_{\mathcal{H}}^2$ required for the proof to work are the strong convexity of Ω on some Hilbert space (the RKHS $\mathcal{H}$) and the fact that the underlying norm on $\mathcal{H}$ dominates the supremum norm. These properties also hold for the ridge or elastic net penalties of linear regression under appropriate conditions on the covariate vector X, which leads to the following open problem.

Open Problem 7.1.4 *Under appropriate moment conditions on the covariate vector X and a lower bound on the ridge regularization parameter λ (or the ridge component of the elastic net penalty), prove an oracle inequality for the hold-out applied to penalized empirical risk minimization with a Lipschitz loss and ridge or elastic-net penalty.*

7.1.3 Extension to empirical risk minimization

The study of the hold-out is essentially the study of empirical risk minimization on the (random) set of functions $(\hat{s}_m(D_n^T))_{m\in\mathcal{M}}$, where D_n^T is the *training sample*. Therefore, advances in the theoretical understanding of the hold-out should have corresponding implications for empirical risk minimization. Theorem 3.7.2 of this thesis is essentially a general oracle inequality for empirical risk minimization on a finite collection of functions $(f_m)_{m\in\mathcal{M}}$. Thus, it is reasonable to expect that Theorem 3.7.2 can be adapted for general empirical risk minimization, where

it can help relax boundedness assumptions, as for the hold-out. The assumption that $\mathcal{M}$ is finite, reasonable in the case of the hold-out, must be discarded to study empirical risk minimization, as ERM estimators used in practice typically perform empirical risk minimization on infinite sets (called models), which may be finite-dimensional vector spaces (as in least-squares regression) or finite-dimensional convex sets ("constraint formulation" of the Lasso, for example). To this end, Theorem 3.7.2 can be applied to a *discretization* of an infinite class of functions $\mathcal{F}$ using, for example, L^∞ balls or *brackets* [76, Section 6.1.3]. The bounds will then involve the L^∞ entropy or the "bracketing entropy" of the class $\mathcal{F}$, as is classical in the litterature. Alternatively, the proof of Theorem 3.7.2 can be modified to take into account the "complexity" of the infinite class $\mathcal{F}$. In the proof of Theorem 3.7.2, the probabilistic tools used to control the empirical process are Bernstein's inequality and a union bound over the finite set $\mathcal{M}$. For an infinite class $\mathcal{F}$, control of the empirical process can instead be achieved using bracketing entropy and chaining to control the expectation of suprema (as in lemma 6.5 of [76], for example) and Bousquet's inequality [21] to control their deviations. Carrying out one of these methods is the subject of the following open problem.

Open Problem 7.1.5 *State an equivalent of Theorem 3.7.2 for general empirical risk minimization over a class of functions $\mathcal{F}$, under some standard assumption limiting the "complexity" or "dimension" of the class $\mathcal{F}$ (such as $L^\infty-$entropy, bracketing entropy, or linear dimension when $\mathcal{F}$ is a subset of a vector space).*

7.2 Agghoo

The first main goal of this thesis was to prove oracle inequalities for the hold-out and Agghoo. The second main goal was to understand more precisely the effect of aggregation, to explain how and to which degree Agghoo improves on the hold-out. By aggregating hold-out estimators corresponding to different ways of splitting the data, Agghoo reduces the variability of the hold-out which is due to the arbitrary choice of training subset T. Thus, the greater the influence of T, the greater the benefits of Agghoo relative to the hold-out. In general, there are three mechanisms by which T causes the hold-out $\widehat{s}^{\mathrm{ho}}_T = \hat{s}_{\hat{m}_T}$ to vary.

1. $\hat{s}_m(D_n^T)$ varies with respect to T, especially if $|T|$ is small or the estimators $\hat{s}_m$ are *unstable*. This is a "bagging effect": Agghoo reduces the variance of the $\hat{s}_m$ with respect to the sample.

2. The hold-out parameter $\hat{m}_T$ may be unstable because the empirical measure $P_n^{T^c}$, used to estimate the risk, is unstable: this happens when T^c is small, i.e when $|T| \approx n$.

3. The hold-out parameter $\hat{m}_T$ may be unstable because the true risk $\left\|\hat{s}_m(D_n^T) - s\right\|^2$ is itself unstable: this happens when T is small or the estimators $\hat{s}_m$ are *unstable*.

Chapters 5 and 6 focus mainly on the second of these mechanisms, which should be dominant when $|T| = n_t$ is close to n and the estimators $\hat{s}_m$ are *stable*. Another possibility, not explored in this thesis, is to choose n_t much smaller than n: although this obviously degrades the performance of the individual estimators $\hat{s}_m$, in special cases where the effect of aggregation is strong, it may nonetheless be interesting. We discuss this possibility in section 7.2.2.

7.2.1 Agghoo for stable estimators with $n_t \sim n$

In chapter 5, the asymptotic distribution of the hold-out was shown to be independent of the training sample D_n^T. Moreover, the proof showed that suitable projections $P_x(\hat{s}_m(D_n^{T_j}))$ of the estimators can be approximated by $P_x(\hat{s}_m(D_n^T))$, for a common subset T. This makes it possible to separate the effect of bagging (mechanism 1) from that of *hyperparameter aggregation* (mechanism 2). More generally, the separate study of the three mechanisms listed above seems like a good way to approach the study of Agghoo, and to make it more tractable. In the case of mechanism 2, this leads to the following simplified model. Let $(f_m)_{m\in\mathcal{M}}$ be a finite family of deterministic functions. For a family $(I_j)_{1\leqslant V}$ of subsets of $\{1,\dots,n\}$, let

$$\hat{m}_j = \operatorname*{argmin}_{m\in\mathcal{M}} P_n^{I_j}\gamma(f_m)$$

and let

$$\hat{f} = \frac{1}{V}\sum_{j=1}^{V} f_{\hat{m}_j}. \tag{7.2}$$

In other words, Agghoo is modeled by subagging applied to empirical risk minimization over some set of functions $(f_m)_{m\in\mathcal{M}}$.

The results of Chapter 6 show that Agghoo performs best when the risk is "flat" around the optimal parameter k_*. Of course, the parametrization $m \mapsto \hat{s}_m$ may be arbitrary and it is unreasonable to expect a direct generalization of this observation. As Agghoo aggregates estimators rather than hyperparameters, the correct generalization of "flatness" should refer to some intrinsic geometry on the space of estimators $\hat{s}_m$, rather than the set of hyperparameters $\mathcal{M}$. The simplified model of equation (7.2) provides some indications. Assume first that the excess risk $\ell(s,t)$ is quadratic, i.e $\ell(s,t) = \|t-s\|^2$ for some Hilbert space norm $\|\cdot\|$. This is the case in least-squares density estimation and least-squares regression. Moreover, it seems reasonable to expect many risk functions to be approximately

quadratic in an L^∞ neighbourhood of their optimum s, as taking expectations smoothes the convex loss function γ. Assume also to simplify that the $(I_j)_{1\leqslant j\leqslant V}$ are *disjoint* (as for V-fold Agghoo). Then $\hat{f}$ is an average of i.i.d functions, hence its excess risk is reduced by an amount proportional to $\mathbb{E}[\|f_{\hat{m}_1} - f_{\hat{m}_2}\|^2]$ compared to the excess risk of $f_{\hat{m}_1}$ (the hold-out). Oracle inequalities for the hold-out suggest that $\hat{m}_1$ is likely to take values which are close to optimal in the sense that $\|f_{\hat{m}_1} - s\|^2 \approx \min_{m\in\mathcal{M}} \|f_m - s\|^2$. Thus, Agghoo is likely to perform well when there are functions f_{m_1}, f_{m_2} that are far apart in norm $\|\cdot\|$ and have nearly optimal risk, $\|f_{m_1} - s\|^2 \approx \|f_{m_2} - s\|^2 \approx \|f_{m_*} - s\|^2$, a hypothesis which we call a "flatness assumption".

A major difficulty in making this heuristic precise is that it requires knowledge of just how far the risk of the hold-out, $\|f_{\hat{m}_1} - s\|^2$, is from the optimum $\min_{m\in\mathcal{M}} \|f_{\hat{m}} - s\|^2$: in other words, not just an oracle inequality (an upper bound on $\|f_{\hat{m}_1} - s\|^2$), but also an "inverse oracle inequality", i.e a lower bound on $\|f_{\hat{m}_1} - s\|^2 - \inf_{m\in\mathcal{M}} \|f_m - s\|^2$.

Towards a rigorous "flatness assumption" in least-squares density estimation In Chapter 5, an "oracle equality" with exact remainder term was established for the hold-out using an asymptotic approximation of the hold-out risk estimator. This argument is specific to least-squares density estimation with orthogonal series estimators (the setting of Chapter 5). More generally, in *least-squares density estimation*, since

$$P_n\gamma(f_m) - P_n\gamma(f_{m_*}) = \|f_m - s\|^2 - \|f_{m_*} - s\|^2 - 2(P_n - P)(f_m - f_{m_*}),$$

a reasonable heuristic is that the hold-out selects f_m with non-negligible probability whenever the variance of the empirical process $(P_n - P)(f_m - f_{m_*})$ outweighs the excess risk $\|f_m - s\|^2 - \|f_{m_*} - s\|^2$. This heuristic was stated and applied successfully in Chapter 5. It implies that, just as margin assumptions provide upper bounds on the excess risk of the hold-out $\|f_{\hat{m}_1} - s\|^2$, a reasonable way to prove a lower bound for $\|f_{\hat{m}_1} - s\|^2$ (an "inverse oracle inequality") is to assume that $\mathrm{Var}\left[(f_m - f_{m_*})(X)\right]$ is larger than some function of $\|f_m - s\|^2 - \|f_{m_*} - s\|^2$, a kind of "inverse margin hypothesis". More generally, one may hope to find interesting risk bounds for Agghoo under assumptions that only involve the quantities $\mathrm{Var}\left((f_m - f_{m'})(X)\right), \|f_m - f_{m'}\|$ and $\|f_m - s\|$, which have a geometrical interpretation. This leads to the following open problem.

Open Problem 7.2.1 *In least-squares density estimation with bounded density s, find (necessary and) sufficient conditions on $\mathrm{Var}\left((f_m - f_{m'})(X)\right)$, $\|f_m - f_{m'}\|$, $\|f_m - s\|$ and n so that*

$$\left\|\hat{f} - s\right\|^2 < \inf_{m\in\mathcal{M}} \|f_m - s\|^2$$

and so that

$$\left\| \hat{f} - s \right\|^2 \leqslant \theta \inf_{m \in \mathcal{M}} \|f_m - s\|^2$$

for some $\theta \in]0;1]$.

An argument in favour of this conjecture is that the empirical process $\sqrt{n_v}(P_{n_v} - P)(f_m)$ may often be approximated by a gaussian process G with the same variance-covariance function, as was done in Chapter 5 using the theory of *strong approximation*. The distribution of G only depends on its variance-covariance function, that is to say, on $\mathrm{Cov}(f_m(X), f_{m'}(X))$ for $(m, m') \in \mathcal{M}^2$, which is the "scalar product" associated with the (pseudo-)norm $\sqrt{\mathrm{Var}(t(X))}$.

As problem 7.2.1 may be hard to solve in general, it is worthwhile to consider some special cases where interesting results about the hold-out may be proved by a more direct approach.

Smoothly parametrized estimators If a collection of estimators $\hat{s}_\lambda$ is twice differentiable with respect to its parameter λ, then it is possible to carry out a Taylor expansion of the risk and of its hold-out estimator in the vicinity of $\lambda_* = \mathrm{argmin}_\lambda P\gamma(\hat{s}_\lambda)$. This makes it possible to find asymptotic approximations for the hold-out parameter $\hat{\lambda}$, its risk, and that of aggregated hold-out, provided that the estimators are stable, such that Agghoo may be approximated by bagged empirical risk minimization, as in equation (7.2)). Thus, an approximate expression for the hold-out parameter $\hat{\lambda}$ can be obtained. A simple geometric argument suggests that local aggregation around λ_* outperforms the oracle if and only if the path $\lambda \mapsto \hat{s}_\lambda$ "curves around" the target s (as opposed to "curving away" from s), or in other words, iff $\langle s - \hat{s}_{\lambda_*}, \partial_\lambda^2 \hat{s}_\lambda|_{\lambda_*} \rangle > 0$ (assuming that $\partial_\lambda^2 \hat{s}_\lambda|_{\lambda_*} \neq 0$). This leads to the following conjecture.

Open Problem 7.2.2 *Prove that if* $\hat{s}_\lambda$ *is a collection of estimators which is twice differentiable with respect to the hyperparameter* λ*, the estimators are* sufficiently stable *(in a sense to be made precise) and*

$$\langle s - \hat{s}_{\lambda_*}, \partial_\lambda^2 \hat{s}_\lambda|_{\lambda_*} \rangle > 0,$$

then Agghoo can outperform the oracle.

Regular histograms Regular histograms $\hat{s}_k$ on $[0;1]$ with k pieces are *unstable* with respect to the parameter k: since the edges of the intervals are shifted with respect to each other, the squared distance $\|\hat{s}_k - \hat{s}_{k'}\|^2$ between any two histograms may be of the same order as the risk $\|\hat{s}_k - s\|^2$ for parameters k, k' close to the optimal one. This suggests that aggregated hold-out may outperform the oracle by as much as a constant factor, in the limit $n_t \sim n$. Results of Genuer [46]

for regressograms support this claim: he shows that when the target regression function s is Lipschitz, aggregating random regressograms asymptotically reduces their variance by at least a factor $\frac{3}{4}$, while the bias term does not increase. This leads to the following conjecture

Open Problem 7.2.3 *For a continuously differentiable pdf s, Monte-Carlo Agghoo applied to the regular histograms $\hat{s}_k$ asymptotically satisfies an oracle inequality with leading constant $C < 1$.*

Analysis of Agghoo is facilitated by the fact that the process $\sqrt{n_v}(P_{n_v} - P)(\hat{s}_k)$ can be naturally expressed in terms of the "empirical brownian bridge" $\sqrt{n_v}(F_{n_v} - F)$ (where F_{n_v} denotes the empirical distribution function), so methods of *strong approximation* can be used to construct a Gaussian process which approximates $\sqrt{n_v}(P_{n_v} - P)(\hat{s}_k)$. Moreover, the variance-covariance function of this process can be explicitly computed in terms of the density s.

7.2.2 Aggregated hold-out with small training samples

The case of small training sample sizes n_t was not explored in this thesis. When $n_t << n$, the validation samples are similar (they include almost the full dataset) but the training samples may be very different. As a result, the estimators $\hat{s}_m(D_n^{T_j})$ cannot be treated like fixed, deterministic functions, in contrast to the previous section, and bagging becomes a significant factor. This does not necessarily mean that the effect of hyperparameter aggregation becomes negligible, since the hold-out risk estimators, $P_n^{T_j^c}\gamma(\hat{s}_m(D_n^{T_j}))$, inherit the variability of the estimators $\hat{s}_m(D_n^{T_j})$. As a result, the selected parameters m_j are also likely to be variable.

Choosing n_t much smaller than n is not recommanded in general, as it is likely to significantly degrade the performance of the estimators $\hat{s}_m(D_n^{T_j})$. The simulations conducted during this thesis mostly support choosing $n_t = 0.8n$ or $n_t = 0.9n$ for sample sizes $n = 500$ or $n = 1000$, which suggests that in the asymptotic, one should have $n_t \sim n$. However, there are special cases where it is reasonable to choose $n_t << n$. We discuss two examples below.

k-nearest neighbours Our simulations in classification (Figure 3.3, Chapter 3) show that the risk of Agghoo can be relatively flat as a function of $\frac{n_t}{n}$ when applied to k-nearest neighbours in classification. Moreover, in regression, the results of Biau and Guyader [14] show that subagging the (inconsistent) $1-$nearest neighbour rule using subsamples of size $n_t << n$ can lead to estimators that converge at the optimal rate. This justifies using Agghoo with $n_t << n$ for nearest neighbour rules. Assuming that the parameter $\hat{k}$ chosen by the hold-out is equivalent to the optimum deterministic choice, $k_*(n_t) = \operatorname{argmin}_k \mathbb{E}[\ell(s, \hat{s}_k(D_n^T))]$, which is plausible

when $n_t \to +\infty$, Agghoo should behave similarly to the (su)bagged $k_*(n_t)-$nearest neighbour rule.

Open Problem 7.2.4 *For the collection of $k-$nearest neighbour estimators in regression, show that Agghoo can be approximated by a weighted nearest neighbour rule, as in [13, 14], when $\frac{n_t}{n} \to 0$ at the right rate. Show that Agghoo converges at the optimal rate for a suitable choice of $n_t(n)$. Find the asymptotically optimal sequence $n_t(n)$. Can Agghoo perform better than the oracle when $\frac{n_t}{n} \to 0$?*

Regular histograms and regressograms Results of Arlot and Genuer for regressograms [4] show that aggregating piecewise constant estimators on "shifted" regular partitions results in improved convergence rates for a sufficiently smooth target s, in least-squares regression. One can reasonably expect analogous results to hold true for histograms in density estimation. The "shifted" regular partitions of [4] consist of intervals of the form $[u + \frac{j}{k}; u + \frac{j+1}{k}]$, where u is a random element of $[0; \frac{1}{k}]$ and k is fixed. For nearby, distinct values of k, regular intervals $[\frac{j}{k}; \frac{j+1}{k}]$ are shifted relative to each other in a similar way as by the construction of Arlot and Genuer. Let $\hat{s}_k$ denote the regular histogram on $[0; 1]$ with k pieces. Let $k_*(n_t) = \operatorname{argmin}_{1 \leqslant k \leqslant n} \mathbb{E}\left[\left\|\hat{s}_k(D_n^T) - s\right\|^2\right]$. Agghoo, which aggregates the regular histograms $\hat{s}_k$ over different values of k close to the optimal $k_*(n_t)$, may therefore be conjectured to behave similarly to the aggregate defined in [4], for $k = k_*(n_t)$. As this aggregate converges *faster* than the individual estimators, the optimal value of k for the aggregate is much smaller than for the regular histogram or regressogram. For Agghoo, this justifies taking $n_t << n$ such that $k_*(n_t)$ is sufficiently small.

Open Problem 7.2.5 *If $s \in L^2([0; 1])$ is sufficiently smooth, show that Agghoo converges faster than the regular histograms, for a properly chosen sequence $n_t(n)$ such that $\frac{n_t(n)}{n} \to 0$.*

Remark that in this scenario, hyperparameter aggregation plays a crucial role, even though $n_t << n$.

7.3 Local aggregation

Chapter 6 shows that Agghoo performs *local aggregation* of the Fourier series estimators $\hat{s}_k$ in some neighbourhood of the optimal parameter,

$$k_*(n_t) = \operatorname*{argmin}_k \mathbb{E}\left[\left\|\hat{s}_k(D_n^T) - s\right\|^2\right].$$

Theorem 6.4.4 shows that the size Δ of this neighbourhood *adapts* to some degree to the density s. For example, if the squared Fourier coefficients θ_j^2 of s decrease

very slowly around $k_*(n_t)$, i.e if the risk is very "flat" around $k_*(n_t)$, the size of Δ will increase, which will reduce Agghoo's risk (at least if $n - n_t$ is large enough).

However, the strongest results, in which Agghoo outperforms the oracle by a constant factor, require Agghoo's parameter n_t to be chosen in a distribution-dependent way. This is not just a limitation of the arguments of Chapter 6: it is not hard to see that to gain a constant factor improvement in the oracle inequality in the setting of Corollary 6.4.6, it is necessary to aggregate two estimators $\hat{s}_k, \hat{s}_{k'}$ that are sufficiently far apart, which here means that $|k-k'|$ must be of order $k_*(n_t)$. This requires the hold-out to be sufficiently variable, which in turn requires n_t to be sufficiently close to n (but not too much). Thus, Agghoo does not fully adapt the size of the neighbourhood in which it aggregates the $\hat{s}_k$ to its (distribution dependent) optimal value.

This is somewhat unsatisfactory, and it would be interesting to find methods which solve this problem. Optimal aggregation is clearly too ambitious a goal in general: results of Tsybakov [102] show that oracle inequalities for optimal aggregation have a remainder term of order $\sqrt{\frac{\log|\mathcal{M}|}{n_v}}$, which is larger than the risk of the oracle when s is sufficiently smooth. Instead, one would like an estimator $\hat{s}$ that functions like an idealized version of Agghoo, i.e an estimator such that

- $\hat{s}$ satisfies a general oracle inequality, similar to the hold-out and Agghoo
- $\hat{s}$ is given by a black box method, which uses no special features of the $\hat{s}_m$.
- $\hat{s}$ makes the best possible use of *local aggregation* around the oracle, and in particular takes advantage of any *flatness condition* when it holds.

To make this last requirement precise, consider a collection of estimators $(\hat{s}_m)_{m\in\mathcal{M}}$ and for any $\varepsilon > 0$, define the sets

$$\mathcal{M}_\varepsilon = \left\{ m \in \mathcal{M} : \|\hat{s}_m - s\|^2 \leqslant \inf_{m'\in\mathcal{M}} \|\hat{s}_{m'} - s\|^2 + \varepsilon \right\}$$

and the aggregated estimators

$$\hat{s}_\varepsilon^{ag} = \frac{1}{|\mathcal{M}_\varepsilon|} \sum_{m\in\mathcal{M}_\varepsilon} \hat{s}_m.$$

$\mathcal{M}_\varepsilon$ is a set of estimators "close" to the oracle in a sense relevant to risk minimization, while $\hat{s}_\varepsilon^{ag}$ is the "local aggregate" around the oracle at the scale ε. The aim is to construct an estimator $\hat{s}$ which performs as well as the best local aggregate $\hat{s}_\varepsilon^{ag}$.

Open Problem 7.3.1 *In the setting of least-squares density estimation, construct an aggregation estimator $\hat{s}$ such that*

$$\frac{\|\hat{s} - s\|^2}{\inf_{\varepsilon>0} \|\hat{s}_\varepsilon^{ag} - s\|^2} \to 1,$$

given any sufficiently stable *collection of estimators $(\hat{s}_m)_{m\in\mathcal{M}}$ (for example, empirical risk minimizers on linear models).*

A possible solution to this problem is the following. Given a cross-validation estimator $\mathrm{CV}_\mathcal{T}$, estimate $\mathcal{M}_\varepsilon$ by

$$\hat{\mathcal{M}}_\varepsilon = \left\{ m \in \mathcal{M} : \mathrm{CV}_\mathcal{T}(\hat{s}_m) \leqslant \inf_{m'\in\mathcal{M}} \mathrm{CV}_\mathcal{T}(\hat{s}_{m'}) + \varepsilon \right\},$$

and define the corresponding "local aggregates",

$$\hat{s}_{\mathcal{T},\varepsilon}^{ag} = \frac{1}{|\hat{\mathcal{M}}_\varepsilon|} \sum_{m\in\hat{\mathcal{M}}_\varepsilon} \hat{s}_m.$$

The risk of the aggregate $\hat{s}_{\mathcal{T},\varepsilon}^{ag}$ may be estimated by

$$\hat{C}(\varepsilon) = -\frac{1}{2|\hat{\mathcal{M}}_\varepsilon|^2} \sum_{(m,m')\in\hat{\mathcal{M}}_\varepsilon^2} \|\hat{s}_m - \hat{s}_{m'}\|^2 + \frac{1}{|\hat{\mathcal{M}}_\varepsilon|} \sum_{m\in\hat{\mathcal{M}}_\varepsilon} \mathrm{CV}_\mathcal{T}(\hat{s}_m).$$

Finally, choosing $\hat{\varepsilon} = \mathrm{argmin}_{\varepsilon>0} \hat{C}(\varepsilon)$ yields the estimator $\hat{s} = \hat{s}_{\mathcal{T},\hat{\varepsilon}}^{ag}$.

Open Problem 7.3.2 *Does $\hat{s}$ solve problem 7.3.1?*

Bibliography

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] Sylvain Arlot. V-fold cross-validation improved: V-fold penalization. 40 pages, plus a separate technical appendix., February 2008.

[3] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statist. Surv.*, 4:40–79, 2010.

[4] Sylvain Arlot and Robin Genuer. Analysis of purely random forests bias. *ArXiv*, abs/1407.3939, 2014.

[5] Sylvain Arlot and Matthieu Lerasle. Choice of V for V-fold cross-validation in least-squares density estimation. *Journal of Machine Learning Research (JMLR)*, 17(208):1–50, 2016.

[6] Jean-Yves Audibert. *PAC-Bayesian aggregation and multi-armed bandits*. Habilitation à diriger des recherches, Université Paris-Est, October 2010.

[7] Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794, 10 2011.

[8] Francis Bach. Bolasso: Model consistent lasso estimation through the bootstrap. *Proceedings of the 25th international conference on Machine learning*, 33-40, 05 2008.

[9] Viorel Barbu and Teodor Precupanu. *Convexity and optimization in Banach spaces*. Springer, 2012.

[10] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[11] Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*. Springer Series in the Data Sciences. Springer, 2015.

[12] Gérard Biau and Erwan Scornet. A random forest guided tour. *TEST*, 25(2):197–227, 2016.

[13] Gérard Biau and Luc 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.

[14] Gérard Biau and Arnaud Guyader. On the rate of convergence of the bagged nearest neighbor estimate. *The Journal of Machine Learning Research*, 11:687–712, 03 2010.

[15] Peter Bickel, Ya'acov Ritov, and Alexandre Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37, 02 2008.

[16] Lucien Birgé and Pascal Massart. *From Model Selection to Adaptive Estimation*, pages 55–87. Springer New York, New York, NY, 1997.

[17] Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 09 1998.

[18] Gilles Blanchard and Pascal Massart. Discussion: Local rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2664–2671, 12 2006.

[19] Kevin Bleakley and Jean-Philippe Vert. The group fused lasso for multiple change-point detection. *Arxiv preprint arXiv:1106.4199*, 06 2011.

[20] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: PS*, 9:323–375, 2005.

[21] Olivier Bousquet. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495 – 500, 2002.

[22] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[23] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[24] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

[25] Peter Bühlmann and Bin Yu. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002.

[26] Andreas Buja and Werner Stuetzle. Observations on bagging. *Statistica Sinica*, 16(2):323–351, 2006.

[27] Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference*. Springer-Verlag, New York, second edition, 2002. A practical information-theoretic approach.

[28] Olivier Catoni. Universal aggregation rules with exact bias bound. 1999.

[29] Olivier Catoni. *Statistical Learning Theory and Stochastic Optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer, Berlin, 2001. Lectures from the 31st Summer School on Probability Theory held in Saint-Flour, July 8 – 25, 2001, with a foreword by Jean Picard.

[30] Alain Celisse. Optimal cross-validation in density estimation with the l^2-loss. *The Annals of Statistics*, 42(5):1879–1910, Oct 2014.

[31] Sourav Chatterjee and Jafar Jafarov. Prediction error of cross-validated Lasso. *arXiv e-prints*, page arXiv:1502.06291, February 2015.

[32] X. Chen, Z. J. Wang, and M. J. McKeown. Asymptotic analysis of robust lassos in the presence of noise with large variance. *IEEE Transactions on Information Theory*, 56(10):5131–5149, Oct 2010.

[33] Denis Chetverikov, Zhipeng Liao, and Victor Chernozhukov. On cross-validated Lasso. *arXiv e-prints*, page arXiv:1605.02214, May 2016.

[34] Geoffrey Chinot, Guillaume Lecué, and Matthieu Lerasle. Robust statistical learning with lipschitz and convex loss functions. *Probability Theory and Related Fields*, 176(3):897–940, Apr 2020.

[35] Clementine Dalelane. Exact oracle inequality for a sharp adaptive kernel density estimator. working paper or preprint, April 2005.

[36] Pascaline Descloux and Sylvain Sardy. Model selection with lasso-zero: adding straw to the haystack to better find needles. *arXiv e-prints*, page arXiv:1805.05133, May 2018.

[37] L. Devroye and T. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979.

[38] Luc P. Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

[39] Thomas G. Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[40] Mona Eberts and Ingo Steinwart. Optimal regression rates for svms using gaussian kernels. *Electron. J. Statist.*, 7:1–42, 2013.

[41] Sam Efromovich. *Nonparametric Curve Estimation*. Springer New York, 1999.

[42] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 04 2004.

[43] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1, part 2):119–139, 1997. EuroCOLT '95.

[44] Jerome H. Friedman and Peter Hall. On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*, 137(3):669 – 683, 2007. Special Issue on Nonparametric Statistics and Related Topics: In honor of M.L. Puri.

[45] Seymour Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328, 1975.

[46] Robin Genuer. Risk bounds for purely uniformly random forests. Research Report RR-7318, INRIA, June 2010.

[47] Eitan Greenshtein and Ya'Acov Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 12 2004.

[48] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer New York, 2002.

[49] Peter Hall. On trigonometric series estimates of densities. *Ann. Statist.*, 9(3):683–685, 05 1981.

[50] Peter Hall. Cross-validation and the smoothing of orthogonal series density estimators. *Journal of Multivariate Analysis*, 21(2):189 – 206, 1987.

[51] Peter Hall and Andrew Robinson. Reducing variability of crossvalidation for smoothing-parameter choice. *Biometrika*, 96:175–186, 01 2009.

[52] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Basis Expansions and Regularization*, pages 139–189. Springer New York, New York, NY, 2009.

[53] Darren Homrighausen and Daniel McDonald. Risk consistency of cross-validation with lasso-type procedures. *Statistica Sinica*, 27, 08 2013.

[54] Andres Hoyos-Idrobo, Yannick Schwartz, Gael Varoquaux, and Bertrand Thirion. Improving sparse recovery on structured images with bagged clustering. In *2015 International Workshop on Pattern Recognition in NeuroImaging*. IEEE, June 2015.

[55] Peter J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964.

[56] P.J. Huber and E. Ronchetti. *Robust Statistics*, pages 1248–1251. Springer, New York, 2011.

[57] A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *Ann. Statist.*, 36(5):2183–2206, 10 2008.

[58] Yoonsuh Jung. Efficient tuning parameter selection by cross-validated score in high dimensional models. *International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering*, 10(1):19–25, 2016.

[59] Yoonsuh Jung and Jianhua Hu. A K-fold averaging cross-validation procedure. *Journal of Nonparametric Statistics*, 27(2):167–179, 2015.

[60] M. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Comput*, 11(6):1427–1453, Aug 1999.

[61] Michael Kearns, Yishay Mansour, Andrew Y. Ng, and Dana Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27(1):7–50, Apr 1997.

[62] J. Komlós, P. Major, and G. Tusnády. An approximation of partial sums of independent rv'-s, and the sample df. i. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 32(1):111–131, Mar 1975.

[63] J. Komlós, P. Major, and G. Tusnády. An approximation of partial sums of independent rv's, and the sample df. ii. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 34(1):33–58, Mar 1976.

[64] Sophie Lambert-Lacroix and Laurent Zwald. Robust regression through the huber's criterion and adaptive lasso penalty. *Electron. J. Statist.*, 5:1015–1053, 2011.

[65] Marc Lavielle. Using penalized contrasts for the change-point problem. *Signal Processing*, 85(8):1501 – 1510, 2005.

[66] Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means : theory and practice. *arXiv e-prints*, page arXiv:1711.10306, November 2017.

[67] Guillaume Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13(4):1000–1022, 11 2007.

[68] Guillaume Lecué and Matthieu Lerasle. Learning from mom's principles: Le cam's approach. *Stochastic Processes and their Applications*, 129(11):4385 – 4410, 2019.

[69] Guillaume Lecué and Charles Mitchell. Oracle inequalities for cross-validation type procedures. *Electron. J. Statist.*, 6:1803–1837, 2012.

[70] Matthieu Lerasle. Optimal model selection for stationary data under various mixing conditions. *Annals of Statistics - ANN STATIST*, 39, 11 2009.

[71] Nelo Molter Magalhães. *Cross-validation and penalization for density estimation*. Theses, Université Paris Sud - Paris XI, May 2015.

[72] Guillaume Maillard. Aggregated hold out for sparse linear regression with a robust loss function. working paper or preprint, February 2020.

[73] Guillaume Maillard, Sylvain Arlot, and Matthieu Lerasle. Aggregated Hold-Out. working paper or preprint, September 2019.

[74] C. L. Mallows. Some comments on cp. *Technometrics*, 15(4):661–675, 1973.

[75] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

[76] Pascal Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[77] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 10 2006.

[78] Charles A. McCarthy. Cp. *Israel Journal of Mathematics*, 5(4):249–271, 1967.

[79] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society Series B*, 72:417–473, 09 2010.

[80] Ron Meir and Tong Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

[81] Shahar Mendelson. Learning without concentration. *J. ACM*, 62:21:1–21:25, 2014.

[82] Léo Miolane and Andrea Montanari. The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *arXiv e-prints*, page arXiv:1811.01212, Nov 2018.

[83] Fabien Navarro and Adrien Saumard. Slope heuristics and v-fold model selection in heteroscedastic regression using strongly localized bases. *ESAIM: Probability and Statistics*, 21:412–451, 2017.

[84] Arkadi Nemirovski. *Topics in Non-parametric Statistics*, volume 1738 of *Lecture Notes in Math.* Springer, Berlin, 2000.

[85] I. Olkin and F. Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257 – 263, 1982.

[86] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, and et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

[87] Maya L. Petersen, Annette M. Molinaro, Sandra E. Sinisi, and Mark J. [van der Laan]. Cross-validated bagged learning. *Journal of Multivariate Analysis*, 98(9):1693 – 1704, 2007.

[88] Leandro P. R. Pimentel. On the location of the maximum of a continuous stochastic process. *J. Appl. Probab.*, 51(1):152–161, 03 2014.

[89] Daniel Revuz and Marc Yor. *Continuous Martingales and Brownian Motion.* Springer Berlin Heidelberg, 1999.

[90] Philippe Rigollet and Alexandre Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39, 03 2010.

[91] Saharon Rosset and Ji Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, 35(3):1012–1030, 2007.

[92] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259 – 268, 1992.

[93] Joseph Salmon and Arnak S. Dalalyan. Optimal aggregation of affine estimators. In *COLT - 24th Conference on Learning Theory - 2011*, Budapest, Hungary, Jul 2011.

[94] Adrien Saumard. Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression. *Electron. J. Statist.*, 6:579–655, 2012.

[95] Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[96] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.

[97] Ingo Steinwart and Andreas Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, 02 2011.

[98] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974.

[99] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[100] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, pages 91–108, 2005.

[101] Ryan J. Tibshirani and Jonathan Taylor. Degrees of freedom in lasso problems. *Ann. Statist.*, 40(2):1198–1232, 04 2012.

[102] Alexandre B. Tsybakov. Optimal rates of aggregation. In Bernhard Schölkopf and Manfred K. Warmuth, editors, *Learning Theory and Kernel Machines*, pages 303–313, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[103] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[104] Sara van de Geer and Johannes Lederer. *The Lasso, correlated design, and improved oracle inequalities*, volume Volume 9 of *Collections*, pages 303–316. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2013.

[105] Aad W. van der Vaart, Sandrine Dudoit, and Mark J. van der Laan. Oracle inequalities for multi-fold cross validation. *Statist. Decisions*, 24(3):351–371, 2006.

[106] V. N. Vapnik. An overview of statistical learning theory. *Transactions on Neural Networks*, 10(5):988–999, sep 1999.

[107] Gaël Varoquaux, Pradeep Reddy Raamana, Denis A. Engemann, Andrés Hoyos-Idrobo, Yannick Schwartz, and Bertrand Thirion. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, 145:166–179, January 2017.

[108] Grace Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, January 1990.

[109] G. Walter and J. Blum. Probability density estimation using delta sequences. *Ann. Statist.*, 7(2):328–340, 03 1979.

[110] Hansheng Wang and Chenlei Leng. Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, 102(479):1039–1048, 2007.

[111] Hansheng Wang, Guodong Li, and Guohua Jiang. Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business and Economic Statistics*, 25(3):347–355, 2007.

[112] Sijian Wang, Bin Nan, Saharon Rosset, and Ji Zhu. Random lasso. *Ann. Appl. Stat.*, 5(1):468–485, 03 2011.

[113] Zhan Wang, Sandra Paterlini, Fuchang Gao, and Yuhong Yang. Adaptive minimax regression estimation over sparse ℓ_q-hulls. *Journal of Machine Learning Research*, 15:1675–1711, 2014.

[114] Marten Wegkamp. Model selection in nonparametric regression. *Ann. Statist.*, 31(1):252–273, 02 2003.

[115] Huan Xu, Constantine Caramanis, and Shie Mannor. Sparse algorithms are not stable: A no-free-lunch theorem. *IEEE transactions on pattern analysis and machine intelligence*, 34, 08 2011.

[116] Yuhong Yang. Combining different procedures for adaptive regression. *Journal of Multivariate Analysis*, 74:135–161, 07 2000.

[117] Yuhong Yang. Mixing strategies for density estimation. *Ann. Statist.*, 28(1):75–87, 02 2000.

[118] Yuhong Yang. Adaptive regression by mixing. *Journal of the American Statistical Association*, 96(454):574–588, 2001.

[119] Congrui Yi and Jian Huang. Semismooth newton coordinate descent algorithm for elastic-net penalized huber loss regression and quantile regression. *Journal of Computational and Graphical Statistics*, 26(3):547–557, 2017.

[120] Hui Zou, Trevor Hastie, and Robert Tibshirani. On the "degrees of freedom" of the lasso. *Annals of Statistics*, 35(5):2173–2192, 2007.

université PARIS-SACLAY

ÉCOLE DOCTORALE
de mathématiques
Hadamard (EDMH)

Titre: Hold-out et Agrégation d'hold-out

Mots clés: Validation Croisée, Sélection de modèles, Sélection d'hyperparamètres, Agrégation

Résumé: En statistiques, il est fréquent d'avoir à choisir entre plusieurs estimateurs (sélection d'estimateurs) ou à les combiner (agrégation). Cela permet notamment d'adapter la complexité d'un modèle statistique en fonction des données (compromis biais-variance). Pour les problèmes de minimisation de risque, une méthode simple et générale, la validation ou hold-out, consiste à consacrer une partie de l'échantillon à l'estimation du risque des estimateurs, dans le but de choisir celui de risque minimal. Cette procédure nécessite de choisir arbitrairement un sous-échantillon "de validation". Afin de réduire l'influence de ce choix, il est possible d'agréger plusieurs estimateurs hold-out en les moyennant (Agrégation d'hold-out). Dans cette thèse, le hold-out et l'agrégation d'hold-out sont étudiés dans différents cadres. Dans un premier temps, les garanties théoriques sur le hold-out sont étendues à des cas où le risque n'est pas borné: les méthodes à noyaux et la régression linéaire parcimonieuse. Dans un deuxième temps, une étude précise du risque de ces méthodes est menée dans un cadre particulier: l'estimation de densité L^2 par des séries de Fourier. Il est démontré que l'agrégation de hold-out peut faire mieux que le meilleur des estimateurs qu'elle agrège, ce qui est impossible pour une méthode qui, comme le hold-out ou la validation croisée, sélectionne un seul estimateur.

Title: Hold-out and Aggregated hold-out

Keywords: Cross-Validation, Model selection, Hyperparameter selection, Aggregation

Abstract: In statistics, it is often necessary to choose between different estimators (estimator selection) or to combine them (aggregation). For risk-minimization problems, a simple method, called hold-out or validation, is to leave out some of the data, using it to estimate the risk of the estimators, in order to select the estimator with minimal risk. This method requires the statistician to arbitrarily select a subset of the data to form the "validation sample". The influence of this choice can be reduced by averaging several hold-out estimators (Aggregated hold-out, Agghoo). In this thesis, the hold-out and Agghoo are studied in various settings. First, theoretical guarantees for the hold-out (and Agghoo) are extended to two settings where the risk is unbounded: kernel methods and sparse linear regression. Secondly, a comprehensive analysis of the risk of both methods is carried out in a particular case: least-squares density estimation using Fourier series. It is proved that aggregated hold-out can perform better than the best estimator in the given collection, something that is clearly impossible for a procedure, such as hold-out or cross-validation, which selects only one estimator.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France