NNT : 2015SACLS216

Université Paris-Saclay
École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : Université Paris-Sud

Laboratoire d'accueil : GREGHEC (UMR 2959 CNRS)

THÈSE DE DOCTORAT DE MATHÉMATIQUES

Spécialité : Mathématiques appliquées

Emilien JOLY

Estimation robuste de distributions à queue lourde

Robust estimation of heavy-tailed distributions

Date de soutenance : 14 Décembre 2015

Après avis des rapporteurs :	Roberto Imbuzeiro OLIVEIRA	IMPA à Rio de Janeiro
	Jean-Michel LOUBES	Université Paul Sabatier

Jury de soutenance :

Olivier CATONI	CNRS & ENSAE	Examinateur
Antoine CHAMBAZ	Université Paris-Ouest	Examinateur
Jean-Michel LOUBES	Université Paul Sabatier	Rapporteur
Gábor LUGOSI	Universitat Pompeu Fabra	Directeur de thèse
Pascal MASSART	Université Paris-Sud	Président du jury
Gilles STOLTZ	CNRS & HEC Paris	Co-directeur de thèse

Thèse préparée sous la direction
de Gábor Lugosi et Gilles Stoltz
au Département de Mathématiques d'Orsay
et au Laboratoire GREGHEC (UMR 2959 CNRS)

Remerciements

Déjà quatre années se sont écoulées et ce qu'il en reste ne pèse pas plus d'une petite centaine de grammes de papier, d'encre et de colle. Et pourtant, il est difficile de réduire cette belle aventure à un simple manuscrit. Je vais essayer, dans ces quelques lignes de donner un aperçu élargi de ce que peut être l'aventure-thèse et aussi et surtout de remercier tous ceux qui y ont participé par leur aide mathématique ou leur soutien.

"Pourquoi faire une thèse de maths?" est une question que l'on m'a souvent posée. Je crois bien que je n'ai jamais donné exactement la même réponse. Je me souviens avoir répondu que les maths sont ce qui fait fonctionner nos écrans, nos téléphones portables et toute la technologie autour de nous. J'ai même répondu, de temps à autre, qu'elles sont ce qui gouverne d'autres disciplines scientifiques et moins scientifiques. Mais récemment, j'ai trouvé la réponse parfaite qui n'est malheureusement pas de moi. "La motivation des mathématiques n'est pas d'être utile au quotidien mais simplement de mieux comprendre le monde." Etienne Ghys

Mes premiers remerciements vont tout naturellement à Gábor. Tu m'auras fait entrevoir ce qu'est la richesse de la recherche scientifique et la simplicité humaine des grands esprits. J'ai toujours trouvé que la manière dont tu envisages la recherche est aussi celle en laquelle je crois. Je te garderai toujours en modèle dans un coin de ma tête. Merci à Gilles, qui m'aura sauvé (et ce n'est pas faible de le dire) maintes fois en me préservant des pièges que je me suis tendus à moi-même. Tu as été le cadre que beaucoup d'autres thésards m'ont jalousé. Je te dois beaucoup.

Merci Pascal. Évidemment, cette thèse te doit aussi beaucoup. C'est sur tes conseils précieux qui m'ont aiguillés vers Gábor puis Gilles que j'ai trouvé un cadre de recherche privilégié. Je me souviendrai de tes conseils empreints de sagesse que je reformule avec mes mots: "Une bonne collaboration est avant tout une rencontre humaine".

Peu de gens liront en détail au delà de ces remerciements. Jean-Michel Loubes et Roberto Imbuzeiro Oliveira en font partie et je vous remercie grandement pour votre temps consacré à la lecture de ce manuscrit. Je suis encore une fois très honoré que vous ayez accepté de me lire.

Ce travail est en grande partie basé sur un article qui m'aura beaucoup influencé au cours de ces années. Il a été écrit par Olivier Catoni que j'ai le grand plaisir de compter parmi les membres de mon jury de thèse. Merci à toi Olivier pour les quelques échanges (mathématiques ou non) que nous avons eus durant mon temps passé au DMA.

Lorsqu'une aventure scientifique se termine, une autre commence. Il y a un peu plus de deux mois, Antoine m'a ouvert les portes d'un nouveau monde, celui des statistiques semi-paramétriques. Merci de ta disponibilité, de ton enthousiasme infaillible et toutes nos discussions. Je suis heureux de te compter dans mon jury.

Nanterre a été une très bonne surprise. Je ne m'attendais pas à être aussi bien accueilli et à me sentir comme un poisson dans l'eau. Merci à toute l'équipe, Nathanaël,

Gabriel, Mélanie, Luc, les deux Laurent, Patrice,... et à l'équipe des thésards/ATER, Marie, Julien, Paul, Mouss, Zaid, Gabriela,...

J'ai commencé cette thèse au DMA de l'ENS Paris. A l'époque, ne connaissant rien aux méandres administratifs de la recherche, j'ai été accompagné par Zaina et Bénédicte. Le laboratoire tout entier est une telle merveille qu'il est impossible de quantifier le génie qui rode dans ces lieux. Je n'oublie pas les belles discussions autour d'un café en salle Cartan qui m'ont permis d'échanger quelques mots avec de vraies stars des mathématiques. Je n'oublie pas non plus les départs groupés pour la "Bodéga". Il fallait bien Igor, Ilaria, Clément, Oriane,... pour hanter agréablement ces lieux.

J'ai fait un cours passage par Barcelone au premier semestre 2014. Juste avant mon départ, Igor est entré dans mon bureau pour me proposer de rentrer dans l'aventure Maths-Park. J'ai foncé tête baissée. Il n'y a rien à regretter. J'ai donc des remerciements spéciaux à envoyer à Roger, Xavier et Igor pour m'avoir donné l'opportunité de transmettre la fièvre des mathématiques aux jeunes et moins jeunes quelques samedis après-midi à l'IHP. Il n'est jamais évident de débarquer dans une ville inconnue sans contacts et parlant à peine la langue. Pourtant, cela a été une de mes périodes préférées de ces années de thèse. Merci à Camille, Maëlle et Santiago pour les moments passés ensemble et les tapas.

A mi-parcours, il a fallu déménager à Orsay, reprendre ses marques et ses habitudes. J'atterris dans le bureau de Lionel, Lucie et Zheng, un bureau jalousé par la moitié des thésards. Par le petit passage "véranda" j'entrapercevais Cagri qui dans les derniers mois de ma thèse a su me dire de rentrer chez moi me reposer lorsque je patinais dans la semoule à 23h un soir de weekend. Je n'oublie pas le bureau 108 avec les incontournables Céline, Emilie, Gabriele et Pierre-Antoine avec qui j'ai toujours autant de plaisir à parler maths. Et ceux du rez-de-chaussé avec Elodie, Morzi, Pierre, Joseph (merci pour la carte de France) et mon ami Bobby.

Mon envie de faire des maths mon métier m'est venue autour de mes 12/13 ans mais un événement particulier m'a vraiment confirmé dans ce choix : La rencontre avec M. Debiesse, mon prof de maths de première. Tu continues de m'influencer quand je fais cours. Tu réussis à faire des cours avec le brio de ne laisser quasiment personne sur la touche. Merci pour m'avoir transmis un échantillon de cette pédagogie. A l'orée de cette thèse, je ne peux pas oublier de remercier Sylvie, Marie-Anne, Pascal, Yohann pour avoir réfléchi ensembles aux exercices, problèmes et (quelque fois) solutions des TDs. Vous avez été une équipe pédagogique de choc!

Il me faut aussi remercier les gens qui ont partagé mes questionnements au cours du M2. En particulier, Cyril, Paul, Bastien, Max, Cécile,... Avant cela, mes années à Cachan ont été grandioses par les rencontres, entre autres, d'Antoine, Nicolas, Xan, Jules, Sarah, Kevin, Romain, Guillaume, Lionel, Caroline, Marc, et plein d'autres...

Puisqu'on régresse, parlons d'amis encore plus vieux. Je garde bien sûr en souvenir toute la bande des fantastiques, Wawan, Moptan, Trotro et plus tard Sisid et Perrine. J'ai aussi une pensée pour Camille (tout jeune papa), Hugo et Florian avec lesquels j'ai grandi et qui comptent énormément.

Et plus récents : la Drôme team. Ce devait être une équipe d'un été seulement, pourtant beaucoup de liens forts se sont construits au fil du temps. Je tiens fort à vous tous. Je pense évidemment à Antoine, Caron, Lucie, Emilie, Valérie, Alice, Maïté, Tigrane, Ileyk, Mikael.

Un grand merci à Catherine et Pierre-Marie pour leur accueil et leur soutien ainsi qu'a Elise, Myriam, Manu et Rastacouette.

Deux personnes ont partagé beaucoup de folles histoires avec moi. El colonel et el Greco se reconnaîtront. Dans mon esprit, il y a toujours un bras mécanique dérivant dans l'espace et un chauffeur de tramway polonais criant "you're off !!"

Trois chaleureux remerciements vont bien sûr à Claude (avec nos labyrinthes) mon père, Louisette (avec nos Sudokus et casses-têtes) ma mère et Théo (avec notre chambre partagée et délimitée, les cartes Magic et le BOW sur la piscine) mon frère. Je chéris notre cohésion, notre façon d'être liés et libres, et nos belles différences. Vous êtes et resterez mes meilleurs partenaires de coinche. Je vous aime.

Je ne peux terminer ces remerciements sans remercier Lucie (mon troisième directeur de thèse). Tu as été un soutien indéfectible lors de mes périodes de doutes, une compagne d'allégresse pour partager mes joies et la personne qui me comprend le mieux du monde. Pour tout ça, merci!

Enfin, toi, elle, lui, vous qui avez poursuivi la lecture de ces remerciements de l'infini jusqu'à ce point, je te, le, vous remercie du fond du cœur.

Table des matières

Chapitre 1

Introduction et motivations

Ce chapitre est une présentation des principaux axes de recherche qui ont conduit à l'élaboration de cette thèse. On y présente brièvement le cadre et les motivations principales de l'estimation robuste. Les diverses contributions originales sont, ensuite, discutées et commentées. Le chapitre se conclura sur les pistes de recherche futures envisagées.

Sommaire

1.1 Motivations de l'estimation robuste

L'estimation de la moyenne d'une quantité aléatoire résumée en une variable aléatoire X est un enjeu central en statistique. L'estimateur le plus naturel pour répondre à cette question est l'estimateur de la moyenne empirique. Pour un échantillon $X_1, \dots, X_n$ à valeurs réelles et de même loi que X, la moyenne empirique est donnée par

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \, .$$

La loi forte des grands nombres assure que la moyenne empirique converge presque sûrement vers l'espérance $\mathbb{E}[X]$. Plus précisément,

$$\lim_{n \to \infty} \overline{X}_n = \mathbb{E}[X] \quad \text{avec probabilité 1} \, .$$

Il est bien connu que le théorème limite central permet de quantifier l'erreur asymptotique de cet estimateur autour de sa moyenne. Plus précisément,

$$\lim_{n \to \infty} \mathbb{P}\left(\frac{\overline{X}_n - \mathbb{E}[X]}{\sigma / \sqrt{n}} \leq z \right) = \Phi(z)$$

où Φ est la fonction de répartition de la loi gaussienne centrée réduite et σ est la variance de X_1. La vitesse de convergence est de l'ordre de $\sigma n^{-1/2}$. Le point de vue choisi dans cette thèse se distingue de l'étude asymptotique découlant des outils liés au théorème limite central. En effet, nous faisons le choix de développer ce que nous appellerons, par la suite, des inégalités de concentration. Elles ont l'avantage de donner des bornes sur la probabilité d'erreur d'estimation $\mathbb{P}(|\overline{X}_n - \mathbb{E}[X]| > t)$ à n (le nombre d'observations) fixé pour un seuil t et sont, par définition, non-asymptotiques.

Parmi toutes les inégalités de concentration, la plus simple est sans doute l'inégalité de Hoeffding.

Theorem 1 (Hoeffding)**.** *Soient $X_1, \dots, X_n$ des variables aléatoires indépendantes telles que pour tout $i \leq n$, X_i est à valeurs dans $[a_i, b_i]$ presque sûrement. Soit $S = \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i])$. Alors pour tout $t > 0$,*

$$\mathbb{P}(S \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right) \, .$$

A elle seule, elle a motivé beaucoup de travaux pour adapter les bornes très fines obtenues dans de nombreux contextes statistiques. L'hypothèse centrale de l'inégalité de Hoeffding est la bornitude des variables aléatoires $(X_i)_i$. Pour des variables aléatoires à valeurs dans $[0, 1]$, la vitesse de convergence est de l'ordre de $n^{-1/2}$ et la concentration obtenue est sous-gaussienne (voir définition 1). L'idée directrice de cette thèse a été de s'affranchir de l'hypothèse de bornitude tout en conservant une vitesse de convergence comparable à celle obtenue dans l'inégalité de Hoeffding.

La plupart des résultats des sections suivantes s'appliquent à des variables aléatoires de variance finie sans être bornées. Nous utiliserons le mot robuste pour parler des estimateurs que nous exposerons dans ce manuscrit. Il est à noter que cette dénomination est différente de la notion de robustesse communément utilisée dans la littérature. Nous en donnerons donc une définition rigoureuse au chapitre 2.

1.1.1 Fiabilité et lois à queue lourde

Le contexte d'utilisation des résultats de concentration présentés plus bas est typiquement donné par les distributions à queue lourde. La définition 2 nous assure que les variables aléatoires à queue lourde ne possèdent pas de moment exponentiel fini. A l'inverse, il est à remarquer que l'hypothèse de bornitude implique immédiatement des moments exponentiels finis (i.e., $\mathbb{E}[\exp(\lambda X)] < \infty$) pour toute valeur du paramètre λ. Nous ferons une hypothèse sur les distributions que nous considérerons : généralement, une variance finie, sauf pour les théorèmes 6 et 40 où nous supposerons seulement un moment d'ordre $1 + \varepsilon$ avec $0 < \varepsilon < 1$.

Nous verrons en section 2.1.2 que la moyenne empirique ne permet pas de donner une estimation fiable de l'espérance dans le cas de variables aléatoires à queue lourde. Un contre-exemple est donné par les variables aléatoires α-stables (voir définition 3). En effet, les lois α-stables possèdent des moments finis d'ordre p pour tout $p < \alpha$ (avec $\alpha > 1$). En outre, la probabilité $\mathbb{P}(|\overline{X}_n - \mathbb{E}[X]| > t)$ que l'erreur d'estimation dépasse un certain seuil t est de l'ordre de $1/(t^\alpha n^{\alpha-1})$ alors que l'on recherche une borne du type $\exp(-t^2/(nC))$ où C est une constante dépendante de la loi de l'échantillon. Un autre candidat naturel d'estimateur de l'espérance est la médiane empirique M_n définie comme le quantile d'ordre 1/2. La vitesse de convergence de M_n vers $\mathbb{E}[X]$ est aussi en $1/\sqrt{n}$ (M_n et $\overline{X}_n$ étant séparés d'au plus $\sigma/\sqrt{n}$). En pratique, la médiane est moins sensible aux valeurs extrêmes que la moyenne empirique $\overline{X}_n$. Néanmoins, cet estimateur peut comporter un biais ce qui limite son utilisation. Pour ces raisons, les estimateurs classiques $\overline{X}_n$ et M_n seront mis de côté au profit de techniques plus robustes présentées plus bas.

1.1.2 Deux estimateurs robustes

Dans le but d'obtenir une erreur de type sous-gaussienne, nous nous sommes tournés vers des estimateurs moins sensibles à la présence de valeurs particulières très éloignées de la moyenne que les estimateurs proposés plus haut. L'inspiration nous a été donnée par deux articles proposant de nouvelles techniques d'estimation de la moyenne.

Le premier définit un estimateur de la moyenne comme le zéro d'une fonction décroissante (voir Catoni [20]). Cette technique très utilisée et très étudiée porte le nom de *M*-estimation. Il se trouve qu'un choix judicieux de cette fonction (voir (2.5)) permet d'obtenir une borne de type sous-gaussienne sur la probabilité d'erreur sous la seule hypothèse de variance finie. Dans la suite nous appellerons cet estimateur, l'estimateur par "soft-troncature" (troncature douce) ou estimateur de Catoni.

Le second article étudie un estimateur portant le nom de “median-of-means” ou médiane des moyennes (voir Lerasle et Oliveira [42]). Introduit plus tôt par Nemirovski et Yudin [53] et plus formellement par Jerrum, Valiant et Vazirani [36], cet estimateur combine les avantages de la moyenne empirique et de la médiane empirique. Nous verrons que cet estimateur assure une erreur d'ordre de grandeur comparable à celui de l'estimateur par soft-troncature.

Ces deux estimateurs définis rigoureusement en sections 2.2.2 et 2.2.3 formeront la base des outils de cette thèse. Nous en donnerons une étude auto-contenue et comparative en section 2.2. Tous les théorèmes originaux présentés plus bas utiliseront l'une ou l'autre de ces techniques.

1.2 Contributions originales

Cette partie est dédiée à l'exposition rapide des résultats originaux de cette thèse. Nous ne rentrerons pas dans les détails de toutes les applications possibles de nos résultats. Les chapitres 3 et 4 ont fait l'objet de deux articles à paraître.

1.2.1 Estimation en grande dimension (c.f. section 2.3)

Les deux articles cités plus haut traitent uniquement du cas unidimensionnel. En section 2.3 nous nous intéresserons à des généralisations de ces techniques en dimension $d > 1$.

L'estimation de la moyenne en grande dimension est autant un défi théorique que pratique. L'élaboration de méthodes d'estimation peu sensibles aux valeurs extrêmes a fait l'objet d'une imposante littérature scientifique. La plupart des articles concernés introduisent un contexte statistique conditionnant fortement le choix de tel ou tel estimateur. Pour des exemples de régression matricielle sous des hypothèses de petit rang voir Candès et Recht [18] et Recht, Fazel et Parrilo [58]. La recherche de techniques robustes pour le problème de l'analyse en composantes principales peut être trouvé dans Candès, Li, Ma et Wright [17] ou dans Zhang et Lerman [67] où la matrice de rang faible est retrouvée exactement. Plus généralement, Hsu et Sabato [32] ainsi que Minsker [52] se sont intéressés à l'estimateur median-of-means en en donnant une version multidimensionnelle.

Contribution. En parallèle, l'étude menée en section 2.3.2 s'intéressera à développer un estimateur en utilisant la technique de soft-troncature pour l'estimation de la moyenne d'une variable aléatoire à valeurs dans $\mathbb{R}^d$. Hormis une hypothèse technique d'isotropie, aucune hypothèse sur la structure statistique de la variable aléatoire de loi inconnue n'est requise. La borne de concentration obtenue sera de type sous-gaussien. En particulier, nous regarderons l'impact du facteur de dimension d dans les bornes de concentration développées.

1.2.2 Minimisation du risque empirique (c.f. chapitre 3)

La méthode de minimisation du risque empirique est une technique très commune en statistique et est couramment utilisée dans de nombreux contextes. Pour une fonction de risque fixée f, le risque empirique est donné par la formule

$$R_{ERM} := \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

où les X_i sont des copies indépendantes de la variable X. La fonction f est choisie dans un ensemble $\mathcal{F}$. Un estimateur réalisant le minimum de R_{ERM} sur l'ensemble $\mathcal{F}$ sera appelé estimateur par minimisation du risque empirique. Dans la suite de cette thèse, nous supposerons l'existence de ce minimum. Si plusieurs fonctions f sont valides,

le choix arbitraire de l'une d'entre elles constituera notre estimateur. Les résultats présentés plus bas restent valides quelque soit ce choix. Une référence très complète sur ce sujet est fournie par Koltchinskii [39].

Dans de très nombreux contextes d'application, la fonction f peut être interprétée comme le coût (en un sens très large) d'un choix de valeurs particulières d'un ensemble de paramètres. Nous citerons cinq exemples classiques d'application de la minimisation du risque empirique où le choix de la classe de fonction $\mathcal{F}$ est explicite. Voir la section 2.4.1.

Pour des fonctions de risque vérifiant des propriétés de convexité et de régularité, cette méthode permet de donner une estimation vérifiant une borne de confiance de type sous-gaussienne. Ce cas est traité en faisant appel à l'estimateur median-of-means dans Hsu et Sabado [32]. Par ailleurs, un article de Audibert et Catoni [7] s'intéresse à l'estimation par minimisation du risque empirique dans le cas paramétrique en utilisant l'estimateur par soft-troncature.

Contribution. Au chapitre 3, nous développerons une procédure de minimisation du risque empirique donnant des bornes de confiance non-asymptotique de type sous-gaussiennes en faisant appel à l'estimateur par soft-troncature introduit par Catoni [20]. Une application à la régression linéaire par moindres carrés et une application à la classification non-supervisée sont données en section 3.4. Nous donnerons aussi quelques simulations.

1.2.3 Estimation multivariée et U-statistiques (c.f. chapitre 4)

Les U-statistiques apparaissent naturellement dans de nombreux problèmes d'estimation multivariée. Le cadre dans lequel nous nous plaçons dans cette thèse est celui du livre de de la Peña et Giné [23]. La raison du choix des U-statistiques pour traiter le problème d'estimation multivarié sera expliqué en section 2.5. Étant donné une fonction $h : \mathbb{R}^m \to \mathbb{R}$ et une collection de variables aléatoires identiquement distribuées et indépendantes $X_1, \dots, X_n$ (avec $n \geq m$), l'enjeu est de déterminer avec la plus grande précision possible la valeur $m_h = \mathbb{E}\left[h(X_1, \dots, X_m)\right]$ inconnue. Cette fonction h sera supposée symétrique (i.e., h reste inchangée par toute permutation des variables). La U-statistique

$$U_n(h) = \frac{(n-m)!}{n!} \sum_{(i_1,\dots,i_m)\in I_n^m} h(X_{i_1}, \dots, X_{i_m})$$

où

$$I_n^m = \left\{(i_1, \dots, i_m) : 1 \leq i_j \leq n,\ i_j \neq i_k \ \text{ if } \ j \neq k\right\}$$

est un estimateur sans biais de m_h. Arcones et Giné [6] montrent que $U_n(h)$ vérifie une inégalité de concentration de type sous-gaussien lorsque la fonction h est bornée. Nous rappelons ce résultat au Théorème 21 en section 2.5.1. L'inégalité de Hoeffding ne s'applique pas immédiatement à $U_n(h)$. En effet, les termes de la somme définissant

$U_n(h)$ ne sont pas tous indépendants entre eux. Nous aurons recours à des techniques comme le découplage ou encore la décomposition de Hoeffding (voir section 4.4).

Contribution. Nous nous sommes intéressés à définir une version robuste de l'estimateur par U-statistique précédent et à montrer des résultats de concentration de type sous-gaussien sous des hypothèses plus faibles que la bornitude de la fonction h. Nous donnerons au chapitre 4 des résultats dans le cas de variance finie et des résultats dans le cas de moment d'ordre $1 + \epsilon$ fini (avec $\epsilon < 1$).

1.3 Perspectives de recherche et travaux futurs

Tout au long de ce manuscrit nous avons fait une recherche systématique de champs d'application des techniques d'estimation robustes (soft-troncature et median-of-means) pour affaiblir les hypothèses d'inégalités de concentration existantes. Dans cette optique, nous avons cherché à diversifier les contextes statistiques étudiés et à rendre compte de l'efficacité des nouvelles techniques proposées. Nous sommes passés par l'estimation robuste en grande dimension, la minimisation de risque empirique et l'estimation multivariée par les U-statistiques. Le travail est bien entendu inachevé et de nombreuses questions restent en suspens. Cette partie est à lire au regard des résultats des chapitres qui suivront.

Convergence en distance L_1

Il n'est, à l'heure actuelle, pas tout à fait clair si les estimateurs de soft-troncature et median-of-means convergent pour la distance $\mathbb{E}\left[|\widehat{\mu} - \mu|\right]$ où μ est la moyenne et $\widehat{\mu}$ est l'estimateur. En effet, les inégalités de concentration données en section 2.2 n'impliquent pas (contrairement à l'habitude) une borne en distance L_1 entre l'estimateur et la moyenne. En effet, le niveau de confiance δ apparaît dans la définition même de nos estimateurs. Il est possible, pour l'estimateur de Catoni, de s'en affranchir quitte à perdre en performance (voir section 2.2.2). Cependant, la condition $n \geq 4\log(\delta^{-1})$ pour l'estimateur de Catoni et la condition $\ln(\delta^{-1}) \leq N \leq \frac{n}{2}$ pour median-of-means nous empêchent d'obtenir directement des bornes en espérance. Cette étape supplémentaire pourrait étendre le champ d'application de ces techniques.

Implémentation pratique de l'estimateur de la section 2.3.2

Il est à noter que dans l'état, l'estimateur défini en section 2.3.2 ne se programme pas facilement. Il n'est pas immédiat (et sans doute de grande complexité) de construire un estimateur en pratique appartenant au polytope convexe $\mathcal{P}_W$. On pourrait penser à un autre estimateur défini comme un minimum de la fonction

$$R(\mu) := \frac{1}{n\alpha} \sum_{i=1}^{n} \Phi\left(\alpha \left\|X_i - \mu\right\|\right)$$

où Φ est une primitive de la fonction croissante ϕ définie en section 2.2.2. En pratique, cette définition alternative fournit un estimateur facile à implémenter (comme minimum d'une fonction convexe) mais il ne nous a pas été possible de développer des bornes théoriques pour cet estimateur. Une possibilité pour de futurs travaux pourrait être de trouver un lien entre ces deux estimateurs et ainsi franchir le fossé qui sépare théorie et pratique.

L'entropie $\gamma_1(\mathcal{F}, D)$

Le théorème 24 fournit une borne de confiance sur la quantité $m_{\widehat{f}} - m_{f^*}$ mesurant la qualité de notre estimation. Ce théorème fait apparaître un premier terme déjà présent dans Catoni [20] et deux termes mesurant la complexité de la classe de fonctions $\mathcal{F}$: l'entropie $\gamma_1(\mathcal{F}, D)$ et l'entropie $\gamma_2(\mathcal{F}, d)$. Or, le terme d'entropie $\gamma_1(\mathcal{F}, D)$ est d'un ordre de convergence (en $1/n$) plus petit que celui du terme d'entropie $\gamma_2(\mathcal{F}, d)$ associée à la distance L_2 (en $1/\sqrt{n}$). Pour s'assurer de la finitude de l'entropie $\gamma_1(\mathcal{F}, D)$, notre théorème suppose que le diamètre de l'espace $\mathcal{F}$ pour la norme infinie est borné. Étant donné son ordre d'importance inférieur dans les bornes, nous nous attendrions à pouvoir prouver une version du théorème 24 sans ce terme d'entropie

$$m_{\widehat{f}} - m_{f^*} \leq 6\left(\alpha v + \frac{2\log(\delta^{-1})}{n\alpha}\right) + L\log(2\delta^{-1})\left(\frac{\gamma_2(\mathcal{F}, d)}{\sqrt{n}}\right).$$

Si une telle borne s'avère impossible à montrer, nous voudrions exhiber un contre exemple de classe $\mathcal{F}$ pourvu d'une entropie finie $\gamma_2(\mathcal{F}, d)$ et d'une variance finie uniformément sur toute la classe $\mathcal{F}$ et pour laquelle la procédure définie au chapitre 3 n'assure pas de convergence de $m_{\widehat{f}}$ vers m_{f^*}. Ceci aiderait à mieux comprendre l'importance du terme d'entropie $\gamma_1(\mathcal{F}, D)$.

Applications à d'autres fonctions de risque

Nous traitons deux applications de la technique de minimisation du risque empirique en section 3.4 pour des données suivant des distributions possiblement à queue lourde. Ces deux illustrations semblent convaincantes mais ne constituent qu'une petite partie des possibles champs d'application. Nous voudrions simuler plus en profondeur notre technique et éprouver l'algorithme sur des jeux de données réels.

Affaiblir une condition d'isotropie

Notre résultat d'estimation en grande dimension contenu dans le théorème 11 formule une hypothèse d'isotropie sur la loi de X. Cette hypothèse s'interprète comme une bonne répartition de la mesure de probabilité associée à X sur toutes les directions de l'espace ambiant. Il est facile de voir que sans cette condition, il est possible d'inclure la variable aléatoire X dans un espace de dimension arbitrairement grande en ajoutant artificiellement des coordonnées nulles sans changer la valeur de la variance V. Dans ce cas, le résultat (2.9) se réduirait à

$$\left\|\widehat{\mu} - \mathbb{E}[X]\right\| \leq 16\sqrt{2}\sqrt{\frac{V\ln(4)}{n}}.$$

En particulier, le terme comportant le niveau de confiance δ disparaît, ce qui rend la borne déterministe. Or dans le cas de variables aléatoires non bornées, il est possible

de trouver une réalisation de l'échantillon $X_1, \dots, X_n$ tel que l'estimateur de Catoni unidimensionnel soit arbitrairement éloigné de l'espérance.

Une amélioration possible serait de remplacer cette condition d'isotropie par des conditions plus naturelles et plus faciles à vérifier sur un échantillon. Nous pourrions par exemple essayer d'utiliser l'inégalité PAC-bayésienne en page 159 de Catoni [19] dans ce but.

1.4 Plan de la thèse

Cette thèse est principalement centrée sur les inégalités de concentration de nouveaux estimateurs robustes de la moyenne. Elle se découpe en trois chapitres.

Le premier chapitre s'intéresse aux estimateurs de soft-troncature et median-of-means pour eux-mêmes. On y définit ensuite des versions généralisées de ces estimateurs pour un cadre multidimensionnel. Nous introduisons ensuite la notion de minimisation du risque empirique pour une classe de fonction donnée. Nous citons en particulier des résultats tirés du domaine des processus empiriques tels que l'inégalité entropique de Dudley [25] ou l'inégalité de "generic chaining" (chaînage générique) de Talagrand [61]. Ces résultats nous permettent de donner un résultat de concentration pour un minimiseur du risque empirique.

Le deuxième chapitre présente deux résultats principaux de concentration de l'estimateur par minimum du risque empirique dans le cas de variables aléatoires à queue lourde. Trois applications sont ensuite discutées : la régression linéaire pour une perte L_1, la régression linéaire pour une perte L_2 et enfin la procédure de classification non-supervisée par l'algorithme "k-means" (ou k-moyennes). Nous terminerons ce chapitre par quelques simulations donnant les performances, en pratique, de notre estimateur sur les applications précédemment introduites.

Le troisième chapitre traite de l'estimation de la moyenne pour des fonctions multivariées. Nous aurons recours aux techniques de U-statistique présentes dans de livre de de la Peña et Giné [23]. Nous adapterons l'estimateur par median-of-means pour définir un estimateur robuste de la moyenne ayant des vitesses de convergence comparables à celle données dans Arcones et Giné [6]. Un exemple d'application au clustering sera développé en fin de chapitre.

Chapter 2

Two robust estimators of the mean

This chapter introduce the mathematical background of this manuscript. The notion of heavy-tailed distributions is defined in Section 2.1. We present in details the two robust estimators of the mean which are used repeatedly throughout the thesis. Sections 2.3, 2.4 and 2.5 discuss applications of robust estimation to various statistical contexts.

Contents

2.1 Why another estimator of the mean?

2.1.1 A first step towards robust estimation

In the task of learning the distribution of a random variable X, perhaps the most fundamental parameter is the mean $\mathbb{E}[X]$. A natural estimate of this quantity is the empirical mean $\overline{X}$. For an independent, identically distributed (i.i.d.) sample $X_1, \dots, X_n$, the empirical mean is given by

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \,. \tag{2.1}$$

In this thesis we are interested in the finite-sample behavior of estimators of the mean. This is in contrast with asymptotic convergence-of-measure theorems such as the central limit theorem. For an important reference on asymptotic theorems we refer to van der Vaart [64]. The results presented here are concentration inequalities. As an example of a non-asymptotic inequality, we start with Hoeffding's inequality [31] for sums of bounded random variables.

Theorem 2 (Hoeffding's inequality)**.** *Let $X_1, \dots, X_n$ be independent random variables such that for each $i \leq n$, X_i takes its values in the interval $[a_i, b_i]$ almost surely. Let $S = \sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])$. Then for any $t > 0$,*

$$\mathbb{P}(S \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) .$$

In particular, let $X_1, \dots, X_n$ be independent random variables taking values in the interval $[a, b]$. Then, for any *confidence level* $\delta \in (0, 1/2)$, with probability at least $1 - 2\delta$,

$$\left|\overline{X} - \mathbb{E}[X]\right| \leq \sqrt{\frac{B \ln(\delta^{-1})}{n}} \tag{2.2}$$

where $B = (b-a)^2/2$. We encounter many inequalities of the form of (2.2) in this thesis. In particular, we are interested in sub-Gaussian estimators of the mean.

Definition 1. *A real-valued random variable X is called sub-Gaussian with variance factor $V > 0$ if for any $\delta \in (0, 1)$,*

$$\left|X - \mathbb{E}[X]\right| \leq \sqrt{V \ln(\delta^{-1})} \tag{2.3}$$

*with probability at least $1 - 2\delta$. For any **fixed** $\delta \in (0, 1)$, we say that a random variable X satisfies a sub-Gaussian concentration inequality with confidence level δ if* (2.3) *holds.*

Remark. Our definition of a Gaussian concentration inequality is dependent of the confidence level δ fixed in advance. In consequence, a Gaussian concentration inequality does not necessary imply that the estimator of interest is a sub-Gaussian random

variable.

For other equivalent definitions of sub-Gaussian random variables, see Boucheron, Lugosi and Massart [14]. Many concentration results for estimators of the mean of a function require a boundedness assumption. However this assumption excludes the case of heavy-tailed random variables. Heavy-tailed distributions are probability distributions supported on $\mathbb{R}$ whose tails are not exponentially bounded.

Definition 2. *A distribution is said to be heavy-tailed at $+\infty$ (resp. at $-\infty$) if, for any $\lambda > 0$,*

$$\lim_{t\to+\infty} e^{\lambda t}\mathbb{P}(X > t) = +\infty \quad \left(resp. \quad \lim_{t\to\infty} e^{\lambda t}\mathbb{P}(-X > t) = +\infty\right).$$

A distribution is heavy-tailed if it is heavy-tailed at $+\infty$ or at $-\infty$.

The results presented in this thesis hold for a particular class of heavy-tailed distributions. We assume a finite-moment of order $1+\epsilon$ with $\epsilon > 0$. In most of the results presented below, ϵ equals to 1 and the finite-moment assumption reduces to a finite variance condition. An estimator of the mean $\widehat{\mu}$ is called δ-*robust* if, for any heavy-tailed sample with –at least– a $1+\epsilon$ finite-moment, it satisfies a sub-Gaussian concentration inequality with a confidence level δ and variance factor of the order of $1/n$.

2.1.2 An example of heavy-tailed distributions: α-stable laws

For $\alpha \in (0,2)$, α-stable laws are examples of heavy-tailed distributions. They have finite moments of small order and infinite higher-order moments.

Definition 3. *Let $\gamma > 0$ and $0 < \alpha \leq 2$. We say that a random variable X has an* α-stable *law $S(\gamma,\alpha)$ if for all $u \in \mathbb{R}$,*

$$\mathbb{E}\exp(iuX) = \exp(-\gamma^\alpha |u|^\alpha).$$

Definition 3 is less general than Definition 1.6 in Nolan [54]. Another reference on α-stable laws is Zolotarev [68]. Gaussian random variables belong to the class of α-stable laws for $\alpha = 2$. Among all α-stable random variables, Gaussian random variables are the only ones with finite moments of all orders. The behavior of α-stable random variables in the case $0 < \alpha < 2$ is different. They are heavy-tailed. See Figure 2.1 for an illustration. We now state a few facts on α-stable laws. We use the notation $h(x) \underset{x\to a}{\sim} g(x)$ for $\lim_{x\to a} h(x)/g(x) = 1$.

Proposition 3. *Let $\gamma > 0$ and $\alpha \in (0,2)$. Let $X_1,\dots,X_n$ be i.i.d. random variables of law $S(\gamma,\alpha)$. Let $f_{\gamma,\alpha} : x \mapsto \mathbb{R}$ be the density function of X_1. Let $S_n = \sum_{1\leq i\leq n} X_i$. Then*

(i) $f_{\gamma,\alpha}(x)$ is an even function ,

(ii) $f_{\gamma,\alpha}(x) \underset{x\to+\infty}{\sim} \alpha\gamma^\alpha c_\alpha x^{-\alpha-1}$ with $c_\alpha = \sin\left(\frac{\pi\alpha}{2}\right)\Gamma(\alpha)/\pi$,

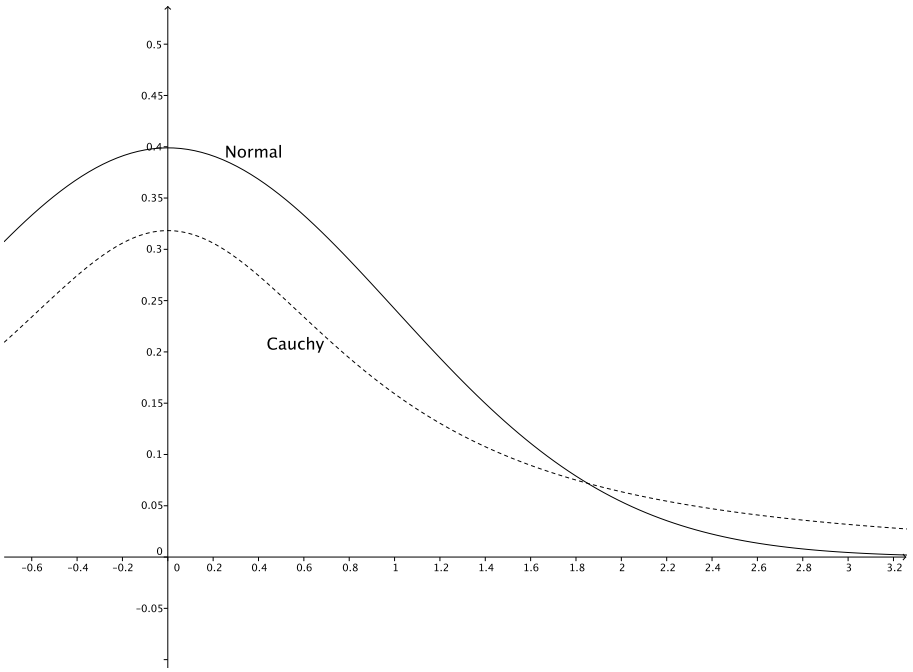


Figure 2.1: Comparison between the standard normal distribution and the Cauchy distribution. The Cauchy distribution is an α-stable distribution with $\alpha = 1$. It has density function $f(x) = \frac{1}{\pi}\frac{\gamma}{\gamma^2+x^2}$.

(iii) $\mathbb{E}\left[X_1^p\right]$ *is finite for any* $p < \alpha$ *and is infinite whenever* $p \geq \alpha$,

(iv) S_n *has a* α*-stable law* $S(\gamma n^{1/\alpha}, \alpha)$.

Proof. (i) and (iv) follow directly from the definition. (ii) is proved in the introduction of Zolotarev [68]. (iii) is a consequence of (ii). □

An α-stable random variable for $\alpha \in (1,2)$ has finite moments of order p up to the order α. Thus, for $\alpha \in (1,2)$, α-stable random variables are heavy tailed. By *(iv)*, the sum S_n is still an α-stable random variable and we know exactly the law of the estimator $\overline{X}$ in that case. By *(i)* we know that the X_i are centered random variables. Finally, by *(ii)*, there exists a constant K depending only on α and γ such that

$$\mathbb{P}\left(\left|\overline{X}\right| > t\right) \underset{t\to+\infty}{\sim} \frac{K}{t^\alpha n^{\alpha-1}} .$$

This last equivalence shows that the tail probabilities of $\overline{X}$ are far from the sub-Gaussian tails in (2.3). In particular, this shows that the empirical mean estimator $\overline{X}$ is not

robust. The two robust procedures that we present in the following section achieve sub-Gaussian concentration inequalities even for heavy-tailed distributions.

The rest of this chapter is organized as follows. In Section 2.2, we introduce two estimators of the mean and the basic performance bounds. Section 2.2.2 is dedicated to an estimator defined by Catoni [20] via a truncation function. Section 2.2.3 is centered on an estimator known as the median-of-means estimator introduced in Nemirovski and Yudin [53]. Section 2.3 is about multidimensional mean estimation and generalizations of robust estimators in $\mathbb{R}^d$ or more generally in Banach spaces. In Section 2.4, we introduce empirical risk minimization and generic chaining. Section 2.5 focuses on the special case of mean estimation of multivariate functions. Ideas from the theory of U-statistics are used to get concentration bounds.

2.2 Robust procedures

In this section we introduce the notion of robust estimators. Section 2.2.1 exposes some estimators commonly used if some outliers are present in a data sample $X_1, \dots, X_n$. These techniques have theoretical limitations. To adress this, two recent estimators of the mean of a real valued random variable are exposed in Sections 2.2.2 and 2.2.3 and their performance bounds are given.

2.2.1 Toward robust estimators: trimming, truncation and flattened M-estimation

In practice, data trimming is commonly used to avoid the effect of outliers. It consists in a reduction of the sample size before computing the standard empirical mean. One orders the data $X_1, \dots, X_n$, chooses a small percentage (say 5%) and removes the first and last 2.5% of the X_i from the sample. This technique is called trimming. An example of an econometric application of trimming techniques is Andersen, Dobrev and Schaumburg [3]. A related technique is to use a truncation function $\psi_K : x \mapsto x\mathbb{1}_{\{|x|\leq K\}} + K\mathbb{1}_{\{|x|>K\}}$ on the sample and to compute the empirical mean of the sample $(\psi_K(X_i))_i$. Hoeffding's inequality (Theorem 2) applies and sub-Gaussian concentration is obtained around $\mathbb{E}[\psi_K(X)]$. Both techniques have the inherent limitation to produce biased estimators. Also, it is a challenge to scale the parameter K for truncation or the percentage of vanished data points. The choice of the parameters is strongly dependent of the underlying distribution. This lack of generality has led researchers to search for alternatives. To our knowledge, Huber [33] was the first to develop robust estimators using the formalism of *M-estimators*. The perhaps most natural M-estimator is the *least-squares* estimator $\widehat{\mu}_{ls}$ defined as the minimizer of the functional

$$\mu \mapsto \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \,. \tag{2.4}$$

The existence of the minimum is ensured by the convexity of the function $x \mapsto x^2$. By simple calculation, the minimum $\widehat{\mu}_{ls}$ equals the empirical mean and the least-squares estimator is just another definition of the empirical mean. More generally, an estimator is said to be an M-estimator if it is obtained by minimizing a functional depending on the data $(X_i)_i$. One idea of Huber [33] was to change the definition of the estimator $\widehat{\mu}_{ls}$ by replacing the function $x \mapsto x^2$ by another convex function. The main focus of Huber and Ronchetti [34] is asymptotic normality of M-estimators but finite-sample results are not discussed. In the next section, we present a particular M-estimator where the choice of the convex function is explicit.

2.2.2 M-estimation through soft truncation

The following estimator has been first introduced by Catoni in [20]. It can be seen as a special case of robust M-estimation defined in Section 2.2.1. It relies on a *soft truncation* function ϕ where ϕ is any non-decreasing function satisfying the following inequalities: For all $x \in \mathbb{R}$,

$$\begin{aligned} \phi(x) &\leq \ln\left(1 + x + \frac{x^2}{2}\right) \\ \phi(x) &\geq -\ln\left(1 - x + \frac{x^2}{2}\right) . \end{aligned} \tag{2.5}$$

By an easy analysis, such functions exist (see Fig. 2.2). For any $\alpha > 0$, we define

$$r_\alpha(\mu_\alpha) = \frac{1}{n\alpha} \sum_{i=1}^{n} \phi\left(\alpha[X_i - \mu_\alpha]\right) .$$

The truncation function ϕ has been chosen in order to have a non-decreasing r_α. This implies that r_α has at least one zero. Here α is some real positive parameter to be chosen later. Define the mean estimator $\widehat{\mu}_\alpha$ as any element of the set $\{\mu_\alpha : r_\alpha(\mu_\alpha) = 0\}$. The function r_α can be seen as the derivative of a concave loss function R_α. An advantage of the soft truncation presented here is the concentration inequality of Theorem 4. The main ingredient for the success of this estimator is the particular form of the function ϕ. One obtains the following:

Theorem 4 (Catoni [20])**.** *Let $X, X_1, \ldots, X_n$ be i.i.d. real-valued random variables. Let V be a real number such that $Var(X) \leq V < \infty$. Let $\delta \in (0, \frac{1}{2})$ and $\alpha = \sqrt{\frac{2 \ln \delta^{-1}}{nV}}$. Suppose $n \geq 4 \ln \delta^{-1}$. Then*

$$|\widehat{\mu}_\alpha - \mathbb{E}[X]| \leq 2\sqrt{2} \sqrt{\frac{V \ln \delta^{-1}}{n}} ,$$

with probability at least $1 - 2\delta$.

We give the proof of Theorem 4 in Section 2.6. Note that this bound does not hold for any confidence level δ. Indeed, for a fixed n, the theorem is valid only for δ in the

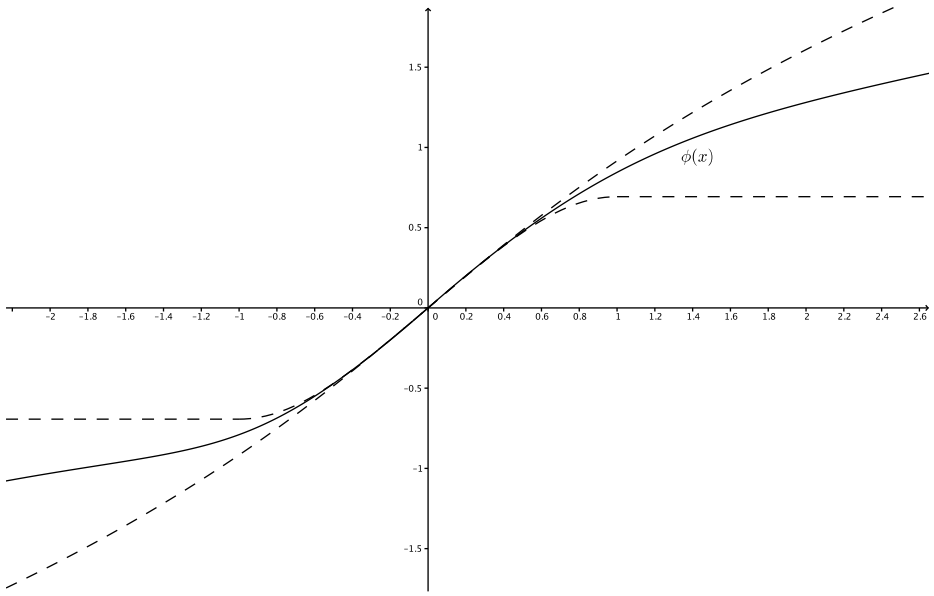


Figure 2.2: A choice (solid line) for the truncation function ϕ is plotted here. The dashed lines correspond to the extremal cases for the function ϕ that satisfy (2.5).

range $\delta \geq e^{-n/4}$. However, for reasonably large sample size n, the term $e^{-n/4}$ is tiny and the conclusion of Theorem 4 holds for a wide range of δ. Note also that the estimator itself depends on the two parameters δ and V. The choice $\alpha = \sqrt{2/nV}$ leads to the slightly poorer bound

$$|\widehat{\mu}_\alpha - \mathbb{E}[X]| \leq \sqrt{\frac{2V}{n}}\left(1 + \ln \delta^{-1}\right)$$

but allows the estimator to be independent of δ in its definition. Also, the estimation procedure can be adapted to the case when no suitable upper bound of the variance is available using PAC-Bayesian tools. This is analyzed in Catoni [20]. We do not discuss this point here.

2.2.3 Median-of-means estimator

In this section, we present another estimator of the mean, the so-called median-of-means estimator. The first occurrence of this estimator in the literature seems to trace back to Nemirovski and Yudin [53] and Jerrum, Valiant and Vazirani [36]. The idea is to divide the data into independent groups on which we compute an estimator (for example the empirical mean) that is only weakly concentrated. Then by selecting the median of the estimators, a robust estimator is obtained. The result described here is contained in Lerasle and Oliveira [42]. Formally, the setting is as follows. Let $X_1, \ldots, X_n$

be i.i.d. real-valued random variables of finite variance. Let $N \leq n$ be an integer and let $\mathcal{B} = (B_1, \dots, B_N)$ be a partition of $\{1, \dots, n\}$ satisfying the regularity property

$$\forall k = 1, \dots, N, \quad \left| |B_k| - \frac{n}{N} \right| \leq 1 \,. \tag{2.6}$$

For all non-empty subsets A of $\{1, \dots, n\}$, we denote by

$$\mu_A = \frac{1}{|A|} \sum_{i \in A} X_i$$

the empirical mean on the subset A. The median $\mathrm{Med}(a)$ of a subset $a = \{a_1, \dots, a_V\}$ of $\mathbb{R}$ is any real number t such that

$$|\{i \,:\, a_i \leq t\}| \geq \frac{N}{2} \quad \text{and} \quad |\{i \,:\, a_i \geq t\}| \geq \frac{N}{2} \,.$$

We finally introduce the median-of-means estimator:

$$\widehat{\mu}_{\mathcal{B}} := \mathrm{Med}(\mu_{B_1}, \dots, \mu_{B_N}) \,.$$

The following result gives a concentration result for the median-of-means estimator.

Theorem 5 (Lerasle & Oliveira [42])**.** *Let $X, X_1, \dots, X_n$ be i.i.d. real-valued random variables. Assume that $Var\,(X) < \infty$. Let $\delta \in (0, 1/2)$, let N be an integer such that $\ln(\delta^{-1}) \leq N \leq \frac{n}{2}$ and let $\mathcal{B}$ be a partition satisfying the regularity condition* (2.6)*. Then we have*

$$|\widehat{\mu}_{\mathcal{B}} - \mathbb{E}\,[X]\,| \leq 2\sqrt{6e}\sqrt{\frac{Var\,(X)\,N}{n}}$$

with probability at least $1 - 2\delta$.

A good choice of N is $\left\lceil \ln(\delta^{-1}) \right\rceil$. In that case, the bound has the same form as in Theorem 4, up to a constant factor. Once again, the estimator $\widehat{\mu}_{\mathcal{B}}$ depends on δ. This time, the dependence cannot be avoided by another choice of the parameters. Unlike the case of the estimator of Section 2.2.2, $\mathrm{Var}\,(X)$ is not a parameter of the median-of-means estimator. An assumption on the $1 + \epsilon$ moment is enough to get the following refinement:

Theorem 6. *Let $X_1, \dots, X_n$ be i.i.d. real-valued random variables. Assume that*

$$\left(\mathbb{E}|X_1 - \mathbb{E}\,[X_1]\,|^{1+\epsilon}\right)^{\frac{1}{1+\epsilon}} \leq V \,.$$

Let $\delta \in (0, 1/2)$, let N be an integer such that $\ln(\delta^{-1}) \leq N \leq \frac{n}{2}$ and let $\mathcal{B}$ be a partition satisfying the regularity condition (2.6)*. Then we have*

$$|\widehat{\mu}_{\mathcal{B}} - \mathbb{E}\,[X]\,| \leq 72^{\frac{1}{1+\epsilon}} V \left(\frac{N}{n}\right)^{\frac{\epsilon}{1+\epsilon}} ,$$

with probability at least $1 - 2\delta$.

Theorems 5 and 6 are proved in Section 2.6.

2.3 Mean estimation in high dimension

We have seen in Section 2.2 two techniques to estimate the mean of a real-valued random variable under the weak assumption of a finite variance. In this thesis, our goal is to apply the estimators defined in Sections 2.2.2 and 2.2.3 in various statistical problems. A first step is to adapt these procedures to the multidimensional case. For an overview of the use of robust statistics in learning theory, see Huber [34] or Hubert, Rousseeuw and Van Aelst [35] and references therein.

2.3.1 Median-of-means estimator in Hilbert spaces

Using an alternative definition of the median called geometric median, Minsker [52] developed a version of the median-of-means estimator in the general context of Banach spaces. We present here a weaker version in Hilbert spaces. Let $\mathbb{X}$ be a Hilbert space with norm $\|\cdot\|$ associated with the scalar product. For any finite collection of points $x_1, \dots, x_k \in \mathbb{X}$, a *geometric median* is any point x_* satisfying

$$\sum_{j=1}^{k} \|x_* - x_j\| = \min_{y \in \mathbb{X}} \sum_{j=1}^{k} \|y - x_j\| \ .$$

We denote by $\mathrm{Med}(x_1, \dots, x_k)$ any such x_*. For $0 < p < \alpha < \frac{1}{2}$, define

$$\psi(\alpha, p) := (1-\alpha) \ln \frac{1-\alpha}{1-p} + \alpha \ln \frac{\alpha}{p} \ .$$

We denote by

$$p^*(\alpha) := \max \{p \in (0, \alpha) \ : \ \psi(\alpha, p) \geq 1\} \ .$$

We now state a theorem of Minsker [52].

Theorem 7 (Minsker [52])**.** *Let $\alpha \in (0, 1/2)$. Let $X_1, \dots, X_n$ be i.i.d. random variables taking values in $(\mathbb{X}, \|\cdot\|)$. Let $\delta \in (0,1)$, let N be an integer such that $\ln(\delta^{-1}) \leq N \leq \frac{n}{2}$ and let $\mathcal{B} = (B_1, \dots, B_N)$ be a partition satisfying the regularity condition* (2.6)*. Denote by μ_{B_i} the empirical mean on the subset B_i. Let $\widehat{\mu} := Med(\mu_{B_1}, \dots, \mu_{B_N})$. Then, with probability at least $1 - \delta$,*

$$\left\| \widehat{\mu} - \mathbb{E}[X] \right\| \leq K_\alpha \sqrt{\frac{\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] N}{n}}$$

where $K_\alpha = (1-\alpha)\sqrt{\frac{2}{(1-2\alpha)p^*(\alpha)}}$.

We give a proof in Section 2.6. For a general version in Banach spaces, see Minsker [52].

Remark 1. *The theorem holds in Banach spaces for $K_\alpha = \frac{2(1-\alpha)}{1-2\alpha}\sqrt{\frac{2}{p^*(\alpha)}}$. See the proof of Lemma 2.1 in Minsker [52].*

2.3.2 A generalized Catoni estimator

Here, we adopt another approach for the estimation of the mean in the multidimensional case. This technique uses the soft truncation function ϕ defined in Section 2.2.2.

A simple robust estimation on each coordinate

We begin with a definition of an estimator based on the estimator defined in Section 2.2.2 for each coordinate. This estimator has the advantage to be simple to implement. The price to pay is a $\log(d)$ factor in the bound. For all $x \in \mathbb{R}$ and $\alpha > 0$, let $\phi_\alpha(x) = \phi(\alpha x)/\alpha$. We denote by $\mathbf{e} = (e_1, \dots, e_d)$ the canonical basis in $\mathbb{R}^d$ and by $\Psi_{\mathbf{e}}(\mu_1, \dots, \mu_d)$ the function:

$$\Psi_{\mathbf{e}}(\mu_1, \dots, \mu_d) = \left(\frac{1}{n} \sum_{i=1}^{n} \phi_{\alpha_j} \left(X_i \cdot e_j - \mu_j \right) \right)_{1 \le j \le d}$$

We denote by $\Psi_{e_j}(\mu_j)$ the j component of $\Psi_{\mathbf{e}}(\mu_1, \dots, \mu_d)$ and by $\hat{\mu}_j$ a solution of $\Psi_{e_j}(\mu_j) = 0$. Then the following result holds.

Theorem 8. *Let $X_1, \dots, X_n$ be i.i.d. random variables taking values in $\mathbb{R}^d$. For any $j \in \{1, \dots, d\}$, let V_j be a bound of the variance of the random variable $X \cdot e_j$ and let $V = \sum_j V_j$. Let $\delta \in (0, \frac{1}{2})$ and $\alpha_j = \sqrt{\left(2 \ln(\frac{d}{\delta})\right) / (nV_j)}$. Let $\hat{\mu}$ be a solution of $\Psi_{\mathbf{e}}(\mu_1, \dots, \mu_d) = 0$. Then, under the condition $n \ge 4 \ln(\frac{d}{\delta})$, with probability at least $1 - 2\delta$,*

$$\left\| \hat{\mu} - \mathbb{E}[X] \right\| \le 2\sqrt{2} \sqrt{\frac{V \ln(\frac{d}{\delta})}{n}}$$

where $\|.\|$ refers to the euclidean norm in $\mathbb{R}^d$.

This result is slightly poorer than Theorem 3.1 in Minsker [52]. This is due to the log factor coming from a union bound over the d coordinates. The estimator is sensitive to the choice of the orthogonal basis $(e_1, \dots, e_d)$. It requires the knowledge of an upper bound of the variance of every coordinate of X in advance.

Proof. The proof follows from Theorem 4 along with a union bound on every coordinate. □

An estimator using convex geometry

In this section we define another estimator using a soft truncation function with a better confidence bound. We proceed by identifying an area where –with high probability– the mean $\mathbb{E}[X]$ is located. This area has a small diameter. The definition of the estimator is based on the following remark.

Remark. HYPERPLANE OF SOLUTION
We denote by $\cdot$ the canonical scalar product in $\mathbb{R}^d$. For any unit vector $w \in \mathbb{R}^d$ (i.e., $\|w\| = 1$) we define $\hat{\mu}_w$ as any solution of

$$\Psi_w(\hat{\mu}_w) = \frac{1}{n}\sum_{i=1}^{n} \phi_\alpha\left(X_i \cdot w - \hat{\mu}_w\right) = 0\,. \tag{2.7}$$

The zeros of the function

$$\overline{\Psi}_w : \left| \begin{array}{ccc} \mathbb{R}^{\mathbf{d}} & \longrightarrow & \mathbb{R} \\ \mu & \longmapsto & \Psi_w(\mu \cdot w) \end{array} \right.$$

form an hyperplane defined by $H_w = \{\mu \in \mathbb{R}^d : \mu \cdot w = \hat{\mu}_w\}$.

The algorithm proposed below invokes some tools of convex geometry. We begin with a generalization of Theorem 4 for a specified direction w. In this part, we assume that the distribution of X is *almost spherical* in the following sense.

Definition 4. *We say that a random variable X is c-almost spherical if*

$$|||\Sigma||| \leq c\frac{Tr(\Sigma)}{d}$$

where Σ is the covariance matrix of X, $|||\cdot|||$ is the canonical matrix norm and Tr is the trace operator.

For any distribution, we always have $|||\Sigma||| \geq \frac{\mathrm{Tr}(\Sigma)}{d}$. Indeed, $|||\Sigma|||$ is equal to the largest eigenvalue of Σ and $\mathrm{Tr}(\Sigma)$ is the sum of the d eigenvalues. Equality is achieved given by isotropic distributions (i.e., distributions invariant under rotation around the expected value $\mathbb{E}[X]$). We adapt Theorem 4 for the case of c-almost spherical distributions.

Lemma 9. *Let w be any unit vector in $\mathbb{R}^d$. Let $X, X_1, \ldots, X_n$ be i.i.d. random variables taking values in $\mathbb{R}^d$ and let $V > 0$ be such that $\mathbb{E}\left[\|X - \mathbb{E}[X]\|\right]^2 \leq V$. Assume that X is c-almost spherical. Let $\Delta \in (0, \frac{1}{2})$, $\alpha = \sqrt{(2d\ln(\Delta^{-1}))/(ncV)}$ and let $\hat{\mu}_w$ be the real number defined in Equation* (2.7)*. Then, under the condition $n \geq 4\ln\Delta^{-1}$, with probability at least $1 - 2\Delta$,*

$$\left|\hat{\mu}_w - \mathbb{E}[X] \cdot w\right| \leq 2\sqrt{2}\sqrt{\frac{cV\ln\Delta^{-1}}{dn}}\,.$$

Proof. We apply Theorem 4 for the independent and identically distributed random variables $(X_i \cdot w)_i$. The variance of the random variable $X \cdot w$ is bounded by $|||\Sigma|||$ and $\mathrm{Tr}(\Sigma) \leq V$. □

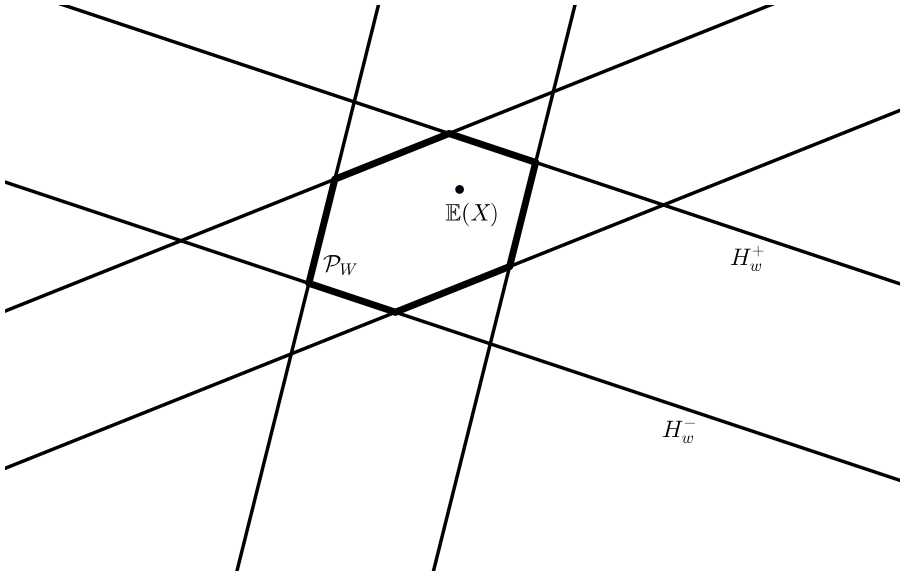


Figure 2.3: The polytope $\mathcal{P}_W$ in the plane.

A consequence of Lemma 9 is that with probability at least $1 - 2\Delta$, $\mathbb{E}[X]$ is squeezed between two parallel hyperplanes H_w^+ and H_w^- (see Fig. 2.3) defined by

$$H_w^+ = \left\{ \mu \,:\, \mu \cdot w = \hat{\mu}_w + 2\sqrt{2}\sqrt{\frac{cV \ln \Delta^{-1}}{dn}} \right\},$$
$$H_w^- = \left\{ \mu \,:\, \mu \cdot w = \hat{\mu}_w - 2\sqrt{2}\sqrt{\frac{cV \ln \Delta^{-1}}{dn}} \right\}.$$

In the spirit of Section 2.3.2, we use Lemma 9 multiple times. The main idea in this section is the choice of the directions w. This choice is given by a 1-net on the unit sphere S^{d-1}. A *1-net* Z is a set of points in the sphere S^{d-1} such that every point on the sphere is at distance at most 1 from any point in Z. The following theorem gives an estimate of the number of elements in a 1-net on the sphere.

Theorem 10. *There exists a 1-net on the sphere S^{d-1} with at most 4^d elements.*

A proof of this fact may be found in Ball [8]. The proof uses a concentration argument and the theory of packing and covering. For more information, see Rogers [60]. Let W be a 1-net on the unit sphere S^{d-1} with at most 4^d elements. We refer to W as the set of directions. Define

$$\mathcal{P}_W := \bigcap_{w \in W} \left\{ \mu \,:\, |\mu \cdot w - \hat{\mu}_w| \le 2\sqrt{2}\sqrt{\frac{cV \ln \Delta^{-1}}{dn}} \right\}. \tag{2.8}$$

We now state the main theorem in this section.

Theorem 11. *Let $X, X_1, \dots, X_n$ be i.i.d. random variables taking values in $\mathbb{R}^d$ and let $V > 0$ be such that $\mathbb{E}\left[\|X_1 - \mathbb{E}[X_1]\|\right]^2 \leq V$. Assume that X is c-almost spherical. Let W be a 1-net on the unit sphere S^{d-1}. Let $\delta \in (0, \frac{1}{2})$ and, for any $w \in W$, let $\hat{\mu}_w$ be a solution of*

$$\sum_{i=1}^{n} \phi_\alpha \left(X_i \cdot w - \hat{\mu}_w\right) = 0$$

with $\alpha = \sqrt{\frac{d \ln(4^d \delta^{-1})}{2ncV}}$. Then, under the condition $n \geq 4 \ln(4^d \delta^{-1})$, with probability at least $1 - 2\delta$, the set $\mathcal{P}_W$ is not empty and for any $\mu \in \mathcal{P}_W$,

$$\left\|\mu - \mathbb{E}[X]\right\| \leq 16\sqrt{2}\sqrt{\frac{cV\ln(4)}{n} + \frac{cV\ln(\delta^{-1})}{dn}} \tag{2.9}$$

where $\|\cdot\|$ is the euclidean norm on $\mathbb{R}^d$.

Any point of the set $\mathcal{P}_W$ is an estimator of the mean $\mathbb{E}[X]$ with an error controlled by the right-hand side of (2.9). The hypotheses of Theorem 11 differ from the ones given in Minsker [52] because of the c-almost spherical assumption. First, observe that the bound in (2.9) becomes less precise if we artificially increase the dimension. Assume that the distribution of the data satisfies the c-almost spherical assumption in $\mathbb{R}^d$. We can naturally embed the distribution into $\mathbb{R}^{2d}$ by adding 0 coordinates to the last d coordinates of the vectors $X_1, \dots, X_n$. The variance factor V is conserved. The new distribution satisfies the $2c$-almost spherical assumption then the conclusion of Theorem 11 holds with an extra factor 2 in front of the first term under the square root. This effect comes from the way that the set $\mathcal{P}_W$ is built. The set W is a 1-net on the unit sphere in $\mathbb{R}^d$ whereas the distribution of X is possibly supported by a sub-vector space of $\mathbb{R}^d$. Another sub-optimality fact is due to Lemma 9 where the variance of $X \cdot w$ is roughly bounded by $|||\Sigma|||$.

If the dimension d is large and if the c-almost spherical assumption holds, the second term becomes negligible and the bound reduces to the term

$$16\sqrt{2}\sqrt{\frac{cV\ln(4)}{n}}$$

which does not involve the confidence level δ. In that case, the bound becomes almost deterministic which is an advantage over the result contained in [52, Corollary 4.1]. The case $c = 1$ is typically given by uncorrelated coordinates of identical variance.

For the proof of Theorem 11, we introduce the notion of polytope associated to a family of vectors.

Definition 5. *Let m be an integer and let $v_1, \dots, v_m$ be vectors in $\mathbb{R}^d$. The polytope of m facets associated with $v_1, \dots, v_m$ is*

$$K = \{x : \langle x, v_i \rangle \leq 1 \text{ for } 1 \leq i \leq m\} .$$

Note that 0 always belongs to K. It is easy to see that the set K is convex. From Definition 5, any 1-net on the sphere defines a polytope. The dissimilarity of a polytope K from the euclidean ball is defined below.

Definition 6. *The dissimilarity $d(K, B)$ between a polytope K and the euclidean ball B is the least positive r such that there exists a positive λ such that*

$$B \subset \lambda K \subset rB$$

Note that $d(K, B)$ is not a distance. The theory of packing and covering leads to the following results.

Theorem 12. *Any 1-net defines a polytope with dissimilarity less than 2 from the euclidean ball.*

The proofs of Theorem 12 and Theorem 10 may be found in Ball [8] in the more general context of linear independent distances between symmetric convex bodies. We give a proof of Theorem 12 in Section 2.6.

Proof of Theorem 11: We invoke Lemma 9 for every w in W. The union bound gives

$$\mathbb{P}\left(\sup_{w\in W}\left|\hat{\mu}_w - \mathbb{E}[X]\cdot w\right| \geq 2\sqrt{2}\sqrt{\frac{cV\ln\Delta^{-1}}{dn}}\right) \leq 4^d\Delta\,. \tag{2.10}$$

Equation (2.10) implies that, with probability at least $1 - 4^d\Delta$ the expected value $\mathbb{E}[X]$ belongs to $\mathcal{P}_W$. In particular, $\mathcal{P}_W$ is not empty. The set $\mathcal{P}_W$ is a polytope but $\mathcal{P}_W$ is not necessarily associated with a 1-net set. We handle this problem by including $\mathcal{P}_W$ into a bigger polytope associated with a 1-net. Let $r = 4\sqrt{2}\sqrt{(V\ln\Delta^{-1})/(dn)}$. Assume that the event in equation (2.10) is satisfied. In particular, the set $\mathcal{P}_W$ is not empty. We fix y to be any point in the polytope $\mathcal{P}_W$. The point y is at distance at most r of each hyperplane $H^+_{R(w)}$ or $H^-_{R(w)}$ for any $w \in W$. Define the polytope

$$\overline{\mathcal{P}}_W(y) := \left\{x \in \mathbb{R}^d \,:\, (x-y)\cdot w \leq r \text{ for } w \in W\right\}\,.$$

By definition, the polytope $\overline{\mathcal{P}}_W(y)$ contains $\mathcal{P}_W$. By Theorem 10, the 1-net property of the class W gives,

$$rB(y,1) \subset \overline{\mathcal{P}}_W(y) \subset 2rB(y,1)$$

where $B(y,1)$ is the euclidean ball centered at y and of radius 1. We have proved that $\mathcal{P}_W$ is included in a euclidean ball of radius $2r$. Inequality (2.10) implies that with probability at least $1 - \Delta 4^d$, $\mathbb{E}[X]$ belongs to $\mathcal{P}_W$ and consequently to $2rB(y,1)$. Since both $\mathbb{E}[X]$ and $\hat{\mu}$ belong to $2rB(y,1)$, $\left\|\hat{\mu} - \mathbb{E}[X]\right\|_2 \leq 4r$. Set $\Delta = \delta/4^d$ and the theorem is proved. □

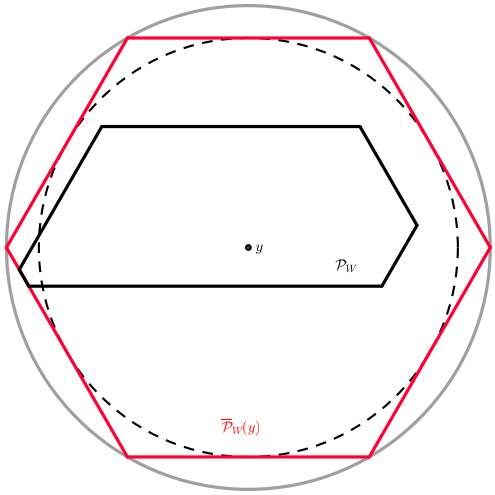


Figure 2.4: The polytope $\overline{\mathcal{P}}_W(y)$. The two circles represent the dissimilarity of $\overline{\mathcal{P}}_W(y)$ from the euclidean ball.

2.4 Empirical risk minimization

In this section we introduce the empirical risk minimization procedure. Section 2.4.1 is dedicated to some examples of application of this procedure. Section 2.4.2 is centered on a central tool for the proofs developped in Chapter 3: the entropy bounds. These can be applied in the context of empirical risk minimization. The obtained result is a performance bound given in Section 2.4.3.

The empirical risk minimization technique can be described as follows. Let X be a random variable in some measurable set $\mathcal{X}$. Let $\mathcal{F}$ be a set of functions $f : \mathcal{X} \mapsto [0, +\infty)$. The *risk* of a function $f \in \mathcal{F}$ is given by $m_f = \mathbb{E}[f(X)]$. The ideal choice is defined by $f^* = \operatorname{argmin}_{f\in\mathcal{F}} m_f$. The statistical learning challenge consists in the elaboration of a technique that approximates the ideal function based on an i.i.d. sample $X_1, \dots, X_n$ drawn from the distribution of X. A fundamental method is *empirical risk minimization* defined by

$$f_{ERM} = \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

where, without loss of generality, a minimizer is assumed to exist. Multiple minimizers may exist and one can choose one of them arbitrarily. This general framework takes its origin in Vapnik and Chervonenkis [65]. For a recent summary on the theory of empirical risk minimization one can consult Koltchinskii [39].

2.4.1 Examples of empirical risk minimization

Empirical risk minimization embrace different statistical designs. We introduce below five examples of the use of empirical risk minimization. The first two examples are treated in detail in Section 3.4. In the following, the minimizers are assumed to exist.

Example 13 (Least-squares regression)**.** *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed random variables taking values in $\mathbb{R}^d \times \mathbb{R}$. Suppose given a class $\mathcal{G}$ of functions $\mathbb{R}^d \to \mathbb{R}$. The risk of each $g \in \mathcal{G}$ is defined by the L_2 loss*

$$R(g) = \mathbb{E}\left[(g(X) - Y)^2\right]$$

where the pair (X, Y) has the same distribution as the (X_i, Y_i) and is independent of them. A least-squares estimator is a minimizer $\widehat{g}$ of the empirical risk

$$\frac{1}{n}\sum_{i=1}^{n}(g(X_i) - Y_i)^2 \, .$$

Example 14 (k-means clustering)**.** *Let $X, X_1, \ldots, X_n$ be independent and identically distributed random variables taking values in $\mathbb{R}^d$. Let P denote the distribution of X. Let $k \geq 2$ be a positive integer fixed. A set of cluster centers $C = \{y_1, \ldots, y_k\} \subset \mathbb{R}^d$ and a quantizer $q : \mathbb{R}^d \to C$ form what is called a clustering scheme. Given a distortion measure $\ell : \mathbb{R}^d \times \mathbb{R}^d \to [0, +\infty)$, the risk of a clustering scheme – also called distortion – is defined by*

$$D_k(P, q) = \mathbb{E}\ell(X, q(X)) \, .$$

In this example, we take $\ell(x, y) = \left\|x - y\right\|^2$. Given a set of cluster centers C, a clustering scheme q is a nearest neighbor quantizer if, for all $x \in \mathbb{R}^d$,

$$\left\|x - q(x)\right\|^2 = \min_{y_i \in C} \left\|x - y_i\right\|^2 \, .$$

It is know that a minimizer of the distortion has to be search among the nearest neighbor quantizers (see Linder [45]). The empirical distortion of a nearest neighbor quantizer is defined by

$$D_k(P_n, q) = \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\ldots,k} \left\|X - y_j\right\|^2$$

and a minimizer of $D_k(P_n, q)$ over the subset $(\mathbb{R}^d)^k$ of all possible cluster centers is called a k-means estimator.

The literature on k-means clustering has become rich in the past decade. See Pollard [55] and Abaya and Wise [1] for reference on the convergence of the distortion of a k-means estimator towards the optimal distortion. The rate of convergence is also of crucial interest (see Levrard [43, 44] and Biau, Devroye and Lugosi [12] and references therein).

Example 15 (Binary classification)**.** *Let (X, Y) be a couple of random variables where X takes its values in a measurable space $\mathcal{X}$ and Y is a random label taking values in $\{-1, 1\}$. A classifier f is a function $\mathcal{X} \to \mathbb{R}$. For any function $\phi : \mathbb{R} \to \mathbb{R}$, the ϕ-risk of a classifier f is given by*

$$R(f) = \mathbb{E}\left[\phi(Y f(X))\right] .$$

The classical choice of ϕ is the 0–1 loss $\phi(x) = \mathbb{1}_{\{x \leq 0\}}$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent random variables distributed as (X, Y) and independent from it. A minimizer of the empirical risk

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i f(X_i))$$

form an estimator obtained by the empirical risk minimization technique.

There is other possible choices for the function ϕ in the definition of the ϕ-risk (see Lecué [40, 41])

Example 16 (Maximum likelihood method)**.** *Let X be a random variable of probability measure P absolutely continuous with respect to the Lebesgue measure λ and let $\mathcal{P}_\Theta = \{p_\theta \, : \, \theta \in \Theta\}$ be a set of probability densities with respect to λ, indexed by a parameter set Θ. The maximum likelihood estimator is defined by*

$$\widehat{p} := \underset{\mathcal{P}_\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \Big(- \log p_\theta(X_i) \Big) .$$

For a study of the convergence of maximum likelihood methods, see van de Geer [63].

Example 17 (A penalized empirical risk minimization)**.** *Recall the notation of Example 16. We assume that the random variable X has a density p^*. The best L_2 approximation of p^* is given by*

$$\bar{p} = \underset{\mathcal{P}_\Theta}{\operatorname{argmin}} \left(-\frac{2}{n} \sum p(X_i) + \|p\|_2^2 \right)$$

since $\|p - p^*\|_2^2 = -2 \int_{\mathbb{R}} p p^* d\lambda + \|p\|_2^2 + \|p^*\|_2^2$ and $\int_{\mathbb{R}} p p^* d\lambda = \mathbb{E}\left[p(X)\right]$ is approximated by $\frac{1}{n} \sum p(X_i)$.

2.4.2 Empirical processes and Dudley's bound

The general theory of empirical processes for M-estimation (see for example van de Geer [63]) may be used to give confidence bounds for the risk of the empirical risk minimizer. One of the key results of empirical processes theory is the chaining bound of Dudley. This machinery may be used to obtain laws of large numbers, central limits theorems, concentration inequalities, expected value upper bounds, etc., where

the common denominator is the metric entropy. It requires the existence of a pseudo metric on the set $\mathcal{F}$. The metric entropy is, in some sense, a measure of the statistical complexity of a set $\mathcal{F}$. In this thesis we use a more general tool called *generic chaining* (see Talagrand [61]).

Definition 7. *Given $\beta > 0$ and a pseudo metric space (T,d), an admissible sequence $(A_n)_n$ is an increasing sequence of partitions of T such that $|A_n| \leq 2^{2^n}$ and $A_0 = T$. For any $t \in T$, $A_n(t)$ denotes the element of the partition A_n that contains t. The generic chaining entropy is*

$$\gamma_\beta(T, d) := \inf \sup_{t \in T} \sum_{n \geq 0} 2^{n/\beta} \Delta(A_n(t))$$

where the infimum is taken on all admissible sequences and $\Delta(A_n(t))$ is the diameter of $A_n(t)$.

We are now ready to state the fundamental theorem of generic chaining.

Theorem 18 (Talagrand [61])**.** *Let (T, d) be a pseudo metric space and let $(X_t)_{t \in T}$ be a centered real-valued process such that for any $s, t \in T$ and any $u \geq 0$,*

$$\mathbb{P}\left(|X_t - X_s| \geq u\right) \leq 2 \exp\left(-\frac{u^2}{d(s,t)^2}\right).$$

Then there exists a universal constant L such that

$$\mathbb{E}\left[\sup_{t \in T} X_t\right] \leq L \gamma_2(T, d).$$

This theorem applies, in particular, to Gaussian processes where the distance is $d(s,t) = \left(\mathbb{E}\left[(X_s - X_t)^2\right]\right)^{1/2}$. In this case, the bound is optimal since, by Sudakov minoration [61, Lemma 2.1.2], $\mathbb{E}\left[\sup_{t \in T} X_t\right]$ is also lower bounded by a constant times $\gamma_2(T, d)$. We recall here a generalization of Theorem 18 under a more general assumption.

Theorem 19 (Talagrand [61])**.** *Let T be a space provided with two pseudo distances d_1 and d_2. Let $(X_t)_{t \in T}$ be a centered real-valued process satisfying for any $s, t \in T$ and any $u \geq 0$,*

$$\mathbb{P}\left(|X_t - X_s| \geq u\right) \leq 2 \exp\left(-\min\left(\frac{u^2}{d_2(s,t)^2}, \frac{u}{d_1(s,t)}\right)\right).$$

Then there exists a universal constant L such that

$$\mathbb{E}\left[\sup_{t \in T} X_t\right] \leq L(\gamma_1(T, d_1) + \gamma_2(T, d_2)).$$

The proof is contained in Talagrand [61, Chapter 1]. It is important to note that there exist analogous theorems for the tail probabilities of $\sup X_t$, see van der Vaart [64]. The proofs of Theorem 18 and 19 are also contained in van der Vaart [64, Chapter 2]. A weaker inequality is Dudley's entropy bound. The generic chaining entropy bound can be bounded by Dudley's entropy: there exists a universal positive constant L such that for any $\beta = 1, 2$, we have

$$\gamma_\beta(T, d) \leq L \int_0^{\Delta(T)} (\log N(T, d, \epsilon))^{1/\beta} d\epsilon \tag{2.11}$$

where $N(T, d, \epsilon)$ refers to the minimal number of balls of radius ϵ needed to cover the set T. $\Delta(T)$ denotes the diameter of the set T.

2.4.3 A risk bound for a sub-Gaussian bounded class $\mathcal{F}$

Most of the bounds in the literature hold for bounded random variables $f(X)$ with $f \in \mathcal{F}$. Here we state one of the simplest results when the random variables $f(X)$ are uniformly bounded for any $f \in \mathcal{F}$.

Theorem 20. *Let $X_1, \ldots, X_n$ be i.i.d random variables taking values in $\mathcal{X}$. Let B be a positive real number and let $\delta \in (0, 1/2)$. We assume that for any $f \in \mathcal{F}$ and for any $x \in \mathcal{X}$, $f(x) \leq B$. For any $f, f' \in \mathcal{F}$, let $d(f, f') = \sup_{x \in \mathcal{X}} |f(x) - f'(x)|$. Let $\widehat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1} f(X_i)$ and let $f^* = \operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E} f(X)$. Then there exists a universal constant K such that*

$$m_{\widehat{f}} - m_{f^*} \leq K(\gamma_2(\mathcal{F}, d) + B) \sqrt{\frac{\ln(2\delta^{-1})}{n}}$$

with probability at least $1 - 2\delta$.

We postpone the proof of Theorem 20 to Section 2.6. Empirical risk minimization method can have a poor performance in case of heavy-tailed distributions. Indeed, the empirical mean is already a poor estimate of the mean for heavy-tailed random variables. In Chapter 3 we study empirical risk minimization for unbounded functions based on minimizing Catoni estimator defined in Section 2.2.2.

2.5 Multivariate estimation with U-statistics

In this section, we investigate the estimation of the mean of functions of several independent random variables via U-statistics. We base our notation on de la Peña and Giné [23]. The object of interest is a multivariate real-valued function $h : \mathcal{X}^m \to \mathbb{R}$. The function h is called a *kernel of order m*. Our goal is to estimate the quantity $\mathbb{E}h(X_1, \ldots, X_m)$ where the random variables X_i are independent and identically distributed. If one has access to an i.i.d. sample $X_1, \ldots, X_n$ and if m divides n, one can split the data into the sample $Y_1 := (X_1, \ldots, X_m), Y_2 := (X_{m+1}, \ldots, X_{2m}), \ldots$ and use a robust estimator defined

in Section 2.2.2 or in Section 2.2.3 as a replacement of the empirical mean estimator in (2.1) directly on this new sample. The $(Y_i)_i$ are independent and so Theorems 4 and 5 apply with no extra work. For example, the Catoni estimator $\widehat{\mu}_\alpha(h)$ is a solution of

$$\frac{m}{n\alpha}\sum_{i=0}^{n/m-1}\phi\left(\alpha\left[h(X_{im+1},\dots,X_{(i+1)m})-\mu_\alpha(h)\right]\right)=0 \tag{2.12}$$

for $\alpha=\sqrt{\frac{2m\ln\delta^{-1}}{nV}}$ and any fixed V such that $V\geq \mathrm{Var}\left(h(X_1,\dots,X_m)\right)$. A direct use of Theorem 4 for the sample $(Y_i)_i$ gives, with probability at least $1-\delta$,

$$|\widehat{\mu}_\alpha(h)-\mathbb{E}h(X_1,\dots,X_m)|\leq 2\sqrt{2}\sqrt{\frac{V\ln\delta^{-1}}{n}}\,. \tag{2.13}$$

The definition of the Catoni estimator in Equation (2.12) is based on the empirical mean estimator of the random variable $\frac{1}{\alpha}\phi\left(\alpha\left[h(X_{im+1},\dots,X_{(i+1)m})-\mu_\alpha(h)\right]\right)$. Nevertheless, the empirical mean estimator is known to have a larger variance (see Hoeffding [30]) than the U-statistics

$$U_n(h)=\frac{(n-m)!}{n!}\sum_{(i_1,\dots,i_m)\in I_n^m}h(X_{i_1},\dots,X_{i_m})\,,$$

where

$$I_n^m=\left\{(i_1,\dots,i_m):1\leq i_j\leq n,\ i_j\neq i_k\ \text{ if }\ j\neq k\right\}\,.$$

The principal idea behind the definition of $U_n(h)$ is to use every value $h(X_{i_1},\dots,X_{i_m})$ in order to reduce the variance of the estimator. But, unlike in (2.12), the terms in the sum are not all independent. In consequence, the theorems giving concentration inequalities for functions of independent random variables cannot be used. In Section 2.5.1, a concentration inequality for the U-statistics estimator for bounded random variables is recalled. Section 2.5.2 is dedicated to two examples of applications of U-statistics in the case of bivariate kernels.

2.5.1 A result for bounded kernels

In this section, we introduce the terminology of U-statistics. In the case of a bounded kernel h, a sub-Gaussian concentration inequality for $U_n(h)$ is obtained.

A kernel is symmetric if for all $x_1,\dots,x_m$ and for any permutation σ of $\{1,\dots,n\}$, $h(x_{\sigma_1},\dots,x_{\sigma_m})=h(x_1,\dots x_m)$. A symmetric kernel h is said to be *P-degenerate* of order $q-1$, $1<q\leq m$, if for all $x_1,\dots,x_{q-1}\in\mathcal{X}$,

$$\int h(x_1,\dots,x_m)dP^{m-q+1}(x_q,\dots,x_m)=\int h(x_1,\dots,x_m)dP^m(x_1,\dots,x_m)$$

and

$$(x_1,\dots,x_q)\mapsto\int h(x_1,\dots,x_m)dP^{m-q}(x_{q+1},\dots,x_m)$$

is not a constant function. Denote by $\mathbb{E}h$ the value of $\int h(x_1, \dots, x_m) dP^m(x_1, \dots, x_m)$. When $q = m$, the kernel h is said to be P-canonical. P-canonical kernels appear in Hoeffding's decomposition of the U-statistics $U_n(h)$. See (4.9) in Chapter 4 for a rigorous definition of Hoeffding's decomposition. The following theorem is due to Arcones and Giné [6].

Theorem 21 (Arcones and Giné [6])**.** *Let h be a kernel of order m such that $h - \mathbb{E}h$ is a symmetric P-canonical kernel. We assume $\|h\|_\infty < \infty$. Then there exist finite positive constants c_1 and c_2 depending only on m such that for any $\delta \in (0, 1)$,*

$$|U_n(h) - \mathbb{E}h| \leq c_1 \|h\|_\infty \left(\frac{\ln\left(\frac{c_2}{\delta}\right)}{n} \right)^{m/2}$$

with probability at least $1 - \delta$.

For any $m > 1$, the rate of convergence in n in Theorem 21 is faster than the result in (2.13). Here, the degenerate property of the kernel implies faster rates of convergence. In Chapter 4, we develop a generalized version of the median-of-means estimator for kernels of order m. A similar speed of convergence for unbounded heavy-tailed kernels – with a P-degenerate property – is obtained in Theorem 38.

2.5.2 Two applications for the case $m = 2$

The case $m = 2$ is of special interest. We present here two statistical learning problems where U-statistics are used.

The ranking problem

Let (X, Y) be a pair of random variables taking values in $\mathcal{X} \times \mathbb{R}$. The random variable X is the observed random variable and Y can be seen as a score of X. Let (X', Y') be a independent copy of (X, Y) and let $Z = Y - Y'$. We think about X being better than X' if $Z > 0$. In the ranking problem, we have access to X and X' but the random variables Y and Y' remain hidden. A ranking rule is a function $r : \mathcal{X} \times \mathcal{X} \to \{-1, 1\}$ where $r(x, x') = 1$ if the rule ranks x higher than x'. The ranking risk is given by

$$L(r) = \mathbb{P}\left(Z \cdot r(X, X') < 0\right) .$$

Assume that $(X_1, Y_1), \dots, (X_n, Y_n)$ are n independent copies of (X, Y). A natural estimate of L is

$$L_n(r) = \frac{1}{n(n-1)} \sum_{i,j \,:\, i \neq j} \mathbb{1}_{\{(Y_i - Y_j) \cdot r(X_i, X_j) < 0\}} . \tag{2.14}$$

The estimate $L_n(r)$ is a U-statistic with a bounded kernel. The kernel in (2.14) is bounded and Theorem 21 applies for the U-statistics $L_n(r)$ for $m = 1$. See Vayatis [66] for a recent

survey on ranking problems.

A clustering problem
Clustering techniques are about partitioning the data $X_1, \dots, X_n$ into a given finite number K of groups so that the observations in the same group are similar. For any partition $\mathcal{P}$ of the space $\mathcal{X}$, we denote by $\phi_{\mathcal{P}} : \mathcal{X}^2 \to \{0, 1\}$ the binary function

$$\forall (x, x') \in \mathcal{X}^2, \quad \phi_{\mathcal{P}}(x, x') = \sum_{C \in \mathcal{P}} \mathbb{1}_{\{(x,x') \in C^2\}}$$

that indicates if the two points x and x' belong to the same cell C of the partition. A measure of dissimilarity $D : \mathcal{X}^2 \to \mathbb{R}^+$ is a symmetric positive-definite function. The performance of a partition can be measured by $\mathbb{E}\left[D(X, X') \cdot \phi_{\mathcal{P}}(X, X')\right]$. A convenient estimate is the U-statistic

$$W_n(\mathcal{P}) = \frac{1}{n(n-1)} \sum_{i \neq j} D(X_i, X_j) \cdot \phi_{\mathcal{P}}(X_i, X_j) \, .$$

If the dissimilarity measure D is unbounded, Theorem 21 does not apply and the standard U-statistic W_n may be heavy-tailed. In Chapter 4, we develop an estimator to replace the U-statistic for possibly heavy-tailed distributions. In Chapter 3, we discuss a special case called K-means clustering.

2.6 Proofs

2.6.1 Proof of Theorem 4

The proof relies on a clever use of Markov's inequality. For any $\alpha > 0$,

$$\begin{aligned}
\mathbb{E}\left[\exp(n\alpha r(\mu))\right] &= \prod_{i=1}^{n} \mathbb{E}\left[\exp(\phi\left(\alpha[X_i - \mu]\right))\right] \\
&\leq \left(1 + \alpha \mathbb{E}\left[X - \mu\right] + \frac{\alpha^2}{2} \mathbb{E}\left[(X - \mu)^2\right]\right)^n \\
&\leq \left(1 + \alpha(\mathbb{E}\left[X\right] - \mu) + \frac{\alpha^2}{2} V + \frac{\alpha^2}{2} (\mathbb{E}\left[X\right] - \mu)^2\right)^n \\
&\leq \exp\left(n\alpha \left(\mathbb{E}\left[X\right] - \mu + \frac{\alpha}{2} V + \frac{\alpha}{2} (\mathbb{E}\left[X\right] - \mu)^2\right)\right)
\end{aligned}$$

By Markov's inequality, with probability at least $1 - \delta$,

$$r(\mu) < \mathbb{E}\left[X\right] - \mu + \frac{\alpha}{2} V + \frac{\alpha}{2} (\mathbb{E}\left[X\right] - \mu)^2 + \frac{\ln(\delta^{-1})}{n\alpha} =: B^+(\mu)$$

The expression $B^+(\mu)$ on the right-hand side has a zero if and only if $1 - 2\alpha(\alpha V/2 + \ln(\delta^{-1})/n\alpha) \geq 0$. Denote by μ^+ the smallest solution of $B^+(\mu) = 0$. It is given by

$$\mu^+ = \mathbb{E}[X] + \frac{1 - \sqrt{1 - 2\alpha\left(\frac{\alpha V}{2} + \frac{\ln(\delta^{-1})}{n\alpha}\right)}}{\alpha} \leq \mathbb{E}[X] + 2\left(\frac{\alpha V}{2} + \frac{\ln(\delta^{-1})}{n\alpha}\right).$$

Optimizing the parameter α gives $\alpha = \sqrt{\frac{2\ln(\delta^{-1})}{nV}}$ and $\mu^+ \leq \mathbb{E}[X] + 2\sqrt{2}\sqrt{\frac{V\ln(\delta^{-1})}{n}}$. Since r is a non-increasing function of μ, for all $\mu \geq \mu^+$, $r(\mu) < 0$ then $\widehat{\mu} \leq \mu^+$. The same argument gives the lower bound for $\widehat{\mu}$ using the second inequality in (2.5).

2.6.2 Proof of Theorem 5

We first show a weak concentration inequality of each empirical mean for each block B_k. By Chebychev's inequality,

$$\forall k = 1\ldots, N \quad \mathbb{P}\left(\mu_{B_k} - \mathbb{E}[X] \geq \sqrt{\frac{V}{|B_k|r}}\right) \leq r\,. \tag{2.15}$$

Let

$$N_r = \left|\left\{k \in \{1,\ldots,N\} \,:\, \mu_{B_k} - \mathbb{E}[X] \geq \sqrt{\frac{V}{|B_k|r}}\right\}\right|.$$

By definition of $\widehat{\mu}$ as a median,

$$\mathbb{P}\left(\widehat{\mu} - \mathbb{E}[X] > \sqrt{\frac{V}{|B_k|r}}\right) \leq \mathbb{P}\left(N_r \geq \frac{N}{2}\right).$$

Note that N_r is a sum of independent Bernoulli random variables of parameter bounded by r. Let B be a binomial random variable with parameters N and r. Then $\mathbb{P}\left(N_r \geq \frac{N}{2}\right) \leq \mathbb{P}\left(B \geq \frac{N}{2}\right)$. The Cramér transform of the Bernoulli distribution with parameter r is given, for $t \in [r, 1]$, by $\psi_r^*(t) = t\ln(\frac{t}{r}) + (1-t)\ln(\frac{1-t}{1-r})$. Then for $t = \frac{1}{2}$,

$$\mathbb{P}\left(B \geq \frac{N}{2}\right) \leq e^{-\frac{N}{2}\ln(\frac{1}{4r(1-r)})}.$$

We choose $r = (1 - \sqrt{1 - e^{-2}})/2$ and $N \geq \ln(\delta^{-1})$ so that $\mathbb{P}\left(B \geq \frac{N}{2}\right) \leq \delta$. The regularity condition ensures that for all k, $|B_k| \geq n/2N$. The constant $2\sqrt{6e} = \sqrt{2}\sqrt{12e}$ comes from $r \geq 1/(12e)$.

2.6.3 Proof of Theorem 6

The proof is similar to the proof of Theorem 5. Chebychev's inequality (2.15) is replaced by the following lemma. This result can be found in Bubeck, Cesa-Bianchi and Lugosi [16, Lemma 3].

Lemma 22. *Let $X_1, \dots, X_n$ be i.i.d. real random variables. Let $\epsilon \in (0,1]$. Assume that $\left(\mathbb{E}|X_1 - \mathbb{E}[X_1]|^{1+\epsilon}\right)^{\frac{1}{1+\epsilon}} < \infty$. Let $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$. Then for any $\delta \in (0,1)$ with probability at least $1-r$, we have*

$$\widehat{\mu} - \mathbb{E}[X] \leq \left(\mathbb{E}|X_1 - \mathbb{E}[X_1]|^{1+\epsilon}\right)^{\frac{1}{1+\epsilon}} \left(\frac{3}{rn^\epsilon}\right)^{\frac{1}{1+\epsilon}}$$

The end of the proof of Theorem 6 is similar with

$$N_r = \left| \left\{ k \in \{1, \dots, N\} \,:\, \mu_{B_i} - \mathbb{E}[X] \geq \left(\mathbb{E}|X_1 - \mathbb{E}[X_1]|^{1+\epsilon}\right)^{\frac{1}{1+\epsilon}} \left(\frac{3}{rn^\epsilon}\right)^{\frac{1}{1+\epsilon}} \right\} \right| .$$

2.6.4 Proof of Theorem 7

The proof uses the following lemma.

Lemma 23. *Let $\mathcal{H}$ be a Hilbert space. Let $x_1, \dots, x_k \in \mathcal{H}$ and let x_* be their geometric median. Fix $\alpha \in (0, 1/2)$ and $r > 0$. Let $C_\alpha = (1-\alpha)(1-2\alpha)^{-1/2}$. Assume that $z \in \mathcal{H}$ is such that $\|x_* - z\| > C_\alpha r$. Then there exists a subset $J \subseteq \{1, \dots, k\}$ of cardinality $|J| > \alpha k$ such that for all $j \in J$, $\left\|x_j - z\right\| > r$.*

Let

$$\varepsilon = \sqrt{\frac{2\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] N}{np^*(\alpha)}} .$$

By Chebychev's inequality, for all $1 \leq i \leq N$, $\mathbb{P}\left(\left\|\mu_{B_i} - \mathbb{E}[X]\right\| > \varepsilon\right) \leq p^*(\alpha)$. Assume that event $\mathcal{E} := \left\{\left\|\widehat{\mu} - \mathbb{E}[X]\right\| > C_\alpha \varepsilon\right\}$ occurs. Lemma 23 implies that there exists a subset $J \subseteq \{1, \dots, N\}$ of cardinality $|J| \geq \alpha N$ such that $\left\|\mu_{B_j} - \mathbb{E}[X]\right\| > \varepsilon$ for all $j \in J$. Hence,

$$\mathbb{P}(\mathcal{E}) \leq \mathbb{P}\left(\sum_{j=1}^N \mathbb{1}_{\{\|\mu_{B_j} - \mathbb{E}[X]\| > \varepsilon\}} > \alpha N\right) \leq \mathbb{P}(W > \alpha N)$$

where W is a binomial random variable with parameters N and $p^*(\alpha)$. Chernoff's bound (e.g., Proposition A.6.1 in [64]) implies that

$$\mathbb{P}(W > \alpha N) \leq \exp(-N\psi(\alpha, p^*(\alpha))) .$$

By definition of $p^*(\alpha)$, $\psi(\alpha, p^*(\alpha)) \geq 1$, then $\mathbb{P}(\mathcal{E}) \leq \delta$ which concludes the proof.

2.6.5 Proof of Theorem 12

Let W be a 1-net on the sphere S^{d-1} and let K_W be the polytope of $|W|$ facets associated with W. We need to show that

$$B(0,1) \subset K_W \subset 2B(0,1) .$$

By definition of the polytope K_W, the first inclusion is immediate. The second inclusion says that if x does not belong to $2B(0,1)$ then x does not belong to K_W. In other words, if $\|x\| > 2$ then there is a $w \in W$ for which $x \cdot w > 1$. Equivalently, for every unit vector $\theta \in S^{d-1}$ there is a $w \in W$ such that

$$\theta \cdot w \geq \frac{1}{2} .$$

The 1/2-cap about w is

$$\left\{ \theta \in S^{d-1} \, : \, \theta \cdot w \geq \frac{1}{2} \right\} .$$

Any θ in the 1/2-cap about w satisfies $\|\theta - w\| \leq 1$ (see Figure 2.5). Then the second

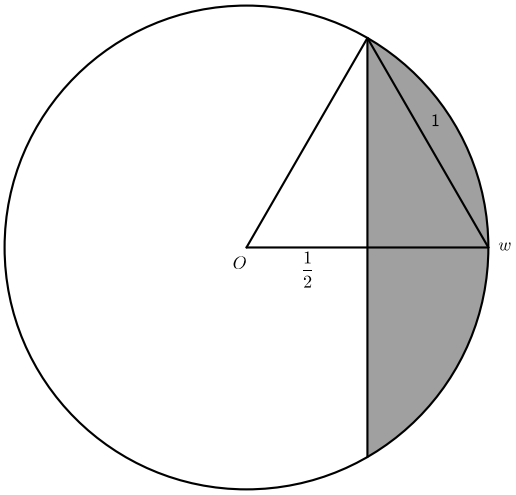


Figure 2.5: The $\frac{1}{2}$-cap about w. This figure is extracted from Ball [8].

inclusion is satisfied if for any vector $\theta \in S^{d-1}$ there exists a $w \in W$ such that θ belongs to the 1/2-cap about w. The definition of a 1-net ensures that such a w always exists.

2.6.6 Proof of Theorem 20

For any $f \in \mathcal{F}$, we define $Z_f = \frac{1}{n} \sum f(X_i)$. Then $\mathbb{E}\widehat{f} - \mathbb{E}f^* = \mathbb{E}\widehat{f} - Z_{\widehat{f}} + Z_{\widehat{f}} - Z_{f^*} + Z_{f^*} - \mathbb{E}f^*$. Note that $Z_{\widehat{f}} - Z_{f^*} \leq 0$. The process $Y_f := Z_f - \mathbb{E}f$ is centered, and, by Theorem 2,

$$\mathbb{P}\left(|Y_f - Y_{f'}| \geq u \right) \leq 2 \exp\left(- \frac{u^2 n}{d(f,f')^2} \right) .$$

A use of Corollary 2.2.5 in [64] ensures that, with probability at least $1-\delta$,

$$\mathbb{E}\widehat{f} - Z_{\widehat{f}} \leq \sup_{f,f' \in \mathcal{F}} |Y_f - Y_{f^*}| + |Z_{f^*} - \mathbb{E}f^*| \leq L\gamma_2(\mathcal{F}, d)\sqrt{\frac{\ln(2\delta^{-1})}{n}} + |Z_{f^*} - \mathbb{E}f^*| \,.$$

Another use of Theorem 2 gives

$$\mathbb{P}\left(|Z_{f^*} - \mathbb{E}f^*| \geq u\right) \leq 2\exp\left(-\frac{u^2 n}{B^2}\right) .$$

Equivalently, with probability at least $1-\delta$,

$$|Z_{f^*} - \mathbb{E}f^*| \leq B\sqrt{\frac{\ln(2\delta^{-1})}{n}} \,.$$

This concludes the proof.

Chapter 3

Empirical risk minimization with heavy tails

The purpose of this chapter is to discuss empirical risk minimization when the losses are not necessarily bounded and may have a distribution with heavy tails. In such situations usual empirical averages may fail to provide reliable estimates and empirical risk minimization may provide large excess risk. However, some robust mean estimators proposed in the literature may be used to replace empirical means. In this chapter we investigate empirical risk minimization based on a robust estimate proposed by Catoni. We develop performance bounds based on chaining arguments tailored to Catoni's mean estimator. This chapter is a joint work with Christian Brownlees and Gábor Lugosi. It is based on a paper [15] to appear in the Annals of Statistics.

Contents

3.1 Introduction

Heavy tailed data are commonly encountered in many fields of research (see, e.g., Embrechts, Klüppelberg and Mikosch [26] and Finkenstadt and Rootzén [28]). For instance, in finance, the influential work of Mandelbrot [47] and Fama [27] documented evidence of power-law behavior in asset prices in the early 1960's. When the data have heavy tails, standard statistical procedures typically perform poorly and appropriate robust alternatives are needed to carry out inference effectively. In this chapter, we propose a class of robust empirical risk minimization procedures for such data that are based on a robust estimator introduced by Catoni [20].

Empirical risk minimization is one of the basic principles of statistical learning that is routinely applied in a great variety of problems such as regression function estimation, classification, and clustering. The general model may be described as follows. Let X be a random variable taking values in some measurable space $\mathcal{X}$ and let $\mathcal{F}$ be a set of non-negative functions defined on $\mathcal{X}$. For each $f \in \mathcal{F}$, define the *risk* $m_f = \mathbb{E}f(X)$ and let $m^* = \inf_{f\in\mathcal{F}} m_f$ denote the optimal risk. In statistical learning n independent random variables $X_1, \dots, X_n$ are available, all distributed as X, and one aims at finding a function with small risk. To this end, one may define the *empirical risk minimizer*

$$f_{\mathrm{ERM}} = \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(X_i)\,,$$

where, for the simplicity of the discussion and essentially without loss of generality, we implicitly assume that the minimizer exists. If the minimum is achieved by more than one function, one may pick one of them arbitrarily.

Remark. (LOSS FUNCTIONS AND RISKS.) The main motivation and terminology may be explained by the following general prediction problem in statistical learning. Let the "training data" $(Z_1, Y_1), \dots, (Z_n, Y_n)$ be independent identically distributed pairs of random variables where the Z_i take their values in, say, $\mathbb{R}^m$ and the Y_i are real-valued. In classification problems the Y_i take discrete values. Given a new observation Z, one is interested in predicting the value of the corresponding response variable Y where the pair (Z, Y) has the same distribution as that of the (Z_i, Y_i). A predictor is a function $g : \mathbb{R}^m \to \mathbb{R}$ whose quality is measured with the help of a *loss function* $\ell : \mathbb{R}\times\mathbb{R} \to \mathbb{R}_+$. The *risk* of g is then $\mathbb{E}\ell(g(Z), Y)$. Given a class $\mathcal{G}$ of functions $g : \mathbb{R}^m \to \mathbb{R}$, empirical risk minimization chooses one that minimizes the *empirical risk* $(1/n)\sum_{i=1}^n \ell(g(Z_i), Y_i)$ over all $g \in \mathcal{G}$. In the simplified notation followed in this chapter, X_i corresponds to the pair (Z_i, Y_i), the function f represents $\ell(g(\cdot), \cdot)$, and m_f substitutes $\mathbb{E}\ell(g(Z), Y)$.

The performance of empirical risk minimization is measured by the *risk* of the selected function,

$$m_{\mathrm{ERM}} = \mathbb{E}\left[f_{\mathrm{ERM}}(X)|X_1, \dots, X_n\right]\,.$$

In particular, the main object of interest for this chapter is the *excess risk* $m_{\mathrm{ERM}} - m^*$. The performance of empirical risk minimization has been thoroughly studied and

well understood using tools of empirical process theory. In particular, the simple observation that

$$m_{\mathrm{ERM}} - m^* \leq 2 \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - m_f \right| ,$$

allows one to apply the rich theory on the suprema of empirical processes to obtain upper performance bounds. The interested reader is referred to Bartlett and Mendelson [10], Boucheron, Bousquet, and Lugosi [13], Koltchinskii [38], Massart [48], Mendelson [51], van de Geer [63] for references and recent results in this area. Essentially all of the theory of empirical minimization assumes either that the functions f are uniformly bounded or that the random variables $f(X)$ have sub-Gaussian tails for all $f \in \mathcal{F}$. For example, when all $f \in \mathcal{F}$ take their values in the interval $[0, 1]$, Dudley's [25] classical metric-entropy bound, together with standard symmetrization arguments, imply that there exists a universal constant c such that

$$\mathbb{E} m_{\mathrm{ERM}} - m^* \leq \frac{c}{\sqrt{n}} \mathbb{E} \int_0^1 \sqrt{\log N_{\mathbb{X}}(\mathcal{F}, \epsilon)} d\epsilon , \tag{3.1}$$

where for any $\epsilon > 0$, $N_{\mathbb{X}}(\mathcal{F}, \epsilon)$ is the *ϵ-covering number* of the class $\mathcal{F}$ under the empirical quadratic distance $d_{\mathbb{X}}(f, g) = \left(\frac{1}{n} \sum_{i=1}^{n} (f(X_i) - g(X_i))^2 \right)^{1/2}$, defined as the minimal cardinality N of any set $\{f_1, \ldots, f_N\} \subset \mathcal{F}$ such that for all $f \in \mathcal{F}$ there exists an $f_j \in \{f_1, \ldots, f_N\}$ with $d_{\mathbb{X}}(f, f_j) \leq \epsilon$. Of course, this is one of the most basic bounds and many important refinements have been established.

A tighter bound may be established by the so-called *generic chaining* method, see Talagrand [61]. Recall the following definition (see, e.g., [61, Definition 1.2.3]). Let T be a (pseudo) metric space. An increasing sequence $(\mathcal{A}_n)$ of partitions of T is called *admissible* if for all $n = 0, 1, 2, \ldots$, $\#\mathcal{A}_n \leq 2^{2^n}$. For any $t \in T$, denote by $A_n(t)$ the unique element of $\mathcal{A}_n$ that contains t. Let $\Delta(A)$ denote the diameter of the set $A \subset T$. Define, for $\beta = 1, 2$,

$$\gamma_\beta(T, d) = \inf_{\mathcal{A}_n} \sup_{t \in T} \sum_{n \geq 0} 2^{n/\beta} \Delta(A_n(t)) ,$$

where the infimum is taken over all admissible sequences. Then one has

$$\mathbb{E} m_{\mathrm{ERM}} - m^* \leq \frac{c}{\sqrt{n}} \mathbb{E} \gamma_2(\mathcal{F}, d_{\mathbb{X}}) , \tag{3.2}$$

for some universal constant c. This bound implies (3.1) as $\gamma_2(\mathcal{F}, d_{\mathbb{X}})$ is bounded by a constant multiple of the entropy integral $\int_0^1 \sqrt{\log N_{\mathbb{X}}(\mathcal{F}, \epsilon)} d\epsilon$ (see, e.g., [61]).

However, when the functions f are no longer uniformly bounded and the random variables $f(X)$ may have a heavy tail, empirical risk minimization may have a much poorer performance. This is simply due to the fact that empirical averages become poor estimates of expected values. Indeed, for heavy-tailed distributions, several estimators of the mean are known to outperform simple empirical averages. It is a natural idea

to define a robust version of empirical risk minimization based on minimizing such robust estimators.

In this chapter we focus on an elegant and powerful estimator proposed and analyzed by Catoni [20]. (A version of) Catoni's estimator may be defined as follows.

Introduce the non-decreasing differentiable *truncation function*

$$\phi(x) = -\mathbb{1}_{\{x<0\}} \log\left(1 - x + \frac{x^2}{2}\right) + \mathbb{1}_{\{x\geq 0\}} \log\left(1 + x + \frac{x^2}{2}\right). \tag{3.3}$$

To estimate $m_f = \mathbb{E}f(X)$ for some $f \in \mathcal{F}$, define, for all $\mu \in \mathbb{R}$,

$$\widehat{r}_f(\mu) = \frac{1}{n\alpha} \sum_{i=1}^{n} \phi(\alpha(f(X_i) - \mu))$$

where $\alpha > 0$ is a parameter of the estimator to be specified below. Catoni's estimator of m_f is defined as the unique value $\widehat{\mu}_f$ for which $\widehat{r}_f(\widehat{\mu}_f) = 0$. (Uniqueness is ensured by the strict monotonicity of $\mu \mapsto \widehat{r}_f(\mu)$). Catoni proves that for any fixed $f \in \mathcal{F}$ and $\delta \in [0, 1]$ such that $n > 2\log(1/\delta)$, under the only assumption that $\mathrm{Var}\,(f(X)) \leq v$, the estimator above with

$$\alpha = \sqrt{\frac{2\log(1/\delta)}{n\left(v + \frac{2v\log(1/\delta)}{n(1-(2/n)\log(1/\delta))}\right)}}$$

satisfies that, with probability at least $1 - 2\delta$,

$$\left|m_f - \widehat{\mu}_f\right| \leq \sqrt{\frac{2v\log(1/\delta)}{n(1 - (2/n)\log(1/\delta))}}. \tag{3.4}$$

In other words, the deviations of the estimate exhibit a sub-Gaussian behavior. The price to pay is that the estimator depends both on the upper bound v for the variance and on the prescribed confidence δ via the parameter α.

Catoni also shows that for any $n > 4(1 + \log(1/\delta))$, if $\mathrm{Var}\,(f(X)) \leq v$, the choice

$$\alpha = \sqrt{\frac{2}{nv}}$$

guarantees that, with probability at least $1 - 2\delta$,

$$\left|m_f - \widehat{\mu}_f\right| \leq (1 + \log(1/\delta)) \sqrt{\frac{v}{n}}. \tag{3.5}$$

Even though we lose the sub-Gaussian tail behavior, the estimator is independent of the required confidence level.

Given such a powerful mean estimator, it is natural to propose an empirical risk minimizer that selects a function from the class $\mathcal{F}$ that minimizes Catoni's mean estimator. Formally, define

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \widehat{\mu}_f$$

where again, for the sake of simplicity we assume that the minimizer exists. (Otherwise one may select an appropriate approximate minimizer and all arguments go through in a trivial way.)

Once again, as a first step of understanding the excess risk $m_{\widehat{f}} - m^*$, we may use the simple bound

$$m_{\widehat{f}} - m^* = \left(m_{\widehat{f}} - \widehat{\mu}_{\widehat{f}} \right) + \left(\widehat{\mu}_{\widehat{f}} - m^* \right) \leq 2 \sup_{f \in \mathcal{F}} \left| m_f - \widehat{\mu}_f \right| .$$

When $\mathcal{F}$ is a finite class of cardinality, say $|\mathcal{F}| = N$, Catoni's bound may be combined, in a straightforward way, with the union-of-events bound. Indeed, if the estimators $\widehat{\mu}_f$ are defined with parameter

$$\alpha = \sqrt{\frac{2 \log(N/\delta)}{n \left(v + \frac{2v \log(N/\delta)}{n(1-(2/n)\log(N/\delta))} \right)}} ,$$

then, with probability at least $1 - 2\delta$,

$$\sup_{f \in \mathcal{F}} \left| m_f - \widehat{\mu}_f \right| \leq \sqrt{\frac{2v \log(N/\delta)}{n(1 - (2/n)\log(N/\delta))}} .$$

Note that this bound requires that $\sup_{f \in \mathcal{F}} \operatorname{Var}(f(X)) \leq v$, that is, the variances are uniformly bounded by a *known* value v. Throughout the chapter we work with this assumption. However, this bound does not take into account the structure of the class $\mathcal{F}$ and it is useless when $\mathcal{F}$ is an infinite class. Our strategy to obtain meaningful bounds is to use *chaining* arguments. However, the extension is nontrivial and the argument becomes more involved. The main results of the chapter present performance bounds for empirical minimization of Catoni's estimator based on generic chaining.

Remark. (MEDIAN-OF-MEANS ESTIMATOR.) Catoni's estimator is not the only one with sub-Gaussian deviations for heavy-tailed distributions. Indeed, the *median-of-means* estimator, proposed by Nemirovsky and Yudin [53] (and also independently by Alon, Matias, and Szegedy [2]) has similar performance guarantees as (3.4). This estimate is obtained by dividing the data in several small blocks, calculating the sample mean within each block, and then taking the median of these means. Hsu and Sabato [32] and Minsker [52] introduce multivariate generalizations of the median-of-means estimator and use it to define and analyze certain statistical learning procedures in the presence of heavy-tailed data. The sub-Gaussian behavior is achieved under various assumptions

on the loss function. Such conditions can be avoided here. As an example, we detail applications of our results in Section 3.4 for three different examples of loss functions. An important advantage of the median-of-means estimate over Catoni's estimate is that the parameter of the estimate (i.e., the number of blocks) only depends on the confidence level δ but not on v and therefore no prior upper bound of the variance v is required to compute this estimate. Also, the median-of-means estimate is useful even when the variance is infinite and only a moment of order $1 + \epsilon$ exists for some $\epsilon > 0$ (see Bubeck, Cesa-Bianchi, and Lugosi [16]). Lerasle and Oliveira [42] consider empirical minimization of the median-of-means estimator and obtain interesting results in various statistical learning problems. However, to establish metric-entropy bounds for minimization of this mean estimate remains to be a challenge.

The rest of the chapter is organized as follows. In Section 3.2 we state and discuss the main results of the chapter. Section 3.3 is dedicated to the proofs. In Section 3.4 we describe some applications to regression under the absolute and squared losses and k-means clustering. Finally, in Section 3.5 we present some simulation results both for regression and k-means clustering. The simulation study gives empirical evidence that the proposed empirical risk minimization procedure improves performance in a significant manner in the presence of heavy-tailed data. Some of the more technical arguments are relegated to the Appendix.

3.2 Main results

The bounds we establish for the excess risk depend on the geometric structure of the class $\mathcal{F}$ under different distances. The $L_2(P)$ distance is defined, for $f, f' \in \mathcal{F}$, by

$$d(f, f') = \left(\mathbb{E}\left[(f(X) - f'(X))^2\right]\right)^{1/2}$$

and the L_∞ distance is

$$D(f, f') = \sup_{x \in \mathcal{X}} \left|f(x) - f'(x)\right| .$$

We also work with the (random) empirical quadratic distance

$$d_{\mathbb{X}}(f, f') = \left(\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - f'(X_i))^2\right)^{1/2} .$$

Denote by f^* a function with minimal expectation

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} m_f .$$

Next we present two results that bound the excess risk $m_{\widehat{f}} - m_{f^*}$ of the minimizer $\widehat{f}$ of Catoni's risk estimate in terms of metric properties of the class $\mathcal{F}$. The first result involves a combination of terms involving the γ_2 and γ_1 functionals under the metrics d and D while the second is in terms of quantiles of γ_2 under the empirical metric $d_{\mathbb{X}}$.

Theorem 24. *Let $\mathcal{F}$ be a class of non-negative functions defined on a set $\mathcal{X}$ and let $X, X_1, \dots, X_n$ be i.i.d. random variables taking values in $\mathcal{X}$. Assume that there exists $v > 0$ such that $\sup_{f\in\mathcal{F}} Var(f(X)) \leq v$. Let $\delta \in (0, 1/3)$. Suppose that $\widehat{f}$ is selected from $\mathcal{F}$ by minimizing Catoni's mean estimator with parameter α. Then there exists a universal constant L such that, under the condition*

$$6\left(\alpha v + \frac{2\log(\delta^{-1})}{n\alpha}\right) + L\log(2\delta^{-1})\left(\frac{\gamma_2(\mathcal{F}, d)}{\sqrt{n}} + \frac{\gamma_1(\mathcal{F}, D)}{n}\right) \leq \frac{1}{\alpha},$$

with probability at least $1 - 3\delta$, the risk of $\widehat{f}$ satisfies

$$m_{\widehat{f}} - m_{f^*} \leq 6\left(\alpha v + \frac{2\log(\delta^{-1})}{n\alpha}\right) + L\log(2\delta^{-1})\left(\frac{\gamma_2(\mathcal{F}, d)}{\sqrt{n}} + \frac{\gamma_1(\mathcal{F}, D)}{n}\right).$$

Theorem 25. *Assume the hypotheses of Theorem 24. We denote by $diam_d(\mathcal{F})$ the diameter of the class $\mathcal{F}$ under the distance d. Set Γ_δ such that $\mathbb{P}\left(\gamma_2(\mathcal{F}, d_{\mathbb{X}}) > \Gamma_\delta\right) \leq \frac{\delta}{8}$. Then there exists a universal constant K such that, under the condition*

$$6\left(\alpha v + \frac{2\log(\delta^{-1})}{n\alpha}\right) + K\max(\Gamma_\delta, diam_d(\mathcal{F}))\sqrt{\frac{\log(\frac{8}{\delta})}{n}} \leq \frac{1}{\alpha},$$

with probability at least $1 - 3\delta$, the risk of $\widehat{f}$ satisfies

$$m_{\widehat{f}} - m_{f^*} \leq 6\left(\alpha v + \frac{2\log(\delta^{-1})}{n\alpha}\right) + K\max(\Gamma_\delta, diam_d(\mathcal{F}))\sqrt{\frac{\log(\frac{8}{\delta})}{n}}.$$

In both theorems above, the choice of α only influences the term $\alpha v + 2\log(\delta^{-1})/(n\alpha)$. By taking $\alpha = \sqrt{2\log(\delta^{-1})/(nv)}$, this term equals

$$2\sqrt{\frac{2v\log(\delta^{-1})}{n}}.$$

For example, in that case, the condition in Theorem 24 reduces to

$$12\sqrt{\frac{2v\log(\delta^{-1})}{n}} + L\log(\delta^{-1})\left(\frac{\gamma_2(\mathcal{F}, d)}{\sqrt{n}} + \frac{\gamma_1(\mathcal{F}, D)}{n}\right) \leq \sqrt{\frac{nv}{2\log(\delta^{-1})}}.$$

This holds for sufficiently large values of n. This choice has the disadvantage that the estimator depends on the confidence level (i.e., on the value of δ). By taking $\alpha = \sqrt{2/(nv)}$, independently of δ, one obtains the slightly worse term

$$\sqrt{\frac{2v}{n}}(1 + \log(\delta^{-1})).$$

Observe that the main term in the second part of the bound of Theorem 24 is

$$L \log(\delta^{-1}) \frac{\gamma_2(\mathcal{F}, d)}{\sqrt{n}}$$

which is comparable to the bound (3.2) obtained under the strong condition of $f(X)$ being uniformly bounded. All other terms are of smaller order. Note that this part of the bound depends on the "weak" distribution-dependent $L_2(P)$ metric d. The quantity $\gamma_1(\mathcal{F}, D) \geq \gamma_2(\mathcal{F}, d)$ also enters the bound of Theorem 24 though only multiplied by $1/n$. The presence of this term requires that $\mathcal{F}$ be bounded in the L_∞ distance D which limits the usefulness of the bound. In Section 3.4 we illustrate the bounds on two applications to regression and k-means clustering. In these applications, in spite of the presence of heavy tails, the covering numbers under the distance D may be bounded in a meaningful way. Note that no such bound can hold for "ordinary" empirical risk minimization that minimizes the usual empirical means $(1/n)\sum_{i=1}^n f(X_i)$ because of the poor performance of empirical averages in the presence of heavy tails.

The main merit of the bound of Theorem 25 is that it does not require that the class $\mathcal{F}$ has a finite diameter under the supremum norm. Instead, the quantiles of $\gamma_2(\mathcal{F}, d_{\mathbb{X}})$ enter the picture. In Section 3.4 we show through the example of L_2 regression how these quantiles may be estimated.

3.3 Proofs

The proofs of Theorems 24 and 25 are based on showing that the excess risk can be bounded as soon as the supremum of the empirical process $\{X_f(\mu) : f \in \mathcal{F}\}$ is bounded for any fixed $\mu \in \mathbb{R}$, where for any $f \in \mathcal{F}$ and $\mu \in \mathbb{R}$, we define $X_f(\mu) = \widehat{r}_f(\mu) - \bar{r}_f(\mu)$ with

$$\bar{r}_f(\mu) = \frac{1}{\alpha}\mathbb{E}\left[\phi(\alpha(f(X) - \mu))\right]$$

and

$$\widehat{r}_f(\mu) = \frac{1}{n\alpha}\sum_{i=1}^n \phi(\alpha(f(X_i) - \mu)) \,.$$

The two theorems differ in the way the supremum of this empirical process is bounded.

Let $A_\alpha(\delta) = \alpha v + 2\log(\delta^{-1})/(n\alpha)$.

Once again, we may assume, essentially without loss of generality, that the minimum exists. In case of multiple minimizers we may choose one arbitrarily. The main result in [20] states that for any $\delta > 0$ such that $\alpha^2 v + 2\log(\delta^{-1})/n \leq 1$, with probability at least $1 - 2\delta$,

$$\left|\widehat{\mu}_{f^*} - m_{f^*}\right| \leq A_\alpha(\delta) \,. \tag{3.6}$$

Let $\Omega_{f^*}(\delta)$ be the event on which inequality (3.6) holds. By definition, $\mathbb{P}\left(\Omega_{f^*}(\delta)\right) \geq 1 - 2\delta$.

3.3.1 A deterministic version of $\widehat{\mu}_f$

We begin with a variant of the argument of Catoni [20]. It involves a deterministic version $\overline{\mu}_f$ of the estimator defined, for each $f \in \mathcal{F}$, as the unique solution of the equation $\overline{r}_f(\mu) = 0$.

In Lemma 26 below we show that $\overline{\mu}_f$ is in a small (deterministic) interval centered at m_f. For any $f \in \mathcal{F}$, $\mu \in \mathbb{R}$, and $\varepsilon \geq 0$, define

$$B_f^+(\mu, \varepsilon) = (m_f - \mu) + \frac{\alpha}{2}(m_f - \mu)^2 + \frac{\alpha}{2}v + \varepsilon \,,$$
$$B_f^-(\mu, \varepsilon) = (m_f - \mu) - \frac{\alpha}{2}(m_f - \mu)^2 - \frac{\alpha}{2}v - \varepsilon$$

and let

$$\mu_f^+(\varepsilon) = m_f + \alpha v + 2\varepsilon \,, \qquad \mu_f^-(\varepsilon) = m_f - \alpha v - 2\varepsilon \,.$$

As a function of μ, $B_f^+(\mu, \varepsilon)$ is a quadratic polynomial such that $\mu_f^+(\varepsilon)$ is an upper bound of the smallest root of $B_f^+(\mu, \varepsilon)$. Similarly, $\mu_f^-(\varepsilon)$ is a lower bound of the largest root of $B_f^-(\mu, \varepsilon)$. Implicitly we assumed that these roots always exist. This is not always the case but a simple condition on α guarantees that these roots exists. In particular, $1 - \alpha^2 v - 2\alpha\varepsilon \geq 0$ guarantees that $B_f^+(\mu, \varepsilon) = 0$ and $B_f^-(\mu, \varepsilon) = 0$ have at least one solution. This condition will always be satisfied by our choice of ϵ and α.

Still following the ideas of [20], the next lemma bounds $\overline{r}_f(\mu)$ by the quadratic polynomials B^+ and B^-. The lemma will help us compare the zero of $\overline{r}_f(\mu)$ to the zeros of these quadratic functions.

Lemma 26. *For any fixed $f \in \mathcal{F}$ and $\mu \in \mathbb{R}$,*

$$B_f^-(\mu, 0) \leq \overline{r}_f(\mu) \leq B_f^+(\mu, 0) \,, \tag{3.7}$$

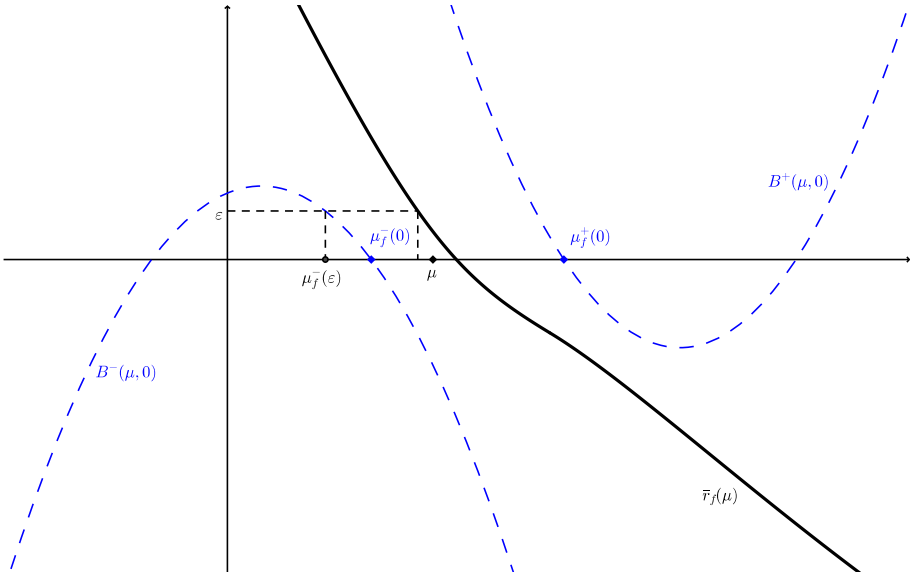
and therefore $m_f - \alpha v \leq \overline{\mu}_f \leq m_f + \alpha v$. In particular,

$$B_{\widehat{f}}^-(\mu, 0) \leq \overline{r}_{\widehat{f}}(\mu) \leq B_{\widehat{f}}^+(\mu, 0) \,.$$

For any μ and ε such that $\overline{r}_{\widehat{f}}(\mu) \leq \varepsilon$, if $1 - \alpha^2 v - 2\alpha\varepsilon \geq 0$, then

$$m_{\widehat{f}} \leq \mu + \alpha v + 2\varepsilon \,. \tag{3.8}$$

Figure 3.1: Representation of $\bar{r}_f(\mu)$ and the quadratic functions $B_f^-(\mu, 0)$ and $B_f^+(\mu, 0)$.



$\bar{r}_f(\mu)$ is squeezed between $B_f^-(\mu, 0)$ and $B_f^+(\mu, 0)$. In particular at $\mu_f^+(0)$ (resp. $\mu_f^-(0)$), $\bar{r}_f(\mu)$ is non-positive (resp. non-negative). Any μ such that $\bar{r}_f(\mu) \leq \varepsilon$ is above $\mu_f^-(\varepsilon)$.

Proof. Writing Y for $\alpha(f(X) - \mu)$ and using the fact that $\phi(x) \leq \log(1 + x + x^2/2)$ for all $x \in \mathbb{R}$,

$$\begin{aligned}
\exp\left(\alpha \bar{r}_f(\mu)\right) &\leq \exp\left(\mathbb{E}\left[\log(1 + Y + \frac{Y^2}{2})\right]\right) \\
&\leq \mathbb{E}\left[1 + Y + \frac{Y^2}{2}\right] \\
&\leq 1 + \alpha(m_f - \mu) + \frac{\alpha^2}{2}\left[v + (m_f - \mu)^2\right] \\
&\leq \exp\left(\alpha B_f^+(\mu, 0)\right) .
\end{aligned}$$

Thus, we have $\bar{r}_f(\mu) - B_f^+(\mu, 0) \leq 0$ (see Figure 3.3.1). Since this last inequality is true for any f, $\sup_f(\bar{r}_f(\mu) - B_f^+(\mu, 0)) \leq 0$ and the second inequality of (3.7) is proved. The second statement of the lemma may be proved by a similar argument.

If $\bar{r}_{\tilde{f}}(\mu) \leq \varepsilon$ then $B_{\tilde{f}}^-(\mu, 0) \leq \varepsilon$ which is equivalent to $B_{\tilde{f}}^-(\mu, \varepsilon) \leq 0$. If $1 - \alpha^2 v - 2\alpha\varepsilon \geq 0$ then a solution of $B_{\tilde{f}}^-(\mu, \varepsilon) = 0$ exists and since $\bar{r}_{\tilde{f}}(\mu)$ is a non-increasing function, μ is above the largest of these two solutions. This implies $\mu_{\tilde{f}}^-(\varepsilon) \leq \mu$ which gives inequality (3.8) (see Figure 3.3.1). □

Inequality (3.8) is the key tool to ensure that the risk $m_{\widehat{f}}$ of the minimizer $\widehat{f}$ can be upper bounded as soon as $\bar{r}_{\widehat{f}}$ is. It remains to find the smallest μ and ε such that $\bar{r}_f(\mu)$ is bounded uniformly on $\mathcal{F}$.

3.3.2 Bounding the excess risk in terms of the supremum of an empirical process

The key to all proofs is that we link the excess risk to the supremum of the empirical process $X_f(\mu) = \widehat{r}_f(\mu) - \bar{r}_f(\mu)$ as f ranges through $\mathcal{F}$ for a suitably chosen value of μ. For fixed $\mu \in \mathbb{R}$ and $\delta \in (0,1)$, define the $1-\delta$ quantile of $\sup_{f\in\mathcal{F}} |X_f(\mu) - X_{f^*}(\mu)|$ by $Q(\mu,\delta)$, that is, the infimum of all positive numbers q such that

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}} |X_f(\mu) - X_{f^*}(\mu)| \le q\right\} \ge 1-\delta\,.$$

First we need a few simple facts summarized in the next lemma.

Lemma 27. *Let $\mu_0 = m_{f^*} + A_\alpha(\delta)$. Then on the event $\Omega_{f^*}(\delta)$, the following inequalities hold:*

1. $\widehat{r}_{\widehat{f}}(\mu_0) \le 0$
2. $\bar{r}_{f^*}(\mu_0) \le 0$
3. $-\widehat{r}_{f^*}(\mu_0) \le 2A_\alpha(\delta)$

Proof. We prove each inequality separately.

1. First note that on $\Omega_{f^*}(\delta)$ inequality (3.6) holds and we have $\widehat{\mu}_{\widehat{f}} \le \widehat{\mu}_{f^*} \le \mu_0$. Since $\widehat{r}_{\widehat{f}}$ is a non-increasing function of μ, $\widehat{r}_{\widehat{f}}(\mu_0) \le \widehat{r}_{\widehat{f}}(\widehat{\mu}_{\widehat{f}}) = 0$.

2. By (3.7), $\overline{\mu}_{f^*} \le m_{f^*} + \alpha v \le m_{f^*} + \alpha v + 2\log(\delta^{-1})/(n\alpha) = \mu_0$. Since $\bar{r}_{f^*}$ is a non-increasing function, $\bar{r}_{f^*}(\mu_0) \le \bar{r}_{f^*}(\overline{\mu}_{f^*}) = 0$.

3. $\widehat{r}_{f^*}$ is a 1-Lipschitz function and therefore

$$\begin{aligned} |\widehat{r}_{f^*}(\mu_0)| = |\widehat{r}_{f^*}(\widehat{\mu}_{f^*}) - \widehat{r}_{f^*}(\mu_0)| &\le |\widehat{\mu}_{f^*} - \mu_0| \\ &\le |\widehat{\mu}_{f^*} - m_{f^*}| + |m_{f^*} - \mu_0| \\ &\le 2A_\alpha(\delta) \end{aligned}$$

   which gives $-\widehat{r}_{f^*}(\mu_0) \le 2A_\alpha(\delta)$.

□

We will use Lemma 26 with μ_0 introduced in Lemma 27. Recall that $\mathbb{P}\left(\Omega_{f^*}(\delta)\right) \geq 1 - 2\delta$.

With the notation introduced above, we see that with probability at least $1 - \delta$,

$$\begin{aligned}
\bar{r}_{\widehat{f}}(\mu_0) &\leq \widehat{r}_{\widehat{f}}(\mu_0) + \bar{r}_{f^*}(\mu_0) - \widehat{r}_{f^*}(\mu_0) + \left|\bar{r}_{\widehat{f}}(\mu_0) - \widehat{r}_{\widehat{f}}(\mu_0) - \bar{r}_{f^*}(\mu_0) + \widehat{r}_{f^*}(\mu_0)\right| \\
&\leq \widehat{r}_{\widehat{f}}(\mu_0) + \bar{r}_{f^*}(\mu_0) - \widehat{r}_{f^*}(\mu_0) + \sup_{f\in\mathcal{F}} \left|\bar{r}_f(\mu_0) - \widehat{r}_f(\mu_0) - \bar{r}_{f^*}(\mu_0) + \widehat{r}_{f^*}(\mu_0)\right| \\
&\leq \widehat{r}_{\widehat{f}}(\mu_0) + \bar{r}_{f^*}(\mu_0) - \widehat{r}_{f^*}(\mu_0) + Q(\mu_0, \delta) \, .
\end{aligned}$$

This inequality, together with Lemma 27, implies that, with probability at least $1 - 3\delta$,

$$\bar{r}_{\widehat{f}}(\mu_0) \leq 2A_\alpha(\delta) + Q(\mu_0, \delta) \, .$$

Now using Lemma 26 with $\varepsilon = 2A_\alpha(\delta) + Q(\mu_0, \delta)$ and under the condition $1 - \alpha^2 v - 4\alpha A_\alpha(\delta) - 2\alpha Q(\mu_0, \delta) \geq 0$, we have

$$\begin{aligned}
m_{\widehat{f}} - m_{f^*} &\leq \alpha v + 5A_\alpha(\delta) + 2Q(\mu_0, \delta) \\
&\leq 6\left(\alpha v + \frac{2\log(\delta^{-1})}{n\alpha}\right) + 2Q(\mu_0, \delta) \, , \qquad (3.9)
\end{aligned}$$

with probability at least $1 - 3\delta$. The condition $1 - \alpha^2 v - 4\alpha A_\alpha(\delta) - 2\alpha Q(\mu_0, \delta) \geq 0$ is satisfied whenever

$$6\left(\alpha v + \frac{2\log(\delta^{-1})}{n\alpha}\right) + 2Q(\mu_0, \delta) \leq \frac{1}{\alpha}$$

holds.

3.3.3 Bounding the supremum of the empirical process

Theorems 24 and 25 both follow from (3.9) by two different ways of bounding the quantile $Q(\mu, \delta)$ of $\sup_{f\in\mathcal{F}} |X_f(\mu) - X_{f^*}(\mu)|$. Here we present these two inequalities. Both of them use basic results of "generic chaining", see Talagrand [61]. Theorem 24 follows from (3.9) and the next inequality:

Proposition 28. *Let $\mu \in \mathbb{R}$ and $\alpha > 0$. There exists a universal constant L such that for any $\delta \in (0, 1)$,*

$$Q(\mu, \delta) \leq L\log(2\delta^{-1})\left(\frac{\gamma_2(\mathcal{F}, d)}{\sqrt{n}} + \frac{\gamma_1(\mathcal{F}, D)}{n}\right) .$$

The proof is an immediate consequence of Theorem 35 and (3.14) in the Appendix and the following lemma.

Lemma 29. *For any $\mu \in \mathbb{R}$, $\alpha > 0$, $f, f' \in \mathcal{F}$, and $t > 0$,*

$$\mathbb{P}\left(|X_f(\mu) - X_{f'}(\mu)| > t\right) \leq 2\exp\left(-\frac{nt^2}{2(d(f,f')^2 + \frac{2D(f,f')t}{3})}\right)$$

where the distances d, D are defined at the beginning of Section 3.2.

Proof. Observe that $n(X_f(\mu) - X_{f'}(\mu))$ is the sum of the independent zero-mean random variables

$$\begin{aligned} C_i(f,f') &= \frac{1}{\alpha}\phi(\alpha(f(X_i)-\mu)) - \frac{1}{\alpha}\phi(\alpha(f'(X_i)-\mu)) \\ &\quad - \left[\frac{1}{\alpha}\mathbb{E}\left[\phi(\alpha(f(X)-\mu))\right] - \frac{1}{\alpha}\mathbb{E}\left[\phi(\alpha(f'(X)-\mu))\right]\right] . \end{aligned}$$

Note that since the truncation function ϕ is 1-Lipschitz, we have $C_i(f,f') \leq 2D(f,f')$. Also,

$$\sum_{i=1}^n \mathbb{E}[C_i(f,f')^2] \leq \sum_{i=1}^n \mathbb{E}\left[((f(X_i)-\mu) - (f'(X_i)-\mu))^2\right] = nd(f,f')^2$$

The lemma follows from Bernstein's inequality (see, e.g., [14, Equation (2.10)]). □

Similarly, Theorem 25 is implied by (3.9) and the following. Recall the notation of Theorem 25.

Theorem 30. *Let $\mu \in \mathbb{R}$, $\alpha > 0$, and $\delta \in (0, 1/3)$. There exists a universal constant K such that*

$$Q(\mu,\delta) \leq K\max(\Gamma_\delta, diam_d(\mathcal{F}))\sqrt{\frac{\log(\frac{8}{\delta})}{n}} .$$

Proof. Assume $\Gamma_\delta \geq \mathrm{diam}_d(\mathcal{F})$. The proof is based on a standard symmetrization argument. Let $(X_1', \ldots, X_n')$ be independent copies of $(X_1, \ldots, X_n)$ and define

$$Z_i(f) = \frac{1}{n\alpha}\phi(\alpha(f(X_i)-\mu)) - \frac{1}{n\alpha}\phi(\alpha(f(X_i')-\mu)) .$$

Introduce also independent Rademacher random variables $(\varepsilon_1, \ldots, \varepsilon_n)$. For any $f \in \mathcal{F}$, denote by $Z(f) = \sum_{i=1}^n \varepsilon_i Z_i(f)$. Then by Hoeffding's inequality, for all $f, g \in \mathcal{F}$ and for every $t > 0$,

$$\mathbb{P}_{(\varepsilon_1,\ldots,\varepsilon_n)}\left(|Z(f) - Z(g)| > t\right) \leq 2\exp\left(-\frac{t^2}{2d_{\mathbb{X},\mathbb{X}'}(f,g)^2}\right) \tag{3.10}$$

where $\mathbb{P}_{(\varepsilon_1,\ldots,\varepsilon_n)}$ denotes probability with respect to the Rademacher variables only (i.e., conditional on the X_i and X_i') and $d_{\mathbb{X},\mathbb{X}'}(f,g) = \sqrt{\sum_{i=1}^n (Z_i(f) - Z_i(g))^2}$ is a random

distance. Using (3.16) in the Appendix with distance $d_{\mathbb{X},\mathbb{X}'}$ and (3.10), we get that, for all $\lambda > 0$,

$$\mathbb{E}_{(\varepsilon_1,\dots,\varepsilon_n)}\left[\exp\left(\lambda \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i[Z_i(f) - Z_i(f^*)]\right|\right)\right] \le 2\exp\left(\lambda^2 L^2 \gamma_2(\mathcal{F}, d_{\mathbb{X},\mathbb{X}'})^2/4\right) \tag{3.11}$$

where L is a universal constant from Proposition 37. Observe that since $x \mapsto \phi(x)$ is Lipschitz with constant 1,

$$\begin{aligned} & d_{\mathbb{X},\mathbb{X}'}(f,g) \\ &= \left(\frac{1}{n^2\alpha^2}\sum_{i=1}^n \Big(\phi\Big(\alpha(f(X_i)-\mu)\Big) - \phi\Big(\alpha(f(X_i')-\mu)\Big)\right. \\ &\quad \left. -\phi\Big(\alpha(g(X_i)-\mu)\Big) + \phi\Big(\alpha(g(X_i')-\mu)\Big)\Big)^2\right)^{1/2} \\ &\le \frac{1}{\sqrt{n}}\left(\frac{1}{n}\sum_{i=1}^n \Big(f(X_i)-g(X_i)\Big)^2\right)^{1/2} + \frac{1}{\sqrt{n}}\left(\frac{1}{n}\sum_{i=1}^n \Big(f(X_i')-g(X_i')\Big)^2\right)^{1/2}. \end{aligned}$$

This implies

$$\gamma_2(\mathcal{F}, d_{\mathbb{X},\mathbb{X}'}) \le \frac{1}{\sqrt{n}}\left(\gamma_2(\mathcal{F}, d_{\mathbb{X}}) + \gamma_2(\mathcal{F}, d_{\mathbb{X}'})\right).$$

Combining this with (3.11), we obtain

$$\begin{aligned} & \mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|Z(f) - Z(f^*)\right| \ge t\right) \\ &\le \mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|Z(f) - Z(f^*)\right| \ge t \;\middle|\; \gamma_2(\mathcal{F}, d_{\mathbb{X}}) \le \Gamma_\delta \;\&\; \gamma_2(\mathcal{F}, d_{\mathbb{X}'}) \le \Gamma_\delta\right) + 2\mathbb{P}\left(\gamma_2(\mathcal{F}, d_{\mathbb{X}}) > \Gamma_\delta\right) \\ &\le \mathbb{E}_{\mathbb{X},\mathbb{X}'}\left[\mathbb{E}_{(\varepsilon_1,\dots,\varepsilon_n)}\left[e^{\lambda \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i[Z_i(f)-Z_i(f^*)]\right|}\right] \;\middle|\; \gamma_2(\mathcal{F}, d_{\mathbb{X}}) \le \Gamma_\delta \;\&\; \gamma_2(\mathcal{F}, d_{\mathbb{X}'}) \le \Gamma_\delta\right] e^{-\lambda t} \\ &\quad + \frac{\delta}{4} \quad \text{(by the definition of } \Gamma_\delta) \\ &\le 2\exp\left(\frac{\lambda^2 L^2}{n}\Gamma_\delta^2 - \lambda t\right) + \frac{\delta}{4}. \end{aligned}$$

Optimization in λ with $t = 2L\Gamma_\delta\sqrt{\log(8/\delta)/n}$ gives

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|Z(f) - Z(f^*)\right| \ge t\right) \le \frac{\delta}{2}.$$

A standard symmetrization inequality of tail probabilities of empirical processes (see, e.g., [63, Lemma 3.3]) guarantees that

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|X_f(\mu) - X_{f^*}(\mu)\right| \ge 2t\right) \le 2\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|Z(f) - Z(f^*)\right| \ge t\right)$$

as long as for any $f \in \mathcal{F}$, $\mathbb{P}\left(\left|X_f(\mu) - X_{f^*}(\mu)\right| \geq t\right) \leq \frac{1}{2}$. Recall that $X_f(\mu) - X_{f^*}(\mu)$ is a zero-mean random variable. Then by Chebyshev's inequality it suffices to have $t \geq \sqrt{2}\text{diam}_d(\mathcal{F})/\sqrt{n}$. Indeed,

$$\begin{aligned}\frac{\text{Var}\left(X_f(\mu) - X_{f'}(\mu)\right)}{t^2} &\leq \frac{\text{Var}\left(\frac{1}{\alpha}\phi(\alpha(f(X)-\mu)) - \frac{1}{\alpha}\phi(\alpha(f^*(X)-\mu))\right)}{nt^2} \\ &\leq \frac{\mathbb{E}\left[(f(X) - f^*(X))^2\right]}{nt^2} \\ &\leq \frac{\text{diam}_d(\mathcal{F})^2}{nt^2}\,.\end{aligned}$$

Without loss of generality, we can assume $L \geq 1$. Since for any choice of $\delta < \frac{1}{3}$, $\sqrt{\log(\frac{8}{\delta})} > \sqrt{2}$ we have $L\Gamma_\delta\sqrt{\log(\frac{8}{\delta})} \geq \text{diam}_d(\mathcal{F})\sqrt{2}$. Thus

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|X_f(\mu) - X_{f^*}(\mu)\right| \geq 2L\Gamma_\delta\sqrt{\frac{\log(\frac{8}{\delta})}{n}}\right) \leq \delta$$

as desired. Now, if $\Gamma_\delta < \text{diam}_d(\mathcal{F})$, $\mathbb{P}\left(\gamma_2(\mathcal{F}, d_{\mathbb{X}}) > \text{diam}_d(\mathcal{F})\right) \leq \frac{\delta}{8}$ and the same argument holds for $\text{diam}_d(\mathcal{F})$ instead of Γ_δ. This concludes the proof. □

3.4 Applications

In this section we describe two applications of Theorems 24 and 25 to simple statistical learning problems. The first is a regression estimation problem in which we distinguish between L_1 and L_2 risks and the second is k-means clustering.

3.4.1 Empirical risk minimization for regression

L_1 regression

Let $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ be independent identically distributed random variables taking values in $\mathcal{Z} \times \mathbb{R}$ where $\mathcal{Z}$ is a bounded subset of (say) $\mathbb{R}^m$. Suppose $\mathcal{G}$ is a class of functions $\mathcal{Z} \to \mathbb{R}$ bounded in the L_∞ norm, that is, $\sup_{g\in\mathcal{G}}\sup_{z\in\mathcal{Z}}|g(z)| < \infty$. We denote by Δ the diameter of $\mathcal{G}$ under the distance induced by this norm. First we consider the setup when the *risk* of each $g \in \mathcal{G}$ is defined by the L_1 loss

$$R(g) = \mathbb{E}|g(Z) - Y|$$

where the pair (Z, Y) has the same distribution of the (Z_i, Y_i) and is independent of them. Let $g^* = \text{argmin}_{g\in\mathcal{G}} R(g)$ be a minimizer of the risk (which, without loss of generality, is assumed to exist). The statistical learning problem we consider here consists of choosing a function $\widehat{g}$ from the class $\mathcal{G}$ that has a risk $R(\widehat{g})$ not much larger than $R(g^*)$.

The standard procedure is to pick $\widehat{g}$ by minimizing the empirical risk $(1/n)\sum_{i=1}|g(Z_i)-Y_i|$ over $g \in \mathcal{G}$. However, if the response variable Y is unbounded and may have a heavy tail, ordinary empirical risk minimization may fail to provide a good predictor of Y as the empirical risk is an unreliable estimate of the true risk.

Here we propose choosing $\widehat{g}$ by minimizing Catoni's estimate. To this end, we only need to assume that the second moment of Y is bounded by a known constant. More precisely, assume that $\mathbb{E}Y^2 \le \sigma^2$ for some $\sigma > 0$. Then $\sup_{g\in\mathcal{G}} \mathrm{Var}\,(|g(Z)-Y|) \le 2\sigma^2 + 2\sup_{g\in\mathcal{G}}\sup_{z\in\mathcal{Z}}|g(z)|^2 =: v$ is a known and finite constant.

Now for all $g \in \mathcal{G}$ and $\mu \in \mathbb{R}$, define

$$\widehat{r}_g(\mu) = \frac{1}{n\alpha}\sum_{i=1}^{n}\phi(\alpha(|g(X_i)-Y_i|-\mu))$$

where ϕ is the truncation function defined in (3.3). Define $\widehat{R}(g)$ as the unique value for which $\widehat{r}_g(\widehat{R}(g)) = 0$. The empirical risk minimizer based on Catoni's risk estimate is then

$$\widehat{g} = \operatorname*{argmin}_{g\in\mathcal{G}} \widehat{R}(g)\,.$$

By Theorem 24, the performance of $\widehat{g}$ may be bounded in terms of covering numbers of the class of functions $\mathcal{F} = \{f(z,y) = |g(z)-y| : g \in \mathcal{G}\}$ based on the distance

$$D(f,f') = \sup_{z\in\mathcal{Z},y\in\mathbb{R}}\left||g(z)-y|-|g'(z)-y|\right| \le \sup_{z\in\mathcal{Z}}\left|g(z)-g'(z)\right|\,.$$

Thus, the covering numbers of $\mathcal{F}$ under the distance D may be bounded in terms of the covering numbers of $\mathcal{G}$ under the L_∞ distance. Denoting by $N_d(A,\epsilon)$ the ϵ-covering number of a set A under the metric d, we obtain the following:

Corollary 31. *Consider the setup described above. We assume* $\int_0^\Delta \log N_\infty(\mathcal{G},\epsilon)d\epsilon < \infty$. *Let* $n \in \mathbb{N}$, $\delta \in (0,1/3)$ *and* $\alpha = \sqrt{2\log(\delta^{-1})/(nv)}$. *There exists an integer* N_0 *and a universal constant* C *such that, for all* $n \ge N_0$, *with probability at least* $1-3\delta$,

$$R(\widehat{g}) - R(g^*) \le 12\sqrt{\frac{2v\log(\delta^{-1})}{n}} + C\log(2\delta^{-1})\left(\frac{1}{\sqrt{n}}\int_0^\Delta \sqrt{\log N_d(\mathcal{G},\epsilon)}d\epsilon + O\left(\frac{1}{n}\right)\right)\,.$$

Proof. Clearly, if two distances d_1 and d_2 satisfy $d_1 \le d_2$, then $\gamma_1(\mathcal{F},d_1) \le \gamma_1(\mathcal{F},d_2)$. Thus, $\gamma_1(\mathcal{F},D) \le \gamma_1(\mathcal{G},\|.\|_\infty) \le L\int_0^\Delta \log N_\infty(\mathcal{G},\epsilon)d\epsilon < \infty$ (see (3.15)) and $\gamma_1(\mathcal{F},D)/n = O(1/n)$. The condition

$$12\sqrt{\frac{2v\log(\delta^{-1})}{n}} + C\log(2\delta^{-1})\left(\frac{1}{\sqrt{n}}\int_0^\Delta \sqrt{\log N_d(\mathcal{G},\epsilon)}d\epsilon + O\left(\frac{1}{n}\right)\right) \le \sqrt{\frac{nv}{2\log(\delta^{-1})}}$$

is satisfied for sufficiently large n. Apply Theorem 24. □

Note that the bound essentially has the same form as (3.1) but to apply (3.1) it is crucial that the response variable Y is bounded or at least has sub-Gaussian tails. We get this under the weak assumption that Y has a bounded second moment (with a known upper bound). The price we pay is that covering numbers under the distance $d_{\mathbb{X}}$ are now replaced by covering numbers under the supremum norm.

L_2 regression

Here we consider the same setup as in Section 3.4.1 but now the risk is measured by the L_2 loss. The *risk* of each $g \in \mathcal{G}$ is defined by the L_2 loss

$$R(g) = \mathbb{E}(g(Z) - Y)^2 .$$

Note that Theorem 1 is useless here as the difference $|R(g) - R(g')|$ is not bounded by the L_∞ distance of g and g' anymore and the covering numbers of $\mathcal{F}$ under the metric D are infinite. However, Theorem 25 gives meaningful bounds. Let $g^* = \operatorname{argmin}_{g\in\mathcal{G}} R(g)$ and again we choose $\widehat{g}$ by minimizing Catoni's estimate.

Here we need to assume that $\mathbb{E}Y^4 \le \sigma^2$ for some $\sigma > 0$. Then $\sup_{g\in\mathcal{G}} \operatorname{Var}\left((g(Z) - Y)^2\right) \le 8\sigma^2 + 8\sup_{g\in\mathcal{G}} \sup_{z\in\mathcal{Z}} |g(z)|^4 =: v$ is a known and finite constant.

By Theorem 25, the performance of $\widehat{g}$ may be bounded in terms of covering numbers of the class of functions $\mathcal{F} = \{f(z, y) = (g(z) - y)^2 : g \in \mathcal{G}\}$ based on the distance

$$d_{\mathbb{X}}(f, f') = \left(\frac{1}{n}\sum_{i=1}^{n}\left((g(Z_i) - Y_i)^2 - (g'(Z_i) - Y_i)^2\right)^2\right)^{1/2}$$

Note that

$$\begin{aligned}\left|(g(Z_i) - Y_i)^2 - (g'(Z_i) - Y_i)^2\right| &= |g(Z_i) - g'(Z_i)||2Y_i - g(Z_i) - g'(Z_i)| \\ &\le 2|g(Z_i) - g'(Z_i)|(|Y_i| + \Delta) \\ &\le 2d_\infty(g, g')(|Y_i| + \Delta) ,\end{aligned}$$

and therefore

$$\begin{aligned}d_{\mathbb{X}}(f, f') &\le 2d_\infty(g, g')\sqrt{\frac{1}{n}\sum_{i=1}^{n}(|Y_i| + \Delta)^2} \\ &\le 2\sqrt{2}d_\infty(g, g')\sqrt{\Delta^2 + \frac{1}{n}\sum_{i=1}^{n}Y_i^2} .\end{aligned}$$

By Chebyshev's inequality,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}Y_i^2 - \mathbb{E}\left[Y^2\right] > t\right) \le \frac{\operatorname{Var}\left(Y^2\right)}{nt^2} \le \frac{\sigma^2}{nt^2}$$

thus $\frac{1}{n}\sum_{i=1}^n Y_i^2 > \mathbb{E}\left[Y^2\right] + \sqrt{8\sigma^2/(n\delta)}$ with probability at most $\delta/8$ and

$$d_{\mathbb{X}}(f,f') > 2\sqrt{2}d_\infty(g,g')\sqrt{\Delta^2 + \mathbb{E}\left[Y^2\right] + \sqrt{\frac{8\sigma^2}{n\delta}}}$$

occurs with a probability bounded by $\frac{\delta}{8}$. Recall again that for two distances d_1 and d_2 such that $d_1 \le cd_2$ one has $\gamma_2(\mathcal{G}, d_1) \le c\gamma_2(\mathcal{G}, d_2)$. Then Theorem 25 applies with

$$\Gamma_\delta = 2\sqrt{2}\sqrt{\Delta^2 + \mathbb{E}\left[Y^2\right] + \sqrt{\frac{8\sigma^2}{n\delta}}}\gamma_2(\mathcal{G}, d_\infty)$$

and $\Gamma_\delta \ge \Delta \ge \mathrm{diam}_d(\mathcal{F})$.

Corollary 32. *Consider the setup described above. Let $n \in \mathbb{N}$, $\delta \in (0, 1/3)$ and $\alpha = \sqrt{2\log(\delta^{-1})/(nv)}$. There exists an integer N_0 and a universal constant C such that, for all $n \ge N_0$, with probability at least $1 - 3\delta$,*

$$\begin{aligned} &R(\widehat{g}) - R(g^*) \\ &\quad\le \quad 12\sqrt{\frac{2v\log(\delta^{-1})}{n}} + C\sqrt{\log\left(\frac{8}{\delta}\right)}\sqrt{\frac{\Delta^2 + \mathbb{E}\left[Y^2\right] + 8\sigma^2/(n\delta)}{n}}\int_0^\Delta \sqrt{\log N_\infty(\mathcal{G}, \epsilon)}d\epsilon\,. \end{aligned}$$

Proof. Apply Theorem 25 and note that the condition holds for sufficiently large n. □

The bound of the corollary essentially matches the best rates of convergence one can get even in the case of bounded regression under such general conditions. For special cases, such as linear regression, better bounds may be proven for other methods, see Audibert and Catoni [7], Hsu and Sabato [32], Minsker [52].

3.4.2 k-means clustering under heavy tailed distribution

In *k-means clustering*—or *vector quantization*—one wishes to represent a distribution by a finite number of points. Formally, let X be a random vector taking values in $\mathbb{R}^m$ and let P denote the distribution of X. Let $k \ge 2$ be a positive integer that we fix for the rest of the section. A clustering scheme is given by a set of k cluster centers $C = \{y_1, \ldots, y_k\} \subset \mathbb{R}^m$ and a *quantizer* $q : \mathbb{R}^m \to C$. Given a *distortion measure* $\ell : \mathbb{R}^m \times \mathbb{R}^m \to [0, \infty)$, one wishes to find C and q such that the expected distortion

$$D_k(P, q) = \mathbb{E}\ell(X, q(X))$$

is as small as possible. The minimization problem is meaningful whenever $\mathbb{E}\ell(X, 0) < \infty$ which we assume throughout. Typical distortion measures are of the form $\ell(x, y) = \|x - y\|^\alpha$ where $\|\cdot\|$ is a norm on $\mathbb{R}^m$ and $\alpha > 0$ (typically α equals 1 or 2). Here, for concreteness and simplicity, we assume that ℓ is the Euclidean distance $\ell(x, y) = \|x - y\|$

though the results may be generalized in a straightforward manner to other norms. In a way equivalent to the arguments of Section 3.4.1, the results may be generalized to the case of the quadratic distortion $\ell(x, y) = \|x - y\|^2$. In order to avoid repetition of arguments, the details are omitted.

It is not difficult to see that if $\mathbb{E}\|X\| < \infty$, then there exists a (not necessarily unique) quantizer q^* that is optimal, that is, q^* is such that for all clustering schemes q,

$$D_k(P, q) \geq D_k(P, q^*) =: D_k^*(P) .$$

It is also clear that q^* is a *nearest neighbor quantizer*, that is,

$$\left\|x - q^*(x)\right\| = \min_{y_i \in C} \left\|x - y_i\right\| .$$

Thus, nearest neighbor quantizers are determined by their cluster centers $C = \{y_1, \ldots, y_k\}$. In fact, for all quantizers with a particular set C of cluster centers, the corresponding nearest neighbor quantizer has minimal distortion and therefore it suffices to restrict our attention to nearest neighbor quantizers.

In the problem of empirical quantizer design, one is given an i.i.d. sample $X_1, \ldots, X_n$ drawn from the distribution P and one's aim is to find a quantizer q_n whose distortion

$$D_k(P, q_n) = \mathbb{E}\left[\|X - q_n(X)\| \middle| X_1, \ldots, X_n\right]$$

is as close to $D_k^*(P)$ as possible. A natural strategy is to choose a quantizer–or equivalently, a set C of cluster centers–by minimizing the *empirical distortion*

$$D_k(P_n, q) = \frac{1}{n} \sum_{i=1}^{n} \|X_i - q(X_i)\| = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\ldots,k} \|X_i - y_j\| ,$$

where P_n denotes the standard empirical distribution based on $X_1, \ldots, X_n$. If $\mathbb{E}\|X\| < \infty$, then the empirically optimal quantizer asymptotically minimizes the distortion. More precisely, if q_n denotes the empirically optimal quantizer (i.e., $q_n = \operatorname{argmin}_q D_k(P_n, q)$), then

$$\lim_{n\to\infty} D_k(P, q_n) = D_k^*(P) \quad \text{with probability 1,}$$

see Pollard [55, 57] and Abaya and Wise [1] (see also Linder [45]). The rate of convergence of $D_k(P, q_n)$ to $D_k^*(P)$ has drawn considerable attention, see, e.g., Pollard [56], Bartlett, Linder, and Lugosi [9], Antos [4], Antos, Györfi, and György [5], Biau, Devroye, and Lugosi [12], Maurer and Pontil [50], and Levrard [43]. Such rates are typically studied under the assumption that X is almost surely bounded. Under such assumptions one can show that

$$\mathbb{E}D_k(P, q_n) - D_k^*(P) \leq C(P, k, m)n^{-1/2}$$

where the constant $C(P, k, m)$ depends on esssup $\|X\|$, k, and the dimension m. The value of the constant has mostly been investigated in the case of quadratic loss $\ell(x, y) = \|x-y\|^2$

but most proofs may be modified for the case studied here. For the quadratic loss, one may take $C(P, k, m)$ as a constant multiple of $B^2 \min(\sqrt{k^{1-2/m}m}, k)$ where $B = \operatorname{esssup} \|X\|$.

However, little is known about the finite-sample performance of empirically designed quantizers under possibly heavy-tailed distributions. In fact, there is no hope to extend the results cited above for distributions with finite second moment simply because empirical averages are poor estimators of means under such general conditions.

In the recent paper of Telgarsky and Dasgupta [62], bounds on the excess risk under conditions on higher moments have been developed. They prove a bound of $O(n^{-1/2+2/p})$ for the excess distortion where p is the number of moments of $\|X\|$ that are assumed to be finite. Here we show that there exists an empirical quantizer $\hat{q}_n$ whose excess distortion $D_k(P, \widehat{q_n}) - D_k^*(P)$ is of the order of $n^{-1/2}$ (with high probability) under the only assumption that $\mathbb{E}\left[\|X\|^2\right]$ is finite. This may be achieved by choosing a quantizer that minimizes Catoni's estimate of the distortion.

The proposed empirical quantizer uses two parameters that depend on the (unknown) distribution of X. For simplicity, we assume that upper bounds for these two parameters are available. (Otherwise either one may try to estimate them or, as the sample size grows, use increasing values for these parameters. The details go beyond the scope of this chapter.)

One of these parameters is the second moment $\operatorname{Var}(X) = \mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right]$ and let V be an upper bound. The other parameter $\rho > 0$ is an upper bound for the norm of the possible cluster centers. The next lemma offers an estimate.

Lemma 33. *(Linder [45].) Let $2 \le j \le k$ be the unique integer such that $D_k^* = \cdots = D_j^* < D_{j-1}^*$ and define $\varepsilon = (D_{j-1}^* - D_j^*)/2$. Let $(y_1, \dots, y_j)$ be a set of cluster centers such that the distortion of the corresponding quantizer is less than $D_j^* + \varepsilon$. Let $B_r = \{x : \|x\| \le r\}$ denote the closed ball of radius $r > 0$ centered at the origin. If $\rho > 0$ is such that*

- $\frac{\rho}{10} P(B_{\frac{\rho}{10}}) > 2\mathbb{E}\|X\|$
- $P(B_{2\rho/5}) > 1 - \frac{\varepsilon^2}{4\mathbb{E}\left[\|X\|^2\right]}$

then for all $1 \le i \le k$, $\left\|y_i\right\| \le \rho$.

Now we are prepared to describe the proposed empirical quantizer. Let C_ρ be the set of all collections $C = \{y_1, \dots, y_k\} \in (\mathbb{R}^m)^k$ of cluster centers with $\|y_j\| \le \rho$ for all $j = 1, \dots, k$. For each $C \in C_\rho$, denote by q_C the corresponding quantizer. Now for all $C \in C_\rho$, we may calculate Catoni's mean estimator of the distortion $D(P, q_C) = \mathbb{E}\|X - q_C(X)\| = \mathbb{E} \min_{j=1,\dots,k} \|X_i - y_j\|$ defined as the unique value $\mu \in \mathbb{R}$ for which

$$\frac{1}{n\alpha} \sum_{i=1}^{n} \phi\left(\alpha\left(\min_{j=1,\dots,k} \|X_i - y_j\| - \mu\right)\right) = 0$$

where we use the parameter value $\alpha = \sqrt{2/(nkV)}$. Denote this estimator by $\widehat{D}(P_n, q_C)$ and let $\widehat{q}_n$ be any quantizer minimizing the estimated distortion. An easy compactness argument shows that such a minimizer exists.

The main result of this section is the following bound for the distortion of the chosen quantizer.

Theorem 34. *Assume that Var* $(X) \leq V < \infty$ *and* $n \geq m$. *Then, with probability at least* $1 - \delta$,

$$D_k(P, \widehat{q}_n) - D_k(P, q^*) \leq C\left(\log \frac{1}{\delta}\right)\left(\sqrt{\frac{Vk}{n}} + \sqrt{\frac{mk}{n}}\right) + O\left(\frac{1}{n}\right),$$

where the constant C *only depends on* ρ.

Proof. The result follows from Theorem 24. All we need to check is that $\mathrm{Var}\left(\min_{j=1,\ldots,k} \|X - y_j\|\right)$ is bounded by kV and estimate the covering numbers of the class of functions

$$\mathcal{F}_\rho = \left\{ f_C(x) = \min_{y \in C} \|x - y\| : C \in \mathcal{C}_\rho \right\}.$$

The variance bound follows simply by the fact that for all $C \in \mathcal{C}$,

$$\begin{aligned}
\mathrm{Var}\left(\min_{j=1,\ldots,k} \|X - y_j\|\right) &\leq \sum_{i=1}^{k} \mathrm{Var}\left(\left\|X - y_i\right\|\right) \\
&\leq \sum_{i=1}^{k} \mathbb{E}\left[\|X - \mathbb{E}X\|^2\right] + \left\|\mathbb{E}X - y_i\right\|^2 - \mathbb{E}\left[\left\|X - y_i\right\|\right]^2 \leq kV .
\end{aligned}$$

In order to use the bound of Theorem 24, we need to bound the covering numbers of the class $\mathcal{F}_\rho$ under both metrics d and D. We begin with the metric

$$D(f_C, f_{C'}) = \sup_{x \in \mathbb{R}^m} |f_C(x) - f_{C'}(x)| .$$

The notation $B_z(\epsilon, d)$ refers to the ball under the metric d of radius ϵ centered at z. Let Z be a subset of B_ρ such that

$$\mathcal{B}_{B_\rho} := \{B_z(\epsilon, \|\cdot\|) : z \in Z\}$$

is a covering of the set B_ρ by balls of radius ϵ under the Euclidean norm. Let $C \in \mathcal{C}_\rho$ and associate to any $y_i \in C$ one of the centers in Z such that $\left\|y_i - z_i\right\| \leq \epsilon$. If there is more than one possible choice for z_i, we pick one of them arbitrarily. We denote by $q_{C'}$ the nearest neighbor quantizer with codebook $C' = (z_i)_i$. Finally, let $S_i = q_{C'}^{-1}(z_i)$. Now clearly, $\forall i,\ \forall x \in S_i$

$$\begin{aligned}
f_C(x) - f_{C'}(x) &= \min_{1 \leq j \leq k} \|x - y_j\| - \min_{1 \leq j \leq k} \|x - z_j\| \\
&= \min_{1 \leq j \leq k} \|x - y_j\| - \|x - z_i\| \\
&\leq \|x - y_i\| - \|x - z_i\| \leq \epsilon
\end{aligned}$$

and similarly, $f_{C'}(x) - f_C(x) \leq \epsilon$. Then $f_C \in B_{f_{C'}}(\epsilon, D)$ and

$$\mathcal{B}_{\mathcal{F}_\rho} := \{B_{f_C}(\epsilon, D) : C \in Z^k\}$$

is a covering of $\mathcal{F}_\rho$. Since Z can be taken such that $|Z| = N_{\|\cdot\|}(B_\rho, \epsilon)$ we obtain

$$N_d(\mathcal{F}_\rho, \epsilon) \leq N_D(\mathcal{F}_\rho, \epsilon) \leq N_{\|\cdot\|}(B_\rho, \epsilon)^k .$$

By standard estimates on the covering numbers of the ball B_ρ by balls of size ϵ under the Euclidean metric,

$$N_{\|\cdot\|}(B_\rho, \epsilon) \leq \left(\frac{4\rho}{\epsilon}\right)^m$$

(see, e.g., Matousek [49]). In other words, there exists a universal constant L and constants C_ρ and C'_ρ that depends only on ρ such that

$$\gamma_2(\mathcal{F}_\rho, d) \leq L \int_0^{2\rho} \sqrt{\log N_d(\mathcal{F}_\rho, \epsilon)} d\epsilon \leq C_\rho \sqrt{km}$$
$$\text{and} \quad \gamma_1(\mathcal{F}_\rho, D) \leq L \int_0^{2\rho} \log N_D(\mathcal{F}_\rho, \epsilon) d\epsilon \leq C'_\rho km .$$

Theorem 24 may now be applied to the class $\mathcal{F}_\rho$. □

3.5 Simulation Study

In this closing section we present the results of two simulation exercises that assess the performance of the estimators developed in this work.

3.5.1 L_2 Regression

The first application is an L_2 regression exercise. Data are simulated from a linear model with heavy-tailed errors and the L_2 regression procedure based on Catoni's risk minimizer introduced in Section 3.4.1 is used for estimation. The procedure is benchmarked against regular ("vanilla") L_2 regression based on the minimization of the empirical L_2 loss.

The simulation exercise is designed as follows. We simulate $(Z_1, Y_1), (Z_2, Y_2), ..., (Z_n, Y_n)$ i.i.d. pairs of random variables in $\mathbb{R}^5 \times \mathbb{R}$. The vector Z_i of explanatory variables is drawn from a multivariate normal distribution with zero mean, unit variance, and correlation matrix equal to an equicorrelation matrix with correlation $\rho = 0.9$. The response variable Y_i is generated as

$$Y_i = Z_i^T \theta^* + \epsilon_i ,$$

where the parameter vector θ^* is set to $(0.25, -0.25, 0.50, 0.70, -0.75)$ and ϵ_i is a zero mean error term. The error term ϵ_i is drawn from a Pareto distribution with tail parameter

β and is appropriately recentered in order to have zero mean. As it is well known, the tail parameter β determines which moments of the Pareto random variable are finite. More specifically, the moment of order k exists only if $k < \beta$. The focus is on finding the value of θ which minimizes the L_2 risk

$$\mathbb{E}\left|Y - Z_i^T\theta\right|^2 .$$

The parameter θ is estimated using the Catoni and the vanilla L_2 regressions. Let $\widehat{R}_C(\theta)$ denote the solution of the equation

$$\hat{r}_\theta(\mu) = \frac{1}{n\alpha}\sum_{i=1}^{n}\phi\left(\alpha\left(\left|Y_i - Z_i^T\theta\right|^2 - \mu\right)\right) = 0 .$$

Then the Catoni L_2 regression estimator is defined as

$$\widehat{\theta}_{n\,C} = \arg\min_\theta \widehat{R}_C(\theta) .$$

The vanilla L_2 regression estimator is defined as the minimizer of the empirical L_2 loss,

$$\widehat{\theta}_{n\,V} = \arg\min_\theta \widehat{R}_V(\theta) = \arg\min_\theta \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - Z_i^T\theta\right|^2 ,$$

which is the classical least squares estimator. The precision of each estimator is measured by their excess risk

$$\begin{aligned} R(\widehat{\theta}_{n\,C}) - R(\theta^*) &= \mathbb{E}|Y - Z^T\widehat{\theta}_{n\,C}|^2 - \mathbb{E}|Y - Z^T\theta^*|^2 \\ R(\widehat{\theta}_{n\,V}) - R(\theta^*) &= \mathbb{E}|Y - Z^T\widehat{\theta}_{n\,V}|^2 - \mathbb{E}|Y - Z^T\theta^*|^2 . \end{aligned}$$

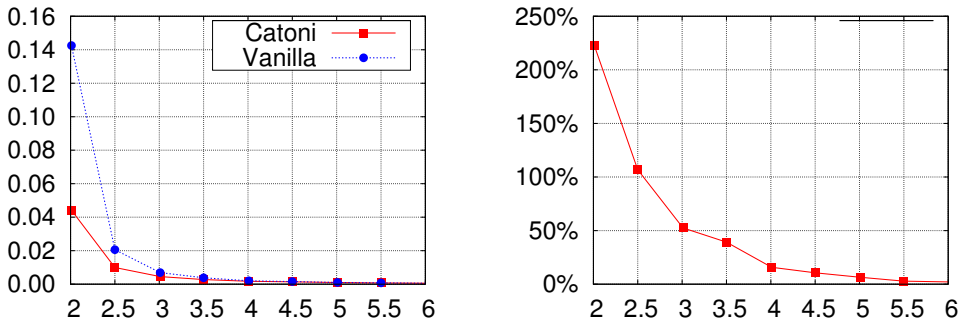
We estimate excess risk by simulation. For each replication of the simulation exercise, we estimate the risk of the estimators and the optimal risk using sample averages based on an i.i.d. sample $(Z_1', Y_1'), \ldots, (Z_m', Y_m')$ that is independent of the one used for estimation, that is

$$\begin{aligned} \widetilde{R}(\widehat{\theta}_{n\,C}) &= \tfrac{1}{m}\textstyle\sum_{i=1}^{m}|Y_i' - Z_i'^T\widehat{\theta}_{n\,C}|^2 \\ \widetilde{R}(\widehat{\theta}_{n\,V}) &= \tfrac{1}{m}\textstyle\sum_{i=1}^{m}|Y_i' - Z_i'^T\widehat{\theta}_{n\,V}|^2 \\ \widetilde{R}(\theta^*) &= \tfrac{1}{m}\textstyle\sum_{i=1}^{m}|Y_i' - Z_i'^T\theta^*|^2 . \end{aligned} \tag{3.12}$$

The simulation experiment is replicated for different values of the Pareto tail parameter β ranging from 2.01 to 6.01 and different values of the sample size n, ranging from 50 to 1,000. For each combination of the tail parameter β and sample size n, the experiment is replicated 1,000 times.

Figure 3.2 displays the Monte Carlo estimate of the excess risk of the Catoni and benchmark regression estimators as functions of the tail parameter β when the sample size n is equal to 500. The left panel shows the level of the excess risks $R(\widehat{\theta}_{n\,C})$ –

Figure 3.2: L_2 Regression Parameter Estimation.



(a) (b)

The figure plots the excess risk of the Catoni and vanilla L_2 regression parameter estimators (a) and the percentage improvement of the Catoni procedure relative to the vanilla (b) as a function of the tail parameter β for a sample size n equal to 500.

Table 3.1: Relative Performance of the Catoni L_2 Parameter Estimator.

β	**n=50**	**n=100**	**n=250**	**n=500**	**n=750**	**n=1000**
2.01	3872.10	440.50	171.30	222.70	218.20	142.80
2.50	169.20	158.70	151.50	106.70	91.70	57.40
3.01	137.60	178.00	89.00	52.50	62.70	63.50
3.50	54.40	20.90	41.30	39.20	38.10	33.50
4.01	30.20	44.40	25.50	15.70	16.30	15.90
4.50	16.50	12.10	11.30	10.60	6.90	13.70
5.01	10.20	7.80	10.20	6.40	5.70	3.10
5.50	6.00	14.80	3.90	2.90	2.10	2.20
6.01	3.90	1.90	2.70	2.10	1.90	1.40

The table reports the percentage improvement of the excess risk of the Catoni L_2 regression estimator relative to the vanilla procedure for different values of the tail parameter β and sample size n.

$R(\theta^*)$ and $R(\widehat{\theta}_{n\,V}) - R(\theta^*)$ as a function of β and the right one shows the percentage improvement of the excess risk of the Catoni procedure over the benchmark calculated as $\left(R(\widehat{\theta}_{n\,V}) - R(\widehat{\theta}_{n\,C})\right) / \left(R(\widehat{\theta}_{n\,C}) - R(\theta^*)\right)$. When the tails are not excessively heavy (high values of β) the difference between the procedures is small. As the tails become heavier (small values of β) the risks of both procedures increase. Importantly, the Catoni estimator becomes progressively more efficient as the tails become heavier and becomes significantly more efficient when the tail parameter is close to 2. Detailed results for different values of n are reported in Table 3.1. Overall the Catoni L_2 regression estimator never performs worse than the benchmark and it is substantially better when the tails of the data are heavy.

3.5.2 k-means

In the second experiment we carry out a k–means clustering exercise. Data are simulated from a heavy-tailed mixture distribution and then cluster centers are chosen by minimizing Catoni's estimate of the L_2 distortion. The performance of the algorithm is benchmarked against the ("vanilla") k–means algorithm procedure where the distortion is estimated by the standard empirical average.

The simulation exercise is designed as follows. An i.i.d. sample of random vectors $X_1, ..., X_n$ in $\mathbb{R}^2$ is drawn from a four-component mixture distribution with equal weights. The means of the mixture components are $(5, 5), (-5, 5), (-5, -5)$, and $(5, -5)$. Each component of the mixture is made up of two appropriately centered independent draws from a Pareto distribution with tail parameter β. The cluster centers obtained by the k–means algorithm based on Catoni and the vanilla k–means algorithm are denoted respectively by $\widehat{q}_{n\,C}$ and $\widehat{q}_{n\,V}$. (Since finding the empirically optimal cluster centers is computationally prohibitive, we use the well-known iterative optimization procedure "k-means" for the vanilla version and a similar variant for the Catoni scheme.) Analogously to the previous exercise, we summarize the performance of the clustering procedures using the excess risk of the algorithms, that is

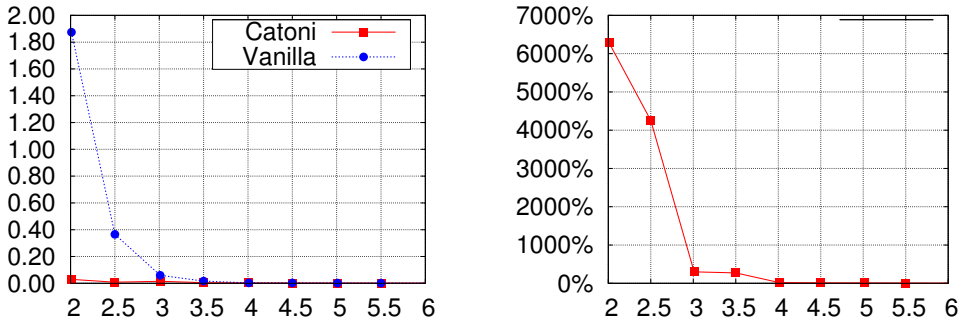
$$\begin{array}{l} D_k(P, \widehat{q}_{n\,C}) - D_k(P, q^*) \\ D_k(P, \widehat{q}_{n\,V}) - D_k(P, q^*) \end{array} \quad ,$$

where q^* denotes the means of the mixture components. We estimate excess risk by simulation. We compute the distortion of the quantizers using an i.i.d. sample $X'_1, \ldots, X'_m$ of vectors that is independent of the ones used for estimation, that is,

$$\begin{array}{rcl} D_k(P_m, \widehat{q}_{n\,C}) & = & \frac{1}{m} \sum_{i=1}^m \min_{j=1,\ldots,k} \left\| X'_i - \widehat{q}_{n\,C}(X'_i) \right\|^2 \\ D_k(P_m, \widehat{q}_{n\,V}) & = & \frac{1}{m} \sum_{i=1}^m \min_{j=1,\ldots,k} \left\| X'_i - \widehat{q}_{n\,V}(X'_i) \right\|^2 \\ D_k(P_m, q^*) & = & \frac{1}{m} \sum_{i=1}^m \min_{j=1,\ldots,k} \left\| X'_i - q^*(X'_i) \right\|^2 \end{array} \quad . \tag{3.13}$$

The experiment is replicated for different values of the tail parameter β ranging from 2.01 to 6.01 and different values of the sample size n ranging from 50 to 1,000. For each

Figure 3.3: k–means Quantizer Estimation.



(a) (b)

The figure plots the excess risk of the Catoni and vanilla k–means quantizer estimator (a) and the percentage improvement of the Catoni procedure relative to the vanilla (b) as a function of the tail parameter β for a sample size n equal to 500.

Table 3.2: Relative Performance of the Catoni k–means Quantizer Estimator.

β	**n=50**	**n=100**	**n=250**	**n=500**	**n=750**	**n=1000**
2.01	823.80	2180.40	3511.60	6278.90	7858.70	10684.60
2.50	404.50	1007.40	2959.80	4255.40	6828.60	9093.60
3.01	301.10	312.20	286.80	298.60	813.60	1560.20
3.50	129.60	188.60	213.30	271.40	448.60	410.00
4.01	73.80	30.90	26.80	20.30	18.20	13.10
4.50	27.60	22.90	16.50	11.70	9.50	10.10
5.01	16.40	10.80	11.60	8.70	6.00	7.20
5.50	9.00	6.80	9.20	5.00	4.10	4.00
6.01	3.50	4.70	5.00	2.70	3.20	3.10

The table reports the improvement of the Catoni k–means quantizer estimator relative to the vanilla procedure for different values of the tail parameter β and sample size n.

combination of tail parameter β and sample size n the experiment is replicated $1,000$ times.

Figure 3.3 displays the Monte Carlo estimate of excess risk of the Catoni and benchmark estimators as a function of tail parameter β for $n = 500$. The left panel shows the estimated excess risk while the right panel shows the percentage improvement of the excess risk of the Catoni procedure, calculated as $(D_k(P, \widehat{q}_{n\,V}) - D_k(P, \widehat{q}_{n\,C}))\,/\,(D_k(P, \widehat{q}_{n\,C}) - D_k(P, q^*))$.

The overall results are analogous to the ones of the L_2 regression application. When the tails of the mixture are not excessively heavy (high values of β) the difference in

the procedures is small. As the tails become heavier (small values of β) the risk of both procedures increases, but the Catoni algorithm becomes progressively more efficient. The percentage gains for the Catoni procedure are substantial when the tail parameter is smaller than 4. Table 3.2 reports detailed results for different values of n. As in the L_2 regression simulation study, the Catoni k–means algorithm never performs worse than the benchmark and it is substantially better when the tails of the mixture are heavy.

3.6 Appendix

3.6.1 A chaining theorem

The following result is a version of standard bounds based on "generic chaining", see Talagrand [61]. We include the proof for completeness.

Recall that if ψ is a non-negative increasing convex function defined on $\mathbb{R}_+$ with $\psi(0) = 0$, then the Orlicz norm of a random variable X is defined by

$$\|X\|_\psi = \inf\left\{c > 0 \ : \ \mathbb{E}\left[\psi\left(\frac{|X|}{c}\right)\right] \leq 1\right\} .$$

We consider Orlicz norms defined by

$$\psi_1(x) = \exp(x) - 1 \quad \text{and} \quad \psi_2(x) = \exp(x^2) - 1 .$$

For further information on Orlicz norms see [64, Chapter 2.2]. First, $\|X\|_{\psi_1} \leq \|X\|_{\psi_2} \sqrt{\log(2)}$ holds. Also note that, by Markov's inequality, $\|X\|_{\psi_1} \leq c$ implies that $\mathbb{P}\{|X| > t\} \leq 2e^{-t/c}$ and similarly, if $\|X\|_{\psi_2} \leq c$, then $\mathbb{P}\{|X| > t\} \leq 2e^{-t^2/c^2}$. Then

$$\begin{aligned} X &\leq \|X\|_{\psi_1} \log(2\delta^{-1}) && \text{with probability at least } 1-\delta , \\ X &\leq \|X\|_{\psi_2} \sqrt{\log(2\delta^{-1})} && \text{with probability at least } 1-\delta . \end{aligned} \tag{3.14}$$

Recall the following definition (see, e.g., [61, Definition 1.2.3]). Let T be a (pseudo) metric space. An increasing sequence $(\mathcal{A}_n)$ of partitions of T is called *admissible* if for all $n = 0, 1, 2, \ldots$, $\#\mathcal{A}_n \leq 2^{2^n}$. For any $t \in T$, denote by $A_n(t)$ the unique element of $\mathcal{A}_n$ that contains t. Let $\Delta(A)$ denote the diameter of the set $A \subset T$. Define, for $\beta = 1, 2$,

$$\gamma_\beta(T, d) = \inf_{\mathcal{A}_n} \sup_{t \in T} \sum_{n \geq 0} 2^{n/\beta} \Delta(A_n(t)) ,$$

where the infimum is taken over all admissible sequences. First of all, we know from [61, Eq. (1.18)] that there exists a universal constant L such that

$$\gamma_\beta(T, d) \leq L \int_0^{\mathrm{diam}_d(T)} (\log N_d(T, \varepsilon))^{\frac{1}{\beta}} \, d\varepsilon \tag{3.15}$$

Theorem 35. *Let $(X_t)_{t\in T}$ be a stochastic process indexed by a set T on which two (pseudo) metrics, d_1 and d_2, are defined such that T is bounded with respect to both metrics. Assume that for any $s,t \in T$ and for all $x > 0$,*

$$\mathbb{P}\{|X_s - X_t| > x\} \leq 2\exp\left(-\frac{1}{2}\frac{x^2}{d_2(s,t)^2 + d_1(s,t)x}\right).$$

Then for all $t \in T$,

$$\left\|\sup_{s\in T}|X_s - X_t|\right\|_{\psi_1} \leq L\left(\gamma_1(T,d_1) + \gamma_2(T,d_2)\right)$$

with $L \leq 384\log(2)$.

The proof of Theorem 35 uses the following lemma:

Lemma 36. ([64, LEMMA 2.2.10].) *Let $a,b > 0$ and assume that the random variables $X_1,\dots,X_m$ satisfy, for all $x > 0$,*

$$\mathbb{P}\{|X_i| > x\} \leq 2\exp\left(-\frac{1}{2}\frac{x^2}{b + ax}\right).$$

Then

$$\left\|\max_{1\leq i\leq m} X_i\right\|_{\psi_1} \leq 48\left(a\log(1+m) + \sqrt{b}\sqrt{\log(1+m)}\right).$$

Proof of Theorem 35. Consider an admissible sequence $(\mathcal{B}_n)_{n\geq 0}$ such that for all $t \in T$,

$$\sum_{n\geq 0} 2^n \Delta_1(B_n(t)) \leq 2\gamma_1(T,d_1)$$

and an admissible sequence $(\mathcal{C}_n)_{n\geq 0}$ such that for all $t \in T$,

$$\sum_{n\geq 0} 2^{n/2} \Delta_2(C_n(t)) \leq 2\gamma_2(T,d_2).$$

Now we may define an admissible sequence by intersection of the elements of $(\mathcal{B}_{n-1})_{n\geq 1}$ and $(\mathcal{C}_{n-1})_{n\geq 1}$: set $\mathcal{A}_0 = \{T\}$ and let

$$\mathcal{A}_n = \{B \cap C \ : \ B \in \mathcal{B}_{n-1} \ \& \ C \in \mathcal{C}_{n-1}\}.$$

$(\mathcal{A}_n)_{n\geq 0}$ is an admissible sequence because each $\mathcal{A}_n$ is increasing and contains at most $(2^{2^{n-1}})^2 = 2^{2^n}$ sets. Define a sequence of finite sets $T_0 = \{t\} \subset T_1 \subset \cdots \subset T$ such that T_n contains a single point in each set of $\mathcal{A}_n$. For any $s \in T$, denote by $\pi_n(s)$ the unique element of T_n in $A_n(s)$. Now for any $s \in T_{k+1}$, we write

$$X_s - X_t = \sum_{k=0}^{\infty}\left(X_{\pi_{k+1}(s)} - X_{\pi_k(s)}\right).$$

Then, using the fact that $\|\cdot\|_{\psi_1}$ is a norm and Lemma 36,

$$\begin{aligned}
&\left\|\sup_{s\in T}|X_s - X_t|\right\|_{\psi_1} \\
&\quad\leq \sum_{k=0}^{\infty}\left\|\max_{s\in T_{k+1}}|X_{\pi_{k+1}(s)} - X_{\pi_k(s)}|\right\|_{\psi_1} \\
&\quad\leq 48\sum_{k=0}^{\infty}\left(d_1(\pi_{k+1}(s),\pi_k(s))\log(1+2^{2^{k+1}}) + d_2(\pi_{k+1}(s),\pi_k(s))\sqrt{\log(1+2^{2^{k+1}})}\right).
\end{aligned}$$

Since $(\mathcal{A}_n)_{n\geq 0}$ is an increasing sequence, $\pi_{k+1}(s)$ and $\pi_k(s)$ are both in $A_k(s)$. By construction, $A_k(s)\subset B_k(s)$, and therefore $d_1(\pi_{k+1}(s),\pi_k(s))\leq \Delta_1(B_k(s))$. Similarly, we have $d_2(\pi_{k+1}(s),\pi_k(s))\leq \Delta_2(C_k(s))$. Using $\log(1+2^{2^{k+1}})\leq 4\log(2)2^k$, we get

$$\begin{aligned}
\left\|\max_{s\in T}|X_s - X_t|\right\|_{\psi_1} &\leq 192\log(2)\left[\sum_{k=0}^{\infty}2^k\Delta_1(B_k(s)) + \sum_{k=0}^{\infty}2^{k/2}\Delta_2(C_k(s))\right] \\
&\leq 384\log(2)\left[\gamma_1(T,d_1)+\gamma_2(T,d_2)\right].
\end{aligned}$$

□

Proposition 37. *Assume that for any $s,t\in T$ and for all $x>0$,*

$$\mathbb{P}\{|X_s - X_t| > x\}\leq 2\exp\left(-\frac{x^2}{2d_2(s,t)^2}\right).$$

Then for all $t\in T$,

$$\left\|\sup_{s\in T}|X_s - X_t|\right\|_{\psi_2}\leq L\gamma_2(T,d_2)$$

where L is a universal constant.

The proof of Proposition 37 is similar to the proof of Theorem 35. One merely needs to replace Lemma 36 by Lemma 2.2.2 in [64] and proceed identically. The details are omitted.

We may use Proposition 37 to bound the moment generating function of $\sup_{s\in T}|X_s - X_t|$ as follows. Set $S=\sup_{s\in T}|X_s - X_t|$. Then using $ab\leq (a^2+b^2)/2$, we have, for every $\lambda>0$,

$$\exp(\lambda S)\leq \exp\left(S^2/\|S\|_{\psi_2}^2 + \lambda^2\|S\|_{\psi_2}^2/4\right),$$

and therefore

$$\mathbb{E}\left[\exp(\lambda\sup_{s\in T}|X_s - X_t|)\right]\leq 2\exp(\lambda^2L^2\gamma_2(T,d_2)^2/4). \tag{3.16}$$

Chapter 4

Robust estimation of U-statistics

In this chapter we discuss the estimation of the mean of multivariate functions in case of possibly heavy-tailed distributions. In such situations, reliable estimates of the mean cannot be obtained by usual U-statistics. We introduce a new estimator, based on the so-called median-of-means technique. We develop performance bounds for this new estimator that generalizes an estimate of [6], showing that the new estimator performs, under minimal moment conditions, as well as classical U-statistics for bounded random variables. We discuss an application of this estimator to clustering. This chapter is a joint work with Gábor Lugosi. It is based on a paper [37] submitted to Stochastic Processes and their Applications.

Contents

4.1 Introduction

Motivated by numerous applications, the theory of U-statistics and U-processes has received considerable attention in the past decades. U-statistics appear naturally in *ranking* [22], *clustering* [21] and *learning on graphs* [11] or as components of higher-order terms in expansions of smooth statistics, see, for example, [59]. The general setting may be described as follows. Let X be a random variable taking values in some measurable space $\mathcal{X}$ and let $h : \mathcal{X}^m \to \mathbb{R}$ be a measurable function of $m \geq 2$ variables. Let P be the probability measure of X. Suppose we have access to $n \geq m$ independent random variables $X_1, \ldots, X_n$, all distributed as X. We define the U-statistics of order m and kernel h based on the sequence $\{X_i\}$ as

$$U_n(h) = \frac{(n-m)!}{n!} \sum_{(i_1,\ldots,i_m)\in I_n^m} h(X_{i_1}, \ldots, X_{i_m}) \,, \tag{4.1}$$

where

$$I_n^m = \Big\{(i_1, \ldots, i_m) : 1 \leq i_j \leq n,\ i_j \neq i_k \ \text{ if } \ j \neq k\Big\}$$

is the set of all m-tuples of different integers between 1 and n. U-statistics are unbiased estimators of the mean $m_h = \mathbb{E}h(X_1, \ldots, X_m)$ and have minimal variance among all unbiased estimators [30]. Understanding the concentration of a U-statistics around its expected value has been subject of extensive study. [23] provide an excellent summary but see also [29] for a more recent development.

By a classical inequality of [31], for a bounded kernel h, for all $\delta > 0$,

$$\mathbb{P}\left(|U_n(h) - m_h| > \|h\|_\infty \sqrt{\frac{\log(\frac{2}{\delta})}{2\lfloor n/m \rfloor}}\right) \leq \delta \,, \tag{4.2}$$

and we also have the "Bernstein-type" inequality

$$\mathbb{P}\left(|U_n(h) - m_h| > \sqrt{\frac{4\sigma^2 \log(\frac{2}{\delta})}{2\lfloor n/m \rfloor}} \vee \frac{4\,\|h\|_\infty \log(\frac{2}{\delta})}{6\lfloor n/m \rfloor}\right) \leq \delta \,,$$

where $\sigma^2 = \mathrm{Var}\,(h(X_1, \ldots, X_m))$.

However, under certain degeneracy assumptions on the kernel, significantly sharper bounds have been proved. Following the exposition of [23], for convenience, we restrict our attention to symmetric kernels. A kernel h is *symmetric* if for all $x_1, \ldots, x_m \in \mathbb{R}$ and all permutations s,

$$h(x_1, \ldots, x_m) = h(x_{s_1}, \ldots, x_{s_m}) \,.$$

A symmetric kernel h is said to be *P-degenerate* of order $q-1$, $1 < q \leq m$, if for all $x_1, \ldots, x_{q-1} \in \mathcal{X}$,

$$\int h(x_1, \ldots, x_m) dP^{m-q+1}(x_q, \ldots, x_m) = \int h(x_1, \ldots, x_m) dP^m(x_1, \ldots, x_m)$$

and

$$(x_1, \ldots, x_q) \mapsto \int f(x_1, \ldots, x_m) dP^{m-q}(x_{q+1}, \ldots, x_m)$$

is not a constant function. In the special case of $m_h = 0$ and $q = m$ (i.e., when the kernel is $(m-1)$-degenerate, h is said to be *P-canonical*. P-canonical kernels appear naturally in the Hoeffding decomposition of a U-statistic, see [23].

[6] proved the following important improvement of Hoeffing's inequalities for canonical kernels: If $h - m_h$ is a bounded, symmetric P-canonical kernel of m variables, there exist finite positive constants c_1 and c_2 depending only on m such that for all $\delta \in (0, 1)$,

$$\mathbb{P}\left(|U_n(h) - m_h| \geq c_1 \|h\|_\infty \left(\frac{\log(\frac{c_2}{\delta})}{n} \right)^{m/2} \right) \leq \delta \,, \tag{4.3}$$

and also

$$\mathbb{P}\left(|U_n(h) - m_h| > \left(\frac{\sigma^2 \log(\frac{c_1}{\delta})}{c_2 n} \right)^{m/2} \vee \frac{\|h\|_\infty}{\sqrt{n}} \left(\frac{\log(\frac{c_1}{\delta})}{c_2} \right)^{(m+1)/2} \right) \leq \delta \,. \tag{4.4}$$

In the special case of P-canonical kernels of order $m = 2$, (4.3) implies that

$$|U_n(h) - m_h| \leq \frac{c_1 \|h\|_\infty}{n} \log\left(\frac{c_2}{\delta} \right) , \tag{4.5}$$

with probability at least $1 - \delta$. Note that this rate of convergence is significantly faster than the rate $O_p(n^{-1/2})$ implied by (4.2).

All the results cited above require boundedness of the kernel. If the kernel is unbounded but $h(X_1, \ldots, X_m)$ has sufficiently light (e.g., sub-Gaussian) tails, then some of these results may be extended, see, for example, [29]. However, if $h(X_1, \ldots, X_m)$ may have a heavy-tailed distribution, exponential inequalities do not hold anymore (even in the univariate $m = 1$ case). However, even though U-statistics may have an erratic behavior in the presence of heavy tails, in this paper we show that under minimal moment conditions, one may construct estimators of m_h that satisfy exponential inequalities analogous to (4.2) and (4.3). These are the main results of the paper. In particular, in Section 4.2 we introduce a robust estimator of the mean m_h. Theorems 38 and 40 establish exponential inequalities for the performance of the new estimator under minimal moment assumptions. More precisely, Theorem 38 only requires that $h(X_1, \ldots, X_m)$ has a finite variance and establishes inequalities analogous to (4.3) for P-degenerate kernels. In Theorem 40 we further weaken the conditions and only assume that there exists $1 < p \leq 2$ such that $\mathbb{E}|h|^p < \infty$.

The next example illustrates why classical U-statistics fail under heavy-tailed distributions.

Example. Consider the special case $m = 2$, $\mathbb{E}X_1 = 0$ and $h(X_1, X_2) = X_1 X_2$. Note that this kernel is P-canonical. We define $Y_1, \ldots, Y_n$ as independent copies of

$X_1, \dots, X_n$. By decoupling inequalities for the tail of U-statistics given in Theorem 3.4.1 in [23] (see also Theorem 45 in the Appendix), $U_n(h)$ has a similar tail behavior to $\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\left(\frac{1}{n-1}\sum_{j=1}^{n-1} Y_j\right)$. Thus, $U_n(h)$ behaves like a product of two independent empirical mean estimators of the same distribution. When the X_i are heavy tailed, the empirical mean is known to be a poor estimator of the mean. As an example, assume that X follows an α-stable law $S(\gamma, \alpha)$ for some $\alpha \in (1, 2)$ and $\gamma > 0$. Recall that a random variable X has an *α-stable* law $S(\gamma, \alpha)$ if for all $u \in \mathbb{R}$,

$$\mathbb{E}\exp(iuX) = \exp(-\gamma^{\alpha}|u|^{\alpha})$$

(see [68], [54]). Then it follows from the properties of α-stable distributions (summarized in Proposition 47 in the Appendix) that there exists a constant $c > 0$ depending only on α and γ such that

$$\mathbb{P}\left(U_n(h) \geq n^{2/\alpha-2}\right) \geq c\,,$$

and therefore there is no hope to reproduce an upper bound like (4.5). Below we show how this problem can be dealt with by replacing the U-statistics by a more robust estimator.

Our approach is based on robust mean estimators in the univariate setting. Estimation of the mean of a possibly heavy-tailed random variable X from i.i.d. sample $X_1, \dots, X_n$ has recently received increasing attention. Introduced by [53], the *median-of-means* estimator takes a confidence level $\delta \in (0, 1)$ and divides the data into $V \approx \log \delta^{-1}$ blocks. For each block $k = 1, \dots, V$, one may compute the empirical mean $\widehat{\mu}_k$ on the variables in the block. The median $\overline{\mu}$ of the $\widehat{\mu}_k$ is the so-called median-of-means estimator. A short analysis of the resulting estimator shows that

$$|\overline{\mu} - m_h| \leq c\sqrt{\mathrm{Var}\,(X)}\sqrt{\frac{\log(1/\delta)}{n}}$$

with probability at least $1 - \delta$ for a numerical constant c. For the details of the proof see [42]. When the variance is infinite but a moment of order $1 < p \leq 2$ exists, the median-of means estimator is still useful, see [16]. This estimator has recently been studied in various contexts. M-estimation based on this technique has been developed by [42] and generalizations in a multivariate context have been discussed by [32] and [52]. A similar idea was used in [2]. An interesting alternative of the median-of-means estimator has been proposed by [20].

The rest of the paper is organized as follows. In Section 4.2 we introduce a robust estimator of the mean m_h and present performance bounds. In particular, Section 4.2.1 deals with the finite variance case. Section 4.2.2 is dedicated to case when h has a finite p-th moment for some $1 < p < 2$ for P-degenerate kernels. Section 4.2.3 is about a generalization of the robust estimator defined in Section 4.2 to functions taking values in a Hilbert space. Finally, in Section 4.3, we present an application to clustering problems.

4.2 Robust U-estimation

In this section we introduce a "median-of-means"-style estimator of $m_h = \mathbb{E}h(X_1, \dots, X_m)$. To define the estimator, one divides the data into V blocks. For any m-tuple of different blocks, one may compute a (decoupled) U-statistics. Finally, one computes the median of all the obtained values. The rigorous definition is as follows.

The estimator has a parameter $V \le n$, the number of blocks. A partition $\mathcal{B} = (B_1, \dots, B_V)$ of $\{1, \dots, n\}$ is called *regular* if for all $K = 1, \dots, V$,

$$\left| |B_K| - \frac{n}{V} \right| \le 1 \, .$$

For any $B_{i_1}, \dots, B_{i_m}$ in $\mathcal{B}$, we set

$$I_{B_{i_1},\dots,B_{i_m}} = \left\{ (k_1, \dots, k_m) : k_j \in B_{i_j} \right\}$$

and

$$U_{B_{i_1},\dots,B_{i_m}}(h) = \frac{1}{|B_{i_1}| \cdots |B_{i_m}|} \sum_{(k_1,\dots,k_m) \in I_{B_{i_1},\dots,B_{i_m}}} h(X_{k_1}, \dots, X_{k_m}) \, .$$

For any integer N and any vector $(a_1, \dots, a_N) \in \mathbb{R}^N$, we define the median $\mathrm{Med}(a_1, \dots, a_N)$ as any number b such that

$$\left| \{ i \le N \, : \, a_i \le b \} \right| \ge \frac{N}{2} \quad \text{and} \quad \left| \{ i \le N \, : \, a_i \ge b \} \right| \ge \frac{N}{2} \, .$$

Finally, we define the robust estimator:

$$\overline{U}_{\mathcal{B}}(h) = \mathrm{Med}\{ U_{B_{i_1},\dots,B_{i_m}}(h) : i_j \in \{1, \dots, V\}, 1 \le i_1 < \dots < i_m \le V \} \, . \tag{4.6}$$

Note that, mostly in order to simplify notation, we only take those values of $U_{B_{i_1},\dots,B_{i_m}}(h)$ into account that correspond to distinct indices $i_1 < \cdots < i_m$. Thus, each $U_{B_{i_1},\dots,B_{i_m}}(h)$ is a so-called decoupled U-statistics (see the Appendix for the definition). One may incorporate all m-tuples (not necessarily with distinct indices) in the computation of the median. However, this has a minor effect on the performance. Similar bounds may be proven though with a more complicated notation.

A simpler alternative is obtained by taking only "diagonal" blocks into account. More precisely, let $U_{B_i}(h)$ be the U-statistics calculated using the variables in block B_i (as defined in (4.1)). One may simply calculate the median of the V different U-statistics $U_{B_i}(h)$. This version is easy to analyze because $\left| \{ i \le V \, : \, U_{B_i}(h) \ge b \} \right|$ is a sum of independent random variables. However, this simple version is wasteful in the sense that only a small fraction of possible m-tuples are taken into account.

In the next two sections we analyze the performance of the estimator $\overline{U}_{\mathcal{B}}(h)$.

4.2.1 Exponential inequalities for P-degenerate kernels with finite variance.

Next we present a performance bound of the estimator $\overline{U}_{\mathcal{B}}(h)$ in the case when σ^2 is finite. The somewhat more complicated case of infinite second moment is treated in Section 4.2.2.

Theorem 38. *Let $X_1, \dots, X_n$ be i.i.d. random variables taking values in $\mathcal{X}$. Let $h : \mathcal{X}^m \mapsto \mathbb{R}$ be a symmetric kernel that is P-degenerate of order $q-1$. Assume Var $(h(X_1, \dots, X_m)) = \sigma^2 < \infty$. Let $\delta \in (0, \frac{1}{2})$ be such that $\lceil \log(1/\delta) \rceil \le \frac{n}{64m}$. Let $\mathcal{B}$ be a regular partition of $\{1, \dots, n\}$ with $|\mathcal{B}| = 32m \lceil \log(1/\delta) \rceil$. Then, with probability at least $1 - 2\delta$, we have*

$$\left| \overline{U}_{\mathcal{B}}(h) - m_h \right| \le K_m \sigma \left(\frac{\lceil \log(1/\delta) \rceil}{n} \right)^{q/2} , \tag{4.7}$$

where $K_m = 2^{\frac{7}{2}m+1} m^{\frac{m}{2}}$.

When $q = m$, the kernel $h - m_h$ is P-canonical and the rate of convergence is then given by $(\log \delta^{-1}/n)^{m/2}$. Thus, the new estimator has a performance similar to standard U-statistics as in (4.3) and (4.4) but without the boundedness assumption for the kernel. It is important to note that a disadvantage of the estimator $\overline{U}_{\mathcal{B}}(h)$ is that it depends on the confidence level δ (through the number of blocks). For different confidence levels, different estimators are used.

Because of its importance in applications, we spell out the special case when $m = q = 2$. In Section 4.3 we use this result in an example of cluster analysis.

Corollary 39. *Let $\delta \in (0, 1/2)$. Let $h : \mathcal{X}^2 \mapsto \mathbb{R}$ be a P-canonical kernel with $\sigma^2 =$ Var $(h(X_1, X_2))$ and let $n \ge 128(1 + \log(1/\delta))$. Then, with probability at least $1 - 2\delta$,*

$$|\overline{U}_{\mathcal{B}}(h) - m_h| \le 512\sigma \frac{1 + \log(1/\delta)}{n} . \tag{4.8}$$

In the proof of Theorem 38 we need the notion of *Hoeffding decomposition* [30] of U-statistics. For probability measures $P_1, \dots, P_m$, define $P_1 \times \cdots \times P_m h = \int h \, d(P_1, \dots, P_m)$. For a symmetric kernel $h : \mathcal{X}^m \mapsto \mathbb{R}$ the *Hoeffding projections* are defined, for $0 \le k \le m$ and $x_1, \dots, x_k \in \mathcal{X}$, as

$$\pi_k h(x_1, \dots, x_k) := (\delta_{x_1} - P) \times \cdots \times (\delta_{x_k} - P) \times P^{m-k} h$$

where δ_x denotes the Dirac measure at the point x. Observe that $\pi_0 h = P^m h$ and for $k > 0$, $\pi_k h$ is a P-canonical kernel. h can be decomposed as

$$h(x_1, \dots, x_m) = \sum_{k=0}^{m} \sum_{1 \le i_1 < \dots < i_k \le m} \pi_k h(x_{i_1}, \dots, x_{i_k}) . \tag{4.9}$$

If h is assumed to be square-integrable (i.e., $P^m h^2 < \infty$), the terms in (4.9) are orthogonal. If h is degenerate of order $q-1$, then for any $1 \le k \le q-1$, $\pi_k h = 0$.

Proof of Theorem 38. We begin with a "weak" concentration result on each $U_{B_{i_1},\dots,B_{i_m}}(h)$. Let $B_{i_1},\dots,B_{i_m}$ be elements of $\mathcal{B}$. For any $B\in\mathcal{B}$, we have $\frac{n}{2|\mathcal{B}|}\le|B|\le\frac{2n}{|\mathcal{B}|}$. We denote by $\mathbf{k}=(k_1,\dots,k_m)$ an element of $I_{B_{i_1},\dots,B_{i_m}}$. We have, by the above-mentioned orthogonality property,

$$\begin{aligned}
&\mathrm{Var}\left(U_{B_{i_1},\dots,B_{i_m}}(h)\right)\\
&\quad=\ \mathbb{E}\left[(U_{B_{i_1},\dots,B_{i_m}}(h)-P^m h)^2\right]\\
&\quad=\ \frac{1}{|B_{i_1}|^2\dots|B_{i_m}|^2}\sum_{\substack{\mathbf{k}\in I_{B_{i_1},\dots,B_{i_m}}\\ \mathbf{l}\in I_{B_{i_1},\dots,B_{i_m}}}}\mathbb{E}\left[(h(X_{k_1},\dots,X_{k_m})-P^m h)(h(X_{l_1},\dots,X_{l_m})-P^m h)\right]\\
&\quad=\ \frac{1}{|B_{i_1}|^2\dots|B_{i_m}|^2}\sum_{\substack{\mathbf{k}\in I_{B_{i_1},\dots,B_{i_m}}\\ \mathbf{l}\in I_{B_{i_1},\dots,B_{i_m}}}}\sum_{s=q}^{m}\binom{|\mathbf{k}\cap\mathbf{l}|}{s}\mathbb{E}\left[\pi_s h(X_1,\dots,X_s)^2\right]\quad\text{(by orthogonality)}\\
&\quad\le\ \frac{1}{|B_{i_1}|^2\dots|B_{i_m}|^2}\sum_{\mathbf{k}\in I_{B_{i_1},\dots,B_{i_m}}}\sum_{s=q}^{m}\sum_{t=0}^{m}\binom{t}{s}\mathbb{E}\left[\pi_s h(X_1,\dots,X_s)^2\right]\times\left(\frac{2n}{|\mathcal{B}|}\right)^{m-t}.
\end{aligned}$$

The last inequality is obtained by counting, for any fixed $\mathbf{k}$ and t, the number of elements $\mathbf{l}$ such that $|\mathbf{k}\cap\mathbf{l}|=t$. Thus,

$$\begin{aligned}
\mathrm{Var}\left(U_{B_{i_1},\dots,B_{i_m}}(h)\right)\ &\le\ \frac{1}{|B_{i_1}|\dots|B_{i_m}|}\sum_{s=q}^{m}\sum_{t=q}^{m}\binom{t}{s}\mathbb{E}\left[\pi_s h(X_1,\dots,X_s)^2\right]\times\left(\frac{2n}{|\mathcal{B}|}\right)^{m-t}\\
&\le\ \frac{1}{|B_{i_1}|\dots|B_{i_m}|}\sum_{s=q}^{m}\binom{m}{s}\mathbb{E}\left[\pi_s h(X_1,\dots,X_s)^2\right]\times\sum_{t=q}^{m}\left(\frac{2n}{|\mathcal{B}|}\right)^{m-t}\\
&\le\ \frac{1}{\left(\frac{n}{2|\mathcal{B}|}\right)^m}\sum_{s=q}^{m}\binom{m}{s}\mathbb{E}\left[\pi_s h(X_1,\dots,X_s)^2\right]\times 2\left(\frac{2n}{|\mathcal{B}|}\right)^{m-q}\\
&\le\ \frac{2^{2m-q+1}|\mathcal{B}|^q}{n^q}\sum_{s=q}^{m}\binom{m}{s}\mathbb{E}\left[\pi_s h(X_1,\dots,X_s)^2\right].
\end{aligned}$$

On the other hand, we have, by (4.9),

$$\begin{aligned}
\mathrm{Var}\,(h)\ &=\ \mathbb{E}\left[\left(\sum_{s=q}^{m}\sum_{1\le i_1<\dots<i_s\le m}\pi_s h(X_{i_1},\dots,X_{i_s})\right)^2\right]\\
&=\ \sum_{s=q}^{m}\sum_{1\le i_1<\dots<i_s\le m}\mathbb{E}\left[\left(\pi_s h(X_{i_1},\dots,X_{i_s})\right)^2\right]\\
&=\ \sum_{s=q}^{m}\binom{m}{s}\mathbb{E}\left[(\pi_s h(X_1,\dots,X_s))^2\right].
\end{aligned}$$

Combining the two displayed equations above,

$$\mathrm{Var}\left(U_{B_{i_1},\dots,B_{i_m}}(h)\right) \;\leq\; \frac{2^{2m-q+1}|\mathcal{B}|^q}{n^q}\sigma^2 \leq \frac{2^{2m}|\mathcal{B}|^q}{n^q}\sigma^2 \,.$$

By Chebyshev's inequality, for all $r \in (0,1)$,

$$\mathbb{P}\left(U_{B_{i_1},\dots,B_{i_m}}(h) - P^m h > 2^m \sigma \frac{|\mathcal{B}|^{q/2}}{n^{q/2} r^{1/2}}\right) \leq r \,. \tag{4.10}$$

We set $x = 2^m \sigma \frac{|\mathcal{B}|^{q/2}}{n^{q/2} r^{1/2}}$, and

$$N_x = \left|\left\{(i_1,\dots,i_m) \in \{1,\dots,V\}^m : 1 \leq i_1 < \dots < i_m \leq |B|,\; U_{B_{i_1},\dots,B_{i_m}}(h) - P^m h > x\right\}\right| \,.$$

The random variable $\frac{1}{\binom{|\mathcal{B}|}{m}} N_x$ is a U-statistics of order m with the symmetric kernel $g : (i_1,\dots,i_m) \mapsto \mathbb{1}_{\{U_{B_{i_1},\dots,B_{i_m}}(h) - P^m h > x\}}$. Thus, Hoeffding's inequality for centered U-statistics (4.2) gives

$$\mathbb{P}\left(N_x - \mathbb{E}N_x \geq t\binom{|\mathcal{B}|}{m}\right) \leq \exp\left(-\frac{|\mathcal{B}|t^2}{2m}\right) \,. \tag{4.11}$$

By (4.10) we have $\mathbb{E}N_x \leq \binom{|\mathcal{B}|}{m} r$. Taking $t = r = \frac{1}{4}$ in (4.11), by the definition of the median, we have

$$\begin{aligned}\mathbb{P}\left(\overline{U}_{\mathcal{B}}(h) - P^m(h) > x\right) &\leq \mathbb{P}\left(N_x \geq \frac{\binom{|\mathcal{B}|}{m}}{2}\right) \\ &\leq \exp\left(-\frac{|\mathcal{B}|}{32m}\right) \,.\end{aligned}$$

Since $|\mathcal{B}| \geq 32m\log(\delta^{-1})$, with probability at least $1-\delta$, we have

$$\overline{U}_{\mathcal{B}}(h) - P^m h \leq K_m \sigma \left(\frac{\left\lceil \log \delta^{-1} \right\rceil}{n}\right)^{q/2}$$

with $K_m = 2^{\frac{7}{2}m+1} m^{\frac{m}{2}}$. The upper bound for the lower tail holds by the same argument. □

4.2.2 Bounded moment of order p with $1 < p \leq 2$

In this section, we weaken the assumption of finite variance and only assume the existence of a centered moment of order p for some $1 < p \leq 2$. The outline of the argument is similar as in the case of finite variance. First we obtain a "weak" concentration inequality for the U-statistics is each block and then use the property of the median

to boost the weak inequality. While for the case of finite variance weak concentration could be proved by a direct calculation of the variance, here we need the randomization inequalities for convex functions of U-statistics established by [24] and [6]. Note that, here, a P-canonical technical assumption is needed.

Theorem 40. *Let h be a symmetric kernel of order m such that $h - m_h$ is P-canonical. Assume that $M_p := \mathbb{E}\left[\left|h(X_1,\dots,X_m) - m_h\right|^p\right]^{1/p} < \infty$ for some $1 < p \leq 2$. Let $\delta \in (0, \frac{1}{2})$ be such that $\left\lceil \log(\delta^{-1}) \right\rceil \leq \frac{n}{64m}$. Let $\mathcal{B}$ be a regular partition of $\{1,\dots,n\}$ with $|\mathcal{B}| = 32m\left\lceil \log(\delta^{-1}) \right\rceil$. Then, with probability at least $1 - 2\delta$, we have*

$$\left|\overline{U}_{\mathcal{B}}(h) - m_h\right| \leq K_m M_p \left(\frac{\left\lceil \log(\delta^{-1}) \right\rceil}{n}\right)^{m(p-1)/p} \tag{4.12}$$

where $K_m = 2^{4m+1} m^{\frac{m}{2}}$.

Proof. Define the centered version of h by $g(x_1,\dots,x_m) := h(x_1,\dots,x_m) - m_h$. Let $\varepsilon_1,\dots,\varepsilon_n$ be i.i.d. Rademacher random variables (i.e., $\mathbb{P}(\varepsilon_1 = -1) = \mathbb{P}(\varepsilon_1 = 1) = 1/2$) independent of $X_1,\dots,X_n$. By the randomization inequalities (see Theorem 3.5.3 in [23] and also Theorem 46 in the Appendix), we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{(k_1,\dots,k_m)\in I_{B_{i_1},\dots,B_{i_m}}} g(X_{k_1},\dots,X_{k_m})\right|^p\right] \\
&\leq 2^{mp}\mathbb{E}_X\mathbb{E}_\varepsilon\left[\left|\sum_{(k_1,\dots,k_m)\in I_{B_{i_1},\dots,B_{i_m}}} \varepsilon_{k_1}\dots\varepsilon_{k_m} g(X_{k_1},\dots,X_{k_m})\right|^p\right] \qquad (4.13)\\
&\leq 2^{mp}\mathbb{E}_X\left[\left|\mathbb{E}_\varepsilon\left[\left(\sum_{(k_1,\dots,k_m)\in I_{B_{i_1},\dots,B_{i_m}}} \varepsilon_{k_1}\dots\varepsilon_{k_m} g(X_{k_1},\dots,X_{k_m})\right)^2\right]\right|^{p/2}\right] \\
&= 2^{mp}\mathbb{E}_X\left[\left|\sum_{(k_1,\dots,k_m)\in I_{B_{i_1},\dots,B_{i_m}}} g(X_{k_1},\dots,X_{k_m})^2\right|^{p/2}\right] \\
&\leq 2^{mp}\sum_{(k_1,\dots,k_m)\in I_{B_{i_1},\dots,B_{i_m}}} \mathbb{E}|g(X_{k_1},\dots,X_{k_m})|^p \\
&= 2^{mp}|B_{i_1}|\cdots|B_{i_m}|\mathbb{E}|g|^p\,. \qquad (4.14)
\end{aligned}$$

Thus, we have $\mathbb{E}\left[|U_{B_{i_1},\dots,B_{i_m}}(h) - m_h|^p\right] \leq 2^{mp}(|B_{i_1}|\dots|B_{i_m}|)^{1-p}\mathbb{E}|g|^p$ and by Markov's in-

equality,

$$\mathbb{P}\left(U_{B_{i_1},\dots,B_{i_m}}(h) - m_h > \frac{2^m M_p}{r^{\frac{1}{p}}} \left(\frac{n}{(2|\mathcal{B}|)} \right)^{m\frac{1-p}{p}} \right) \leq r \,. \tag{4.15}$$

Another use of (4.11) with $t = r = \frac{1}{4}$ gives

$$\overline{U}_{\mathcal{B}}(h) - P^m h \leq 2^{4m+1} m^{\frac{m}{2}} M_p \left(\frac{\left\lceil \log \delta^{-1} \right\rceil}{n} \right)^{m\frac{p-1}{p}} \,.$$

□

To see why the bound of Theorem 40 gives essentially the right order of magnitude, consider again the example described in the introduction, when $m = 2$, $h(X_1, X_2) = X_1 X_2$, and the X_i have an *α-stable* law $S(\gamma, \alpha)$ for some $\gamma > 0$ and $1 < \alpha \leq 2$. Note that an α-stable random variable has finite moments up to (but not including) α and therefore we may take any $p = \alpha - \epsilon$ for any $\epsilon \in (0, 1 - \alpha)$. As we noted it in the introduction, there exists a constant c depending on α and γ only such that for all $1 \leq i_1 < i_2 \leq V$,

$$\mathbb{P}\left(\left| U_{B_{i_1},B_{i_2}}(h) - m_h \right| \geq c \left(\frac{n}{|\mathcal{B}|} \right)^{2/\alpha - 2} \right) \geq 2/3 \,,$$

and therefore (4.15) is essentially the best rate one can hope for.

4.2.3 Generalization to Hilbert space valued kernels

This section is dedicated to the more general context of kernels taking values in a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$. The notions of *symmetric* kernels, *P-canonical* kernels and *Hoeffding decomposition* can be defined for a measurable function $h : \mathcal{X}^m \mapsto \mathcal{H}$ by the same formulas than in Section 4.1. Let $\mathcal{B}$ be a regular partition of $\{1, \dots, n\}$. For any $B_{i_1}, \dots, B_{i_m}$ in $\mathcal{B}$, we set

$$I_{B_{i_1},\dots,B_{i_m}} = \left\{ (k_1, \dots, k_m) : k_j \in B_{i_j} \right\}$$

and

$$U_{B_{i_1},\dots,B_{i_m}}(h) = \frac{1}{|B_{i_1}| \cdots |B_{i_m}|} \sum_{(k_1,\dots,k_m) \in I_{B_{i_1},\dots,B_{i_m}}} h(X_{k_1}, \dots, X_{k_m}) \,.$$

We define an estimator of the mean m_h in $\mathcal{H}$ by

$$\overline{U}_{\mathcal{B}}(h) = \mathrm{GeoMed}\{ U_{B_{i_1},\dots,B_{i_m}}(h) : i_j \in \{1, \dots, |\mathcal{B}|\}, 1 \leq i_1 < \dots < i_m \leq |\mathcal{B}| \}$$

where GeoMed refers to the geometric median defined in Section 2.3.1. We now state a generalization of Theorem 38.

Theorem 41. *Let $X_1, \dots, X_n$ be i.i.d. random variables taking values in $\mathcal{X}$. Let $h : \mathcal{X}^m \mapsto \mathcal{H}$ be a symmetric kernel that is P-degenerate of order $q-1$. Assume $\mathbb{E}\left[\|h(X_1, \dots, X_m) - m_h\|^2\right] = \sigma^2 < \infty$. Let $\delta \in (0, \frac{1}{2})$ be such that $\lceil \log(1/\delta) \rceil \leq \frac{n}{64m}$. Let $\mathcal{B}$ be a regular partition of $\{1, \dots, n\}$ with $|\mathcal{B}| = 32m\lceil \log(1/\delta) \rceil$. Then, with probability at least $1 - \delta$, we have*

$$\left\| \overline{U}_{\mathcal{B}}(h) - m_h \right\| \leq K_m \sigma \left(\frac{\lceil \log(1/\delta) \rceil}{n} \right)^{q/2} , \tag{4.16}$$

where $K_m = 2^{\frac{7}{2}m+2} m^{\frac{m}{2}}$.

Proof. From the proof of Theorem 38, we obtain that

$$\mathbb{E}\left[\left\| U_{B_{i_1}, \dots, B_{i_m}}(h) - m_h \right\|^2 \right] \leq \frac{2^{2m-q+1} |\mathcal{B}|^q}{n^q} \sum_{s=q}^{m} \binom{m}{s} \mathbb{E}\left[\|\pi_s h(X_1, \dots, X_s)\|^2 \right]$$

and

$$\mathbb{E}\left[\|h(X_1, \dots, X_m) - m_h\|^2 \right] = \sum_{s=q}^{m} \binom{m}{s} \mathbb{E}\left[\|\pi_s h(X_1, \dots, X_s)\|^2 \right] .$$

Then, by Markov's inequality, for all $r \in (0, 1)$,

$$\mathbb{P}\left(\left\| U_{B_{i_1}, \dots, B_{i_m}}(h) - m_h \right\| > 2^m \sigma \frac{|\mathcal{B}|^{q/2}}{n^{q/2} r^{1/2}} \right) \leq r .$$

Let $\alpha \in (0, 1/2)$ and let

$$\varepsilon = 2^m \sigma \frac{|\mathcal{B}|^{q/2}}{n^{q/2} r^{1/2}} .$$

Assume that event $\mathcal{E} := \left\{ \left\| \overline{U}_{\mathcal{B}}(h) - m_h \right\| > C_\alpha \varepsilon \right\}$ occurs where C_α is defined as in Lemma 23. By Lemma 23, there exists a subset $J \subseteq \left\{ (i_j)_{1 \leq j \leq m} \in \{1, \dots, |\mathcal{B}|\} \, : \, 1 \leq i_1 < \dots < i_m \leq |\mathcal{B}| \right\}$ of cardinality $|J| \geq \alpha \binom{|\mathcal{B}|}{m}$ such that $\left\| U_{B_{i_1}, \dots, B_{i_m}}(h) - m_h \right\| > \varepsilon$ for all $(i_j)_{1 \leq j \leq m} \in J$. We set

$$N_\varepsilon = \left| \left\{ (i_1, \dots, i_m) \in \{1, \dots, V\}^m : 1 \leq i_1 < \dots < i_m \leq |B|, \, \left\| U_{B_{i_1}, \dots, B_{i_m}}(h) - P^m h \right\| > \varepsilon \right\} \right| .$$

Hence,

$$\mathbb{P}(\mathcal{E}) \leq \mathbb{P}\left(N_\varepsilon > \alpha \binom{|\mathcal{B}|}{m} \right) .$$

The random variable $\frac{1}{\binom{|\mathcal{B}|}{m}} N_\varepsilon$ is a U-statistics and (4.2) implies that, for any $t > 0$,

$$\mathbb{P}\left(N_\varepsilon - \mathbb{E}N_\varepsilon \geq t \binom{|\mathcal{B}|}{m} \right) \leq \exp\left(-\frac{|\mathcal{B}| t^2}{2m} \right)$$

with $\mathbb{E}N_\varepsilon \le r\binom{|\mathcal{B}|}{m}$. We set $\alpha = 5/12$, $r = 1/6$ and $t = 1/4$ then

$$\mathbb{P}\left(N_\varepsilon > \alpha\binom{|\mathcal{B}|}{m}\right) \le \exp\left(-\frac{|\mathcal{B}|}{32m}\right) \le \delta\,.$$

In that case, $\frac{C_\alpha}{r^{1/2}} = \frac{42}{12} \le 4$ and the constant $K_m = 2^{\frac{7}{2}m+2}m^{\frac{m}{2}}$ is justified. □

4.3 Cluster analysis with U-statistics

In this section we illustrate the use of the proposed mean estimator in a clustering problem when the presence of possibly heavy-tailed data requires robust techniques.

We consider the general statistical framework defined by [21], described as follows: Let X, X' be i.i.d. random variables taking values in $\mathcal{X}$ where typically but not necessarily, $\mathcal{X}$ is a subset of $\mathbb{R}^d$). For a partition $\mathcal{P}$ of $\mathcal{X}$ into K disjoint sets–the so-called "cells"–, define $\Phi_\mathcal{P}(x, x') = \sum_{C\in\mathcal{P}} \mathbb{1}_{\{(x,x')\in C^2\}}$ the $\{0,1\}$-valued function that indicates whether two elements x and x' belong to the same cell C. Given a *dissimilarity measure* $D : \mathcal{X}^2 \to \mathbb{R}_+^*$, the clustering task consists in finding a partition of $\mathcal{X}$ minimizing the *clustering risk*

$$W(\mathcal{P}) = \mathbb{E}\left[D(X, X')\Phi_\mathcal{P}(X, X')\right]\,.$$

Let Π_K be a finite class of partitions $\mathcal{P}$ of $\mathcal{X}$ into K cells and define $W^* = \min_{\mathcal{P}\in\Pi_K} W(\mathcal{P})$.

Given $X_1, \dots, X_n$ be i.i.d. random variables distributed as X, the goal is to find a partition $\mathcal{P} \in \Pi_K$ with risk as close to W^* as possible. A natural idea–and this is the approach of [21]–is to estimate $W(\mathcal{P})$ by the U-statistics

$$\widehat{W}_n(\mathcal{P}) = \frac{2}{n(n-1)} \sum_{1\le i<j\le n} D(X_i, X_j)\Phi_\mathcal{P}(X_i, X_j)$$

and choose a partition minimizing the empirical clustering risk $\widehat{W}_n(\mathcal{P})$. [21] uses the theory of U-processes to analyze the performance of such minimizers of U-statistics. However, in order to control uniform deviations of the form $\sup_{\mathcal{P}\in\Pi_K} |\widehat{W}_n(\mathcal{P}) - W(\mathcal{P})|$, exponential concentration inequalities are needed for U-statistics. This restricts one to consider bounded dissimilarity measures $D(X, X')$. When $D(X, X')$ may have a heavy tail, we propose to replace U-statistics by the median-of-means estimators of $W(\mathcal{P})$ introduced in this paper.

Let $\mathcal{B}$ be a regular partition of $\{1, \dots, n\}$ and define the median-of-means estimator $\overline{W}_\mathcal{B}(\mathcal{P})$ of $W(\mathcal{P})$ as in (4.6). Then Theorem 38 applies and we have the following simple corollary.

Corollary 42. *Let Π_K be a class of partitions of cardinality $|\Pi_K| = N$. Assume that $\sigma^2 := \mathbb{E}\left[D(X_1, X_2)^2\right] < \infty$. Let $\delta \in (0, 1/2)$ such that $n \ge 128\lceil\log(N/\delta)\rceil$. Let $\mathcal{B}$ be a regular*

partition of $\{1, \dots, n\}$ *with* $|\mathcal{B}| = 64 \lceil \log(N/\delta) \rceil$. *Then there exists a constant* C *such that, with probability at least* $1 - 2\delta$,

$$\sup_{\mathcal{P} \in \Pi_K} |\overline{W}_{\mathcal{B}}(\mathcal{P}) - W(\mathcal{P})| \leq C\sigma \left(\frac{\lceil \log(N/\delta) \rceil}{n} \right)^{1/2} . \tag{4.17}$$

Proof. Since $\Phi_{\mathcal{P}}(x, x')$ is bounded by 1, $\mathrm{Var}\left(D(X_1, X_2)\Phi_{\mathcal{P}}(X_1, X_2)\right) \leq \mathbb{E}\left[D(X_1, X_2)^2\right]$. For a fixed $\mathcal{P} \in \Pi_K$, Theorem 38 applies with $m = 2$ and $q = 1$. The inequality follows from the union bound. □

Once uniform deviations of $\overline{W}_{\mathcal{B}}(\mathcal{P})$ from its expected value are controlled, it is a routine exercise to derive performance bounds for clustering based on minimizing $\overline{W}_{\mathcal{B}}(\mathcal{P})$ over $\mathcal{P} \in \Pi_K$.

Let $\widehat{\mathcal{P}} = \mathrm{argmin}_{\mathcal{P} \in \Pi_K} \overline{W}_{\mathcal{B}}(\mathcal{P})$ denote the empirical minimizer. (In case of multiple minimizers, one may select one arbitrarily.) Now for any $\mathcal{P}_0 \in \Pi_K$,

$$\begin{aligned} W(\widehat{\mathcal{P}}) - W^* &= W(\widehat{\mathcal{P}}) - \overline{W}_{\mathcal{B}}(\widehat{\mathcal{P}}) + \overline{W}_{\mathcal{B}}(\widehat{\mathcal{P}}) - W^* \\ &\leq W(\widehat{\mathcal{P}}) - \overline{W}_{\mathcal{B}}(\widehat{\mathcal{P}}) + \overline{W}_{\mathcal{B}}(\mathcal{P}_0) - W(\mathcal{P}_0) + W(\mathcal{P}_0) - W^* \\ &\leq 2 \sup_{\mathcal{P} \in \Pi_K} |\overline{W}_{\mathcal{B}}(\mathcal{P}) - W(\mathcal{P})| + W(\mathcal{P}_0) - W^* . \end{aligned}$$

Taking the infimum over Π_K,

$$W(\widehat{\mathcal{P}}) - W^* \leq 2 \sup_{\mathcal{P} \in \Pi_K} |\overline{W}_{\mathcal{B}}(\mathcal{P}) - W(\mathcal{P})| . \tag{4.18}$$

Finally, (4.17) implies that

$$W(\widehat{\mathcal{P}}) - W^* \leq 2C\sigma \left(\frac{1 + \log(N/\delta)}{n} \right)^{1/2} .$$

This result is to be compared with Theorem 2 of [21]. Our result holds under the only assumption that $D(X, X')$ has a finite second moment. (This may be weakened to assuming the existence of a finite p-th moment for some $1 < p \leq 2$ by using Theorem 40). On the other hand, our result holds only for a finite class of partitions while [21] uses the theory of U-processes to obtain more sophisticated bounds for uniform deviations over possibly infinite classes of partitions. It remains a challenge to develop a theory to control processes of median-of-means estimators–in the style of [6]–and not having to resort to the use of simple union bounds.

In the rest of this section we show that, under certain "low-noise" assumptions, analogous to the ones introduced by [46] in the context of classification, to obtain faster rates of convergence. In this part we need bounds for P-canonical kernels and use the full power of Corollary 39. Similar arguments for the study of minimizing U-statistics appear in [22], [21].

We assume the following conditions, also considered by [21]:

1. There exists $\mathcal{P}^*$ such that $W(\mathcal{P}^*) = W^*$
2. There exist $\alpha \in [0,1]$ and $\kappa < \infty$ such that for all $\mathcal{P} \in \Pi_K$ and for all $x \in \mathcal{X}$,

$$\mathbb{P}\left(\Phi_{\mathcal{P}}(x, X) \neq \Phi_{\mathcal{P}^*}(x, X)\right) \leq \kappa (W(\mathcal{P}) - W^*)^{\alpha} .$$

Note that $\alpha \leq 2$ since by the Cauchy-Schwarz inequality,

$$W(\mathcal{P}) - W^* \leq \mathbb{E}\left[D(X_1, X_2)^2\right]^{1/2} \mathbb{P}\left(\Phi_{\mathcal{P}}(X_1, X_2) \neq \Phi_{\mathcal{P}^*}(X_1, X_2)\right)^{1/2} .$$

Corollary 43. *Assume the conditions above and that* $\sigma^2 := \mathbb{E}\left[D(X_1, X_2)^2\right] < \infty$. *Let* $\delta \in (0, 1/2)$ *such that* $n \geq 128 \lceil \log(N/\delta) \rceil$. *Let* $\mathcal{B}$ *be a regular partition of* $\{1, \ldots, n\}$ *with* $|\mathcal{B}| = 64 \lceil \log(N/\delta) \rceil$. *Then there exists a constant* C *such that, with probability at least* $1 - 2\delta$,

$$W(\widehat{\mathcal{P}}) - W^* \leq C\sigma^{2/(2-\alpha)} \left(\frac{\lceil \log(N/\delta) \rceil}{n} \right)^{1/(2-\alpha)} . \tag{4.19}$$

The proof Corollary 43 is postponed to the Appendix.

4.4 Appendix

4.4.1 Decoupling and randomization

Here we summarize some of the key tools for analyzing U-statistics that we use in the paper. For an excellent exposition we refer to [23].

Let $\{X_i\}$ be i.i.d. random variables taking values in $\mathcal{X}$ and let $\{X_i^k\}$, $k = 1, \ldots, m$, be sequences of independent copies. Let Φ be a non-negative function. As a corollary of Theorem 3.1.1 in [23] we have the following:

Theorem 44. *Let* $h : \mathcal{X}^m \to \mathbb{R}$ *be a measurable function with* $\mathbb{E}|h(X_1, \ldots, X_m)| < \infty$. *Let* $\Phi : [0, \infty) \to [0, \infty)$ *be a convex nondecreasing function such that* $\mathbb{E}\Phi\left(|h(X_1, \ldots, X_m)|\right) < \infty$. *Then*

$$\mathbb{E}\Phi\left(\left| \sum_{I_n^m} h(X_{i_1}, \ldots, X_{i_m}) \right| \right) \leq \mathbb{E}\Phi\left(C_m \left| \sum_{I_n^m} h(X_{i_1}^1, \ldots, X_{i_m}^m) \right| \right)$$

where $C_m = 2^m(m^m - 1)((m-1)^{m-1} - 1) \times \cdots \times 3$. *Moreover, if the kernel* h *is symmetric, then,*

$$\mathbb{E}\Phi\left(c_m \left| \sum_{I_n^m} h(X_{i_1}^1, \ldots, X_{i_m}^m) \right| \right) \leq \mathbb{E}\Phi\left(\left| \sum_{I_n^m} h(X_{i_1}, \ldots, X_{i_m}) \right| \right)$$

where $c_m = 1/(2^{2m-2}(m-1)!)$.

An equivalent result for tail probabilities of U-statistics is the following (see Theorem 3.4.1 in [23]):

Theorem 45. *Under the same hypotheses as Theorem 44, there exists a constant C_m depending on m only such that, for all $t > 0$,*

$$\mathbb{P}\left(\left|\sum_{I_n^m} h(X_{i_1}, \dots, X_{i_m})\right| > t\right) \leq C_m \mathbb{P}\left(C_m \left|\sum_{I_n^m} h(X_{i_1}^1, \dots, X_{i_m}^m)\right| > t\right).$$

If moreover, the kernel h is symmetric then there exists a constant c_m depending on m only such that, for all $t > 0$,

$$c_m \mathbb{P}\left(c_m \left|\sum_{I_n^m} h(X_{i_1}^1, \dots, X_{i_m}^m)\right| > t\right) \leq \mathbb{P}\left(\left|\sum_{I_n^m} h(X_{i_1}, \dots, X_{i_m})\right| > t\right).$$

The next Theorem is a direct corollary of Theorem 3.5.3 in [23].

Theorem 46. *Let $1 < p \leq 2$. Let $(\varepsilon_i)_{i \leq n}$ be i.i.d Rademacher random variables independent of the $(X_i)_{i \leq n}$. Let $h : \mathcal{X} \to \mathbb{R}$ be a P-degenerate measurable function such that $\mathbb{E}\left(|h(X_1, \dots, X_m)|^p\right) < \infty$. Then*

$$\begin{aligned} c_m \mathbb{E}\left|\sum_{I_n^m} \varepsilon_{i_1} \dots \varepsilon_{i_m} h(X_{i_1}, \dots, X_{i_m})\right|^p &\leq \mathbb{E}\left|\sum_{I_n^m} h(X_{i_1}, \dots, X_{i_m})\right|^p \\ &\leq C_m \mathbb{E}\left|\sum_{I_n^m} \varepsilon_{i_1} \dots \varepsilon_{i_m} h(X_{i_1}, \dots, X_{i_m})\right|^p, \end{aligned}$$

where $C_m = 2^{mp}$ and $c_m = 2^{-mp}$.

The same conclusion holds for decoupled U-statistics.

4.4.2 α-stable distributions

Proposition 47. *Let $\alpha \in (0, 2)$. Let $X_1, \dots, X_n$ be i.i.d. random variables of law $S(\gamma, \alpha)$. Let $f_{\gamma,\alpha} : x \mapsto \mathbb{R}$ be the density function of X_1. Let $S_n = \sum_{1 \leq i \leq n} X_i$. Then*

(i) $f_{\gamma,\alpha}(x)$ is an even function.

(ii) $f_{\gamma,\alpha}(x) \underset{x \to +\infty}{\sim} \alpha \gamma^\alpha c_\alpha x^{-\alpha-1}$ with $c_\alpha = \sin\left(\frac{\pi\alpha}{2}\right)\Gamma(\alpha)/\pi$.

(iii) $\mathbb{E}\left[X_1^p\right]$ is finite for any $p < \alpha$ and is infinite whenever $p \geq \alpha$.

(iv) S_n has a α-stable law $S(\gamma n^{1/\alpha}, \alpha)$.

Proof. (i) and (iv) follow directly from the definition. (ii) is proved in the introduction of [68]. (iii) is a consequence of (ii). □

4.4.3 Proof of Corollary 43

Define $\Lambda_n(\mathcal{P}) = \widehat{W}_n(\mathcal{P}) - W^*$, the U-statistics based on the sample $X_1, \dots, X_n$, with symmetric kernel

$$h_{\mathcal{P}}(x, x') = D(x, x') \left(\Phi_{\mathcal{P}}(x, x') - \Phi_{\mathcal{P}^*}(x, x')\right) .$$

We denote by $\Lambda(\mathcal{P}) = W(\mathcal{P}) - W^*$ the expected value of $\Lambda_n(\mathcal{P})$. The main argument in the following analysis is based on the Hoeffding decomposition. For all partitions $\mathcal{P}$,

$$\Lambda_n(\mathcal{P}) - \Lambda(\mathcal{P}) = 2L_n(\mathcal{P}) + M_n(\mathcal{P})$$

for $L_n(\mathcal{P}) = \frac{1}{n}\sum_{i\le n} h^{(1)}(X_i)$ with $h^{(1)}(x) = \mathbb{E}\left[h_{\mathcal{P}}(X, x)\right] - \Lambda(\mathcal{P})$ and $M_n(\mathcal{P})$ the U-statistics based on the canonical kernel given by $h^{(2)}(x, x') = h_{\mathcal{P}}(x, x') - h^{(1)}(x) - h^{(1)}(x') - \Lambda(\mathcal{P})$. Let $\mathcal{B}$ be a regular partition of $\{1, \dots, n\}$. For any $B \in \mathcal{B}$, $\Lambda_B(\mathcal{P})$ is the U-statistics on the kernel $h_{\mathcal{P}}$ restricted to the set B and $\overline{\Lambda}_B(\mathcal{P})$ is the median of the sequence $(\Lambda_B(\mathcal{P}))_{B\in\mathcal{B}}$. We define similarly $L_B(\mathcal{P})$ and $M_B(\mathcal{P})$ on the variables $(X_i)_{i\in B}$. For any $B \in \mathcal{B}$,

$$\begin{aligned}
\mathrm{Var}\left(\Lambda_B(\mathcal{P})\right) &= 4\mathrm{Var}\left(L_B(\mathcal{P})\right) + \mathrm{Var}\left(M_B(\mathcal{P})\right) \\
&= \frac{4}{|B|}\mathrm{Var}\left(h^{(1)}(X)\right) + \frac{2}{|B|(|B|-1)}\mathrm{Var}\left(h^{(2)}(X_1, X_2)\right) .
\end{aligned}$$

Simple computations show that $\mathrm{Var}\left(h^{(2)}(X_1, X_2)\right) = 2\mathrm{Var}\left(h^{(1)}(X)\right)$ and therefore,

$$\mathrm{Var}\left(\Lambda_B(\mathcal{P})\right) \le \frac{8}{|B|}\mathrm{Var}\left(h^{(1)}(X)\right) .$$

Moreover,

$$\begin{aligned}
\mathrm{Var}\left(h^{(1)}(X)\right) &\le \mathbb{E}_{X'}\left[\mathbb{E}_X\left[h_{\mathcal{P}}(X, X')\right]^2\right] \\
&\le \mathbb{E}_{X'}\left[\mathbb{E}_X\left[D(X, X')^2\right]\mathbb{E}_X\left[\left(\Phi_{\mathcal{P}}(X, X') - \Phi_{\mathcal{P}^*}(X, X')\right)^2\right]\right] \\
&= \mathbb{E}_{X'}\left[\mathbb{E}_X\left[D(X, X')^2\right]\mathbb{P}_X\left(\Phi_{\mathcal{P}}(X, X') \neq \Phi_{\mathcal{P}^*}(X, X')\right)\right] \\
&\le \sigma^2\kappa(W(\mathcal{P}) - W^*)^\alpha
\end{aligned}$$

where $\mathbb{E}_X$ (resp. $\mathbb{E}_{X'}$) refers to the expectation taken with respect to X (resp. X'). Chebyshev's inequality gives, for $r \in (0, 1)$,

$$\mathbb{P}\left(\Lambda_B(\mathcal{P}) - \Lambda(\mathcal{P}) > \sigma(W(\mathcal{P}) - W^*)^{\alpha/2}\sqrt{\frac{8\kappa}{r|B|}}\right) \le r .$$

Using again (4.11) with $r = \frac{1}{4}$, by $|B| \ge \frac{n}{128\lceil\log(N/\delta)\rceil}$, there exists a constant C such that for any $\mathcal{P} \in \Pi_K$, with probability at least $1 - 2\delta/N$,

$$|\overline{\Lambda}_{\mathcal{B}}(\mathcal{P}) - \Lambda(\mathcal{P})| \le C\sigma(W(\mathcal{P}) - W^*)^{\alpha/2}\sqrt{\frac{\lceil\log(N/\delta)\rceil}{n}} .$$

This implies by the union bound, that

$$|\overline{W}_{\mathcal{B}}(\widehat{\mathcal{P}}) - W(\widehat{\mathcal{P}})| \leq K\sigma(W(\widehat{\mathcal{P}}) - W^*)^{\alpha/2} \sqrt{\frac{\lceil \log(N/\delta) \rceil}{n}}$$

with probability at least $1 - 2\delta$. Using (4.18), we obtain

$$(W(\widehat{\mathcal{P}}) - W^*)^{1-\alpha/2} \leq 2K\sigma \sqrt{\frac{\lceil \log(N/\delta) \rceil}{n}} \,,$$

concluding the proof.

Bibliography

[1] E. A. Abaya and G. L. Wise. Convergence of vector quantizers with applications to optimal quantization. *SIAM Journal on Applied Mathematics*, 44:183–189, 1984.

[2] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58:137–147, 2002.

[3] T. Andersen, D. Dobrev, and E. Schaumburg. Jump-robust volatility estimation using nearest neighbor truncation. *Journal of Econometrics*, 169(1):75–93, 2012.

[4] A. Antos. Improved minimax bounds on the test and training distortion of empirically designed vector quantizers. *IEEE Transactions on Information Theory*, 51:4022–4032, 2005.

[5] A. Antos, L. Györfi, and A. György. Improved convergence rates in empirical vector quantizer design. *IEEE Transactions on Information Theory*, 51:4013–4022, 2005.

[6] M. A. Arcones and E. Giné. Limit theorems for U-processes. *The Annals of Probability*, 21:1494–1542, 1993.

[7] J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39:2766–2794, 2011.

[8] K. Ball. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31:1–58, 1997.

[9] P. Bartlett, T. Linder, and G. Lugosi. The minimax distortion redundancy in empirical quantizer design. *IEEE Trans. Inform. Theory*, 44:1802–1813, Sep. 1998.

[10] P.L. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory Related Fields*, 135:311–334, 2006.

[11] G. Biau and K. Bleakley. Statistical inference on graphs. *Statistics & Decisions*, 24(2):209–232, 2006.

[12] G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory*, 54:781–790, 200.

[13] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM. Probability and Statistics*, 9:323–375, 2005.

[14] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford, 2013.

[15] C. Brownlees, E. Joly, and G. Lugosi. Empirical risk minimization for heavy-tailed losses. *arXiv:1406.2462v2*, 2015.

[16] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Badits with heavy tail. *IEEE Transactions on Information Theory*, 59:7711–7717, 2013.

[17] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[18] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[19] O. Catoni. *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI-2001*. Springer, 2004.

[20] O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012.

[21] S. Clémençon. A statistical view of clustering performance through the theory of U-processes. *Journal of Multivariate Analysis*, 124:42 – 56, 2014.

[22] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, pages 844–874, 2008.

[23] V. de la Peña and E. Giné. *Decoupling: from dependence to independence*. Springer, New York, 1999.

[24] V. de la Peña. Decoupling and Khintchine's inequalities for U-statistics. *The Annals of Probability*, pages 1877–1892, 1992.

[25] R.M. Dudley. Central limit theorems for empirical measures. *Annals of Probability*, 6:899–929, 1978.

[26] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance anf Finance*. Springer-Verlang, Frankfurt, 1997.

[27] E. F. Fama. Mandelbrot and the stable Paretian hypothesis. *The Journal of Business*, 36:420–429, 1963.

[28] B. Finkenstadt and H. Rootzén. *Extreme Values in Finance, Telecommunications and the Enviroment*. Chapman and Hall, New York, 2003.

[29] E. Giné, R. Latała, and J. Zinn. Exponential and moment inequalities for U-statistics. In *High Dimensional Probability II—Progress in Probability*, pages 13–38. Birkhauser, 2000.

[30] W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, pages 293–325, 1948.

[31] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[32] D. Hsu and S. Sabato. Approximate loss minimization with heavy tails. *Computing Research Repository*, abs/1307.1827, 2013.

[33] P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[34] P. J. Huber and E. Ronchetti. *Robust Statistics*. Wiley-Interscience, 2nd edition edition, 2009.

[35] M. Hubert, P. J. Rousseeuw, and S. Van Aelst. High-breakdown robust multivariate methods. *Statistical Science*, pages 92–119, 2008.

[36] M. Jerrum, L. Valiant, and V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:186–188, 1986.

[37] E. Joly and G. Lugosi. Robust estimation of u-statistics. *arXiv:1504.04580*, 2015.

[38] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 36:00–00, 2006.

[39] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*. Springer Science & Business Media, 2011.

[40] G. Lecué. Suboptimality of penalized empirical risk minimization in classification. *COLT07*, pages 142–156.

[41] G. Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, (4):1000–1022, 2007.

[42] M. Lerasle and R. I. Oliveira. Robust empirical mean estimators. *arXiv:1112.3914*, 2011.

[43] C. Levrard. Fast rates for empirical vector quantization. *Electronic Journal of Statistics*, pages 1716–1746, 2013.

[44] C. Levrard. Nonasymptotic bounds for vector quantization in hilbert spaces. *The Annals of Statistics*, 43(2):592–619, 2015.

[45] T. Linder. Learning-theoretic methods in vector quantization. In L. Györfi, editor, *Principles of nonparametric learning*, number 434 in CISM Courses and Lecture Notes. Springer-Verlag, New York, 2002.

[46] E. Mammen and A. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

[47] B. Mandelbrot. The variation of certain speculative prices. *The Journal of Business*, 36:394–419, 1963.

[48] P. Massart. *Concentration inequalities and model selection*. Ecole d'été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics. Springer, 2006.

[49] J. Matoušek. *Lectures on discrete geometry*. Springer, 2002.

[50] A. Maurer and M. Pontil. *k*-dimensional coding schemes in Hilbert spaces. *IEEE Transactions on Information Theory*, 56:5839–5846, 2010.

[51] S. Mendelson. Learning without concentration. *arXiv preprint arXiv:1401.0304*, 2014.

[52] S. Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 2015.

[53] A.S. Nemirovsky and D.B. Yudin. Problem complexity and method efficiency in optimization. 1983.

[54] J. P. Nolan. *Stable Distributions - Models for Heavy Tailed Data*. Birkhauser, Boston, 2015. In progress, Chapter 1 online at academic2.american.edu/~jpnolan.

[55] D. Pollard. Strong consistency of *k*-means clustering. *Annals of Statistics*, 9, no. 1:135–140, 1981.

[56] D. Pollard. A central limit theorem for *k*-means clustering. *Annals of Probability*, 10(4):919–926, 1982.

[57] D. Pollard. Quantization and the method of *k*-means. *IEEE Trans. Inform. Theory*, IT-28:199–205, 1982.

[58] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[59] J. Robins, L. Li, E. Tchetgen, and A. van der Vaart. Quadratic semiparametric von Mises calculus. *Metrika*, 69(2-3):227–247, 2009.

[60] C. Rogers. *Packing and covering*. Cambridge University Press, 1964.

[61] M. Talagrand. *The generic chaining*. Springer, 2005.

[62] M. Telgarsky and S. Dasgupta. Moment-based uniform deviation bounds for k-means and friends. *arXiv preprint arXiv:1311.1903*, 2013.

[63] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge, UK, 2000.

[64] A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes*. Springer-Verlag, New York, 1996.

[65] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

[66] N. Vayatis. Applications of concentration inequalities for statistical scoring and ranking problems. In *ESAIM: Proceedings*, volume 44, pages 99–109. EDP Sciences, 2014.

[67] T. Zhang and G. Lerman. A novel m-estimator for robust pca. *The Journal of Machine Learning Research*, 15(1):749–808, 2014.

[68] VM. Zolotarev. *One-dimensional stable distributions*, volume 65. American Mathematical Soc., 1986.

Estimation de distributions à queue lourde

Nous nous intéressons à estimer la moyenne d'une variable aléatoire de loi à queue lourde. Nous adoptons une approche plus robuste que la moyenne empirique classique communément utilisée. L'objectif est de développer des inégalités de concentration de type sous-gaussien sur l'erreur d'estimation. En d'autres termes, nous cherchons à garantir une forte concentration sous une hypothèse plus faible que la bornitude : une variance finie. Deux estimateurs de la moyenne pour une loi à support réel sont invoqués et leurs résultats de concentration sont rappelés. Plusieurs adaptations en dimension supérieure sont envisagées. L'utilisation appropriée de ces estimateurs nous permet d'introduire une nouvelle technique de minimisation du risque empirique pour des variables aléatoires à queue lourde. Quelques applications de cette technique sont développées. Nous appuyons ces résultats sur des simulations sur des jeux de données simulées. Dans un troisième temps, nous étudions un problème d'estimation multivarié dans le cadre des U-statistiques où les estimateurs précédents offrent, là aussi, une généralisation naturelle d'estimateurs présents dans la littérature.

Robust estimation of heavy-tailed distributions

In this thesis, we are interested in estimating the mean of heavy-tailed random variables. We focus on a robust estimation of the mean approach as an alternative to the classical empirical mean estimation. The goal is to develop sub-Gaussian concentration inequalities for the estimating error. In other words, we seek strong concentration results usually obtained for bounded random variables, in the context where the bounded condition is replaced by a finite variance condition. Two existing estimators of the mean of a real-valued random variable are invoked and their concentration results are recalled. Several new higher dimension adaptations are discussed. Using those estimators, we introduce a new version of empirical risk minimization for heavy-tailed random variables. Some applications are developed. These results are illustrated by simulations on artificial data samples. Lastly, we study the multivariate case in the U-statistics context. A natural generalization of existing estimators is offered, once again, by previous estimators.